Further Topics in Mathematics of Data Science (HS 2025)

Almut Rödder

December 2025

Contents

Disclaimer: These are the notes for Further Topics in Mathematics of Data Science accompanying the lecture held by Afonso Bandeira in Autumn 2025 based on this book, written by Afonso Bandeira, Amit Singer and Thomas Strohmer. Parts of these notes are based on previous years's notes when the course was held by Antoine Maillard.

These are personal lecture notes, shared for convenience and not intended as a polished or authoritative reference. They may contain errors, omissions, or informal phrasing, and have not been reviewed or verified for accuracy. They do not represent the official views of any institution, course, or instructor. Use at your own discretion — always consult primary sources, textbooks, or your instructor for anything critical.

1. Feige's Conjecture

We will start this session with a remarkable conjecture about small deviations stated by Feige [Fei06] in 2006 that is easy to believe, but it seems harder to prove:

Conjecture 1.1. *Let $X_1, \dots, X_n$ be non-negative independent random variables with $\mathbb{E}[X_i] = 1$. For $X := \sum_{i=1}^n X_i$ holds*

$$\mathbb{P}(X < \mathbb{E}[X] + 1) \geq \frac{1}{e}. \tag{1.1}$$

When Feige stated this conjecture, he showed that it holds for $\alpha = \frac{1}{13}$ instead of $\alpha = \frac{1}{e}$ by a case-by-case analysis. This already means that it the probability for a sum of independent random variables to exceed the mean is bounded away from one.
Feige came up with this conjecture to construct algorithms and guarantees to determine the average degree d of a graph in sublinear ($O(\sqrt{n})$) time (when the graph has n vertices and the average degree is bounded away from 0). The original paper [Fei06] provides a very intuitive explanation of this procedure. The upper bound is indeed easy to prove when 1.1 holds:
Take the following algorithm: In every step $i \in [N]$, choose a vertex uniformly at random and count the number of neighbors, d_i. The choice is independent, so with replacement. If the algorithm runs for N steps, the estimator is given by

$$\hat{d} = \frac{1}{N} \sum_{i=1}^{N} d_i.$$

Note that it is unbiased and with 1.1, we have

$$\mathbb{P}(\hat{d} \leq (1 + \frac{1}{N})d) \geq \alpha.$$

Now, repeat this procedure $\frac{2}{\alpha}$ times and take the minimal number $\hat{d}_{\min}$ as estimator. Then, it holds

$$\mathbb{P}(\hat{d}_{\min} \leq (1 + \frac{1}{N})d) \geq 1 - (1 - \alpha)^{\frac{2}{\alpha}} \geq 1 - \frac{1}{e^2} \geq \frac{5}{6}.$$

Feige [Fei06] shows in his paper that the algorithm that produces $\hat{d}_{\min}$ also guarantees

$$\mathbb{P}(\hat{d}_{\min} \geq (1 - \varepsilon)\frac{d}{2}) \geq \frac{5}{6}$$

if $N = \frac{72}{\alpha\varepsilon}\sqrt{\frac{n}{d_0}}$ and $d \geq d_0$ for a fixed $d_0 > 0$.
Let us go back to the conjecture and convince ourselves that the conjecture is intuitive: First, if X is symmetric around its mean, then it is clear that the probability is bounded by $\frac{1}{2}$. So, we have to construct examples that are not symmetric around its mean. One thing that might come into our mind is Markov's inequality:

$$\mathbb{P}(X < \mathbb{E}[X] + 1) \geq 1 - \frac{n}{n+1}.$$

Note that this expression converges to 0 for $n \to \infty$ and therefore is very uninformative. This is not so surprising: Markov's inequality does not need any assumptions and is therefore quite loose. But this is not the only problem here: For such small deviations, concentration inequalities like Markov, Chebyshev or Hoeffding are not capable. Instead, they are useful for tail bounds.

An example that does not achieve $\alpha = \frac{1}{2}$ cannot be symmetric around its mean. Try to construct it as an exercise:

Exercise 1.2. *Construct an example for which* $\alpha = \frac{1}{e}$.

Remark 1.3. *Feige's bound of* $\frac{1}{13}$ *was improved to* $\frac{1}{8}$ *by He, Zhang and Zhang [HZZ10] in 2010. Garnett improved it in 2020 to* $\approx \frac{7}{50}$. *Both used moment bounds (Garnett included a bound on third moments, instead of first, second and fourth). This is possible because Feige showed in his original paper that* X_i *just need to be supported on two values. Nevertheless, the proofs are non trivial and require many case distinctions. If you are interested, you should definitely have a look at these papers! The current best known bound is due to Guo et al. [GHLL20] and yields* 0.1798. *Some of the techniques used in this paper you should be able to recognize after a few lectures!*

Instead of diving deeper into improving the gap for Feige's conjecture (1.1), we will focus on a subclass of distributions for which the constant $\frac{1}{e}$ indeed holds, namely discrete and continuous log-concave functions. For this we will follow the work of Alqasem et al. [AAMM24] from 2024. They showed that (1.1) holds true for X when it has a discrete log-concave distributions, even with negative support. This generalizes to sums of independent log-concave observations like in the original conjecture and to continuous log-concave distributions. The continuous case was already known due to a sharper bound of Grünbaum [Grü60] from the 1960s.

Let us introduce discrete log-concave random variables:

Definition 1.4. *An integer-valued distribution with probabilities* $\{p(k)\}_{k\in\mathbb{Z}}$ *is said to be log-concave if for every* $k \in \mathbb{Z}$ *it holds*

$$p(k)^2 \geq p(k-1)p(k+1)$$

and it has a connected support.

First, try to come up with some examples:

Exercise 1.5. *Show that the following distributions are log-concave: Discrete Uniform, Geometric, Poisson, Bernoulli, Binomial. Also try to come up with an example of an integer-valued distribution that is not log-concave.*

The main theorem from [AAMM24] is the following:

Theorem 1.6. *Let* X *be discrete log-concave, then*

$$\mathbb{P}(X < \mathbb{E}[X] + 1) \geq \frac{1}{e}. \tag{1.2}$$

They also show that this bound is sharp, i.e. that discrete log-concave random variables can be constructed that reach $\frac{1}{e}$ in a limit. By studying the proof, we will see this implicitly.

Remark 1.7. *Note that the Theorem is stated without assumptions on* $\mathbb{E}[X]$, X *could potentially be negative and not necessarily a sum of independent random variables is mentioned. But note that the sum of independent log-concave distributions is log-concave as well. The discrete version of this is known as Hoggar's Theorem and was proven in 1974. We will focus on the* $X \geq 0$ *case because it allows us to skip one step of the proof. If you are interested in the details, see the original work [AAMM24].*

We leave the implication for continuous log-concave distributions as a standard probability-theory exercise:

Exercise 1.8. *Show that the discrete log-concave case implies the continuous log-concave case.*

We will now turn to the proof of Theorem 1.6. The first step reduces the set of log-concave distributions to the ones with compact support only and is left as an exercise:

Exercise 1.9. *Show that it is sufficient to show* (1.2) *for compactly supported* $X \geq 0$ *only.*

Remark 1.10. *This part becomes a bit more tricky if we allow X to be supported on all $\mathbb{Z}$.*

With this fact in mind, we will show that Theorem 1.6 holds for every log-concave distribution with (arbitrary) compact support. We can further reduce this class by applying the Krein-Milman Theorem which allows us to study truncated geometric functionals only. It essentially says that extreme points for compactly supported log-concave distributions are attained by probability distributions of the type

$$p(k) = Cp^k, \quad k \in \{M, \dots, N\},$$

where C is the normalizing constant depending on $0 < p \leq 1$, M and N. We can shift M and N to make X be supported on $\{1, \dots, n\}$ in general. First, let $p = 1$. It is easy to show that (1.2) holds. Therefore, assume $p < 1$ now.

Exercise 1.11. *Compute C and show that*

$$\mathbb{E}[X] = \frac{1}{1-p} - \frac{np^n}{1-p^n}$$

for $p < 1$ and $n \in \mathbb{N}$.

Having this expression for $\mathbb{E}[X]$, we can compute

$$\mathbb{P}(X < \mathbb{E}[X] + 1) = \sum_{k=1}^{\lceil \mathbb{E}[X] \rceil} Cp^k = \frac{1 - p^{\lceil \mathbb{E}[X] \rceil}}{1 - p^n}.$$

Now, plug in $\mathbb{E}[X]$ and since the expression is non-increasing in n, show that the limit $n \to \infty$ is reaches $\frac{1}{e}$.

2. Discrepancy Theory

In class you already learned about *Komlos Conjecture*. We want to explore the background of this, coming from Discrepancy Theory. Let us start with an example to motivate this Theory: Suppose you have a set of n observations X_i with m binary observables $X_{ij} \in \{\pm 1\}$ each. For each of the m observables we define the set $S_i = \{j \in [n] : X_{ij} = 1\}$. The goal is to split $[n]$ into two groups such that the characteristics are equally split in both groups, i.e. if we have the two groups $G_1, G_2 \subseteq [n]$, it should hold that

$$|G_1 \cap S_i| \approx |G_2 \cap S_i| \quad \forall i \in [m].$$

We denote the mapping into the two groups by a function $\chi : [n] \to \{\pm 1\}$, where the preimage of $\{1\}$ defines G_1, and $\{-1\}$ for G_2 respectively. The discrepancy of this map is the worst difference in sizes of the groups and formally defined as

$$\operatorname{disc}_X(\chi) = \max_{i \in m} \left| \sum_{j \in S_i} \chi(j) \right|$$

where $\left| \sum_{j \in S_i} \chi(j) \right|$ is simply the difference of numbers of elements in G_1 and G_2 that have characteristic j. We define the discrepancy of the dataset X as the minimum of discrepancies over all maps:

$$\operatorname{disc}(X) = \min_{\chi : [n] \to \{\pm 1\}} \operatorname{disc}_X(\chi)$$

Example 2.1 (Discrepancy with four students). *Consider four students: Alice (1), Ben (2), Carol (3), and Dan (4), who are to be split into two study groups G_1, G_2. Each student has three binary traits: likes basketball $(+1)$ or not (-1), likes chess $(+1)$ or not (-1) and likes coding $(+1)$ or not (-1). We record traits as a matrix $X \in \{\pm 1\}^{3 \times 4}$:*

$$X = \begin{bmatrix} +1 & +1 & -1 \\ +1 & -1 & +1 \\ +1 & -1 & +1 \\ -1 & +1 & -1 \end{bmatrix}^\top.$$

For each observable i, define

$$S_1 = \{1, 2, 3\} \quad \textit{(basketball)}, \qquad S_2 = \{1, 4\} \quad \textit{(chess)}, \qquad S_3 = \{2, 3\} \quad \textit{(coding)}.$$

Since $|S_1| = 3$ is odd, it is impossible to split S_1 evenly between two groups. Hence for any partition $\chi : [4] \to \{\pm 1\}$ the imbalance

$$\big|\, |G_1 \cap S_1| - |G_2 \cap S_1| \,\big|$$

is at least 1. Thus the discrepancy of this instance is ≥ 1.

Choose the signing $\chi = (+1, +1, -1, -1)$, i.e.

$$G_1 = \{1, 2\}, \qquad G_2 = \{3, 4\}.$$

Then:

$$\begin{aligned} \textit{For basketball } S_1 &: \ |G_1 \cap S_1| = 2,\ |G_2 \cap S_1| = 1 \ \Rightarrow\ \textit{imbalance } = 1, \\ \textit{For chess } S_2 &: \ |G_1 \cap S_2| = 1,\ |G_2 \cap S_2| = 1 \ \Rightarrow\ \textit{imbalance } = 0, \\ \textit{For coding } S_3 &: \ |G_1 \cap S_3| = 1,\ |G_2 \cap S_3| = 1 \ \Rightarrow\ \textit{imbalance } = 0. \end{aligned}$$

Thus the maximum imbalance is 1, which equals the lower bound. Therefore the minimal discrepancy of this example is exactly 1.

This simple example is easily solved by hand - but it is not so clear how to find bounds if there are more students and arbitrarily many characteristics. The general discrepancy theory looks at the **discrepancy of vectors**:

Definition 2.2 (Discrepancy of Vectors). *Given a set of vectors $\{u_1, \dots, u_n\}$, define the discrepancy as*

$$\operatorname{disc}(u_1, \dots, u_n) := \min_{\varepsilon \in \{\pm 1\}^n} \| \sum_{j=1}^{n} \varepsilon_j u_j \|_\infty$$

Remark 2.3. *This is consistent with what we defined before: The vector u_i can be defined as $(u_j)_i := \mathbb{1}_{X_{ij}=1}$ and ε defines the mapping χ. Therefore, $\varepsilon_j (u_i)_j = \chi(j) \mathbb{1}_{j \in S_i}$.*

In 1985, Spencer [Spe85] showed the following remarkable upper bound for any choice of vectors u_j with $\|u_j\|_\infty \leq 1$ and where $n = m$:

Theorem 2.4 (Spencer: "Six Deviations Suffice"). *Let $n \geq 1$. There exists a constant $C > 0$ such that*

$$\operatorname{disc}(u_1, \dots, u_n) \leq C\sqrt{n}$$

for all $u_1, \dots, u_n \in \mathbb{R}^n$ with $\|u_j\|_\infty \leq 1$ for all $j \in [n]$.

It is possible to show $C \leq 6$. A nice sketch of the proof can be found in the lecture notes from last year's course held by Antoine Maillard. One might ask whether the scaling in $\sqrt{n}$ is really necessary. We will provide an example here:

Example 2.5. *A* Hadamard matrix *of order n is a square matrix $H \in \{\pm 1\}^{n \times n}$ such that its rows are mutually orthogonal, i.e.*

$$H H^\top = n I_n.$$

Suppose these objects exist for a fixed $n \in \mathbb{N}$. Show that for $u_1, \dots, u_n$ being the rows of H, it holds that $\operatorname{disc}(u_1, \dots, u_n) \geq \sqrt{n}$.

The construction of such Hadamard matrices is not trivial at all. Try to rebuild it for exponentials of 2, known as Sylvesters construction:

Exercise 2.6. *Can you show that Hadamard matrices exist for $n = 2^k$, $k \geq 1$? Start with $k = 1$ and construct them inductively using H_1 and H_{2^k} to construct $H_{2^{k+1}}$.*

Actually, the existence of Hadamard matrices for general $n = 4k$ (which is necessary for the existence) is still open:

Conjecture 2.7 (Hadamard Conjecture). *For every $n = 4k$, $k \geq 1$, there exists a Hadamard matrix.*

Note that in the previous example, we constructed the vectors u with values in $\{\pm 1\}$. In Theorem 2.4 we learned that the discrepancy restricted to the ℓ_∞-ball can be bounded by order $\sqrt{n}$. Note that the ℓ_∞-ball is a subset of $\sqrt{n} B_2$ where B_2 denotes the ℓ_2-ball. It is still open if the bound from 2.4 generalizes to the ℓ_2-ball. This is indeed the statement of Komlos Conjecture that you learned about in the course:

Conjecture 2.8 (Komlos). *For every set of vectors with $\|u_j\|_2 \leq 1$, it holds that*

$$\operatorname{disc}(u_1, \dots, u_n) \leq K$$

for a universal constant $K > 0$.

Note that the conjecture implies Spencer's Theorem. Recent improvent by Bansal and Jiang [BJ25] shows that $K \leq O(\log(n)^{\frac{1}{4}})$ while it was previously open to improver over $\sqrt{\log(n)}$ for more than 25 years.

A lower bound on the discrepancy for the discrepancy for binary vectors is also not known for an infinite family of vectors:

Open Problem 2.9. *Is the value of*

$$\sup_{n \in \mathbb{N}} \sup_{\{u_j\}_{j \in [n]} \in \{\frac{\pm 1}{\sqrt{n}}\}^n} \operatorname{disc}(u_1, \dots, u_n) > 1?$$

3. Concentration Inequalities

3.1. Subgaussian Random Variables

In the lecture you proved Hoeffding's ineuqality for Subgaussian random variables. Let us first recall what a Subgaussian random variable is:

Definition 3.1 (Sub-Gaussian random variable). *A real-valued random variable X is called* sub-Gaussian *if it satisfies any (and hence all) of the following equivalent conditions:*

1. ***Moment generating function (MGF) bound:*** *There exists a constant $C_1 > 0$ such that for all $t \in \mathbb{R}$,*
$$\mathbb{E}[\exp(tX)] \leq \exp\Big(\frac{C_1^2 t^2}{2}\Big).$$
2. ***Tail bound:*** *There exists a constant $C_2 > 0$ such that for all $t \geq 0$,*
$$\mathbb{P}(|X| \geq t) \leq 2\exp\Big(-\frac{t^2}{C_2^2}\Big).$$
3. ***Moment growth:*** *There exists a constant $C_3 > 0$ such that for all integers $p \geq 1$,*
$$(\mathbb{E}|X|^p)^{1/p} \leq C_3\sqrt{p}.$$

Exercise 3.2. *Try to prove the equivalence of the three definitions.*

Remark 3.3. *We say X is σ-Subgaussian if* $\mathbb{E}[\exp(tX)] \leq \exp\Big(\frac{\sigma^2 t^2}{2}\Big)$.

Hoeffding's Inequality states the following:

Theorem 3.4 (Hoeffding's Inequality). *Let $X_1, X_2, \ldots, X_n$ be independent, mean-zero, σ_i-sub-Gaussian random variables. Define the sum $S_n = \sum_{i=1}^n X_i$. Then, for any $t > 0$,*

$$\mathbb{P}(|S_n| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right),$$

where $c > 0$ is an absolute constant.

We will see now that Hoeffding's inequality can help us with proving high-dimensional phenomena, but it also helps us to study the quality of simple estimators by using concentration of i.i.d. observables. We start with the following exercise:

Exercise 3.5. *Find a lower bound for the following question: How many vectors $u \in \mathbb{R}^d$ exist such that*

$$|\langle u_i, u_j\rangle| \leq \varepsilon$$

for all pairwise u_i, u_j, $i \neq j$?

We can also bound the expected maximum of *i.i.d.* Subgaussian random variables:

Exercise 3.6. *Show that for n i.i.d. copies of a σ-Subgaussian random variable it holds:*

$$\mathbb{E}[\max_{i\in[n]} |X_i|] \leq C\sigma\sqrt{\log(n)}$$

for a fixed constant $C > 0$.

The following example is taken from [Ver09]. We will use Hoeffding's inequality to show concentration of the *Median of Means Estimator*, independent of higher than second moments. This is possible by combining Chebyshevs inequality and Hoeffding for bounded random variables.

Theorem 3.7 (Median-of-Means Concentration). *Let X be a random variable with mean μ and variance σ^2, and let $X_1, \dots, X_N$ be independent copies of X. Then for any $0 \le t \le \sqrt{N}$, there exists an estimator $\hat{\mu} = \hat{\mu}(X_1, \dots, X_N)$ such that*

$$\mathbb{P}\Big(|\hat{\mu} - \mu| \ge t\frac{\sigma}{\sqrt{N}}\Big) \le 2\exp(-ct^2),$$

where $c > 0$ is an absolute constant.

Proof: For simplicity, assume $N = BL$ for integers B and L. Divide the sample $X_1, \dots, X_N$ into B blocks of length L and compute the block means

$$\mu_b = \frac{1}{L}\sum_{i=(b-1)L+1}^{bL} X_i, \quad b = 1, \dots, B,$$

then define the estimator as their median:

$$\hat{\mu} = \mathrm{Med}(\mu_1, \dots, \mu_B).$$

Step 1: Variance of block means. Each μ_b is an average of L independent copies of X, so

$$\mathbb{E}[\mu_b] = \mu, \qquad \mathrm{Var}(\mu_b) = \frac{\sigma^2}{L}.$$

Step 2: Chebyshev inequality for a single block. By Chebyshev's inequality,

$$\mathbb{P}\Big(\mu_b \ge \mu + t\frac{\sigma}{\sqrt{N}}\Big) \le \frac{\mathrm{Var}(\mu_b)}{(t\sigma/\sqrt{N})^2} = \frac{\sigma^2/L}{t^2\sigma^2/N} = \frac{N}{t^2 L} = \frac{B}{t^2}.$$

Choosing the number of blocks $B = t^2/4$, we get

$$\mathbb{P}\Big(\mu_b \ge \mu + t\frac{\sigma}{\sqrt{N}}\Big) \le \frac{1}{4}.$$

Step 3: Median-of-means bound. By definition of the median,

$$\mathbb{P}\Big(\hat{\mu} \ge \mu + t\frac{\sigma}{\sqrt{N}}\Big) \le \mathbb{P}\Big(\text{at least half of the } \mu_b \text{ satisfy } \mu_b \ge \mu + t\frac{\sigma}{\sqrt{N}}\Big).$$

The events $\{\mu_b \ge \mu + t\sigma/\sqrt{N}\}$ are independent and each has probability at most $1/4$. Applying Hoeffding's inequality for the sum of B independent Bernoulli indicators Z_i with expectation bounded by $\frac{1}{4}$ each yields

$$\mathbb{P}\Big(\hat{\mu} \ge \mu + t\frac{\sigma}{\sqrt{N}}\Big) = \mathbb{P}\Big(\sum_{i=1}^{B} Z_i \ge \frac{B}{2}\Big) \le \exp(-\frac{(\frac{B}{4})^2}{2B}) = \exp(-t^2/128).$$

Step 4: Symmetry. Similarly,

$$\mathbb{P}\Big(\hat{\mu} \le \mu - t\frac{\sigma}{\sqrt{N}}\Big) \le \exp(-c_0 t^2/4).$$

Step 5: Combine bounds. By a union bound, we get

$$\mathbb{P}\Big(|\hat{\mu} - \mu| \geq t\frac{\sigma}{\sqrt{N}}\Big) \leq 2\exp(-ct^2),$$

where c is an absolute constant, completing the proof. ■

3.2. Subexponential and Bernstein

We will close this section with Bernstein's inequality applied to subexponential random variables. In contrast to Subgaussian random variables, these can have heavy tails and therefore for large deviations, tail bounds are governed by a weaker behaviour than in the Subgaussian case. For small deviations, we will see Subgaussian behaviour, consistent with the Central Limit Theorem. Let us introduce Subexponential random variables:

Definition 3.8 (Subexponential Random Variables). *A real-valued random variable X is called* sub-Exponential *if it satisfies any (and hence all) of the following equivalent conditions:*

1. ***Moment generating function (MGF) bound:*** *There exists a constant $C_1 > 0$ such that for all $t \in \mathbb{R}$,*
$$\mathbb{E}[\exp(tX)] \leq \exp\Big(C_1^2 t^2\Big) \quad \forall |t| \leq \frac{1}{C_1}.$$
2. ***Tail bound:*** *There exists a constant $C_2 > 0$ such that for all $t \geq 0$,*
$$\mathbb{P}(|X| \geq t) \leq 2\exp\Big(-\frac{t}{C_2}\Big).$$
3. ***Moment growth:*** *There exists a constant $C_3 > 0$ such that for all integers $p \geq 1$,*
$$(\mathbb{E}|X|^p)^{1/p} \leq C_3 p.$$

Exercise 3.9. *Show that the following Random Variables are Subexponential*

- *Any Subgaussian*
- *Poisson*
- *Exponential*

Why are we interested in Subexponential Random Variables at all? If we want to show concentration of $\|X\|_2^2$ for a random Gaussian vector X consisting of n independent Gaussian entries, we can see by Gaussian Tail bounds that

$$\mathbb{P}(\|X\|_2^2 \geq t) \sim \exp(-Ct) \tag{3.1}$$

instead of $-Ct^2$. This is because X_i^2 does not behave Subgaussian anymore and has heavy tails. We still expect concentration in (3.1), but with slower rate. We will make this rigorous now.

Theorem 3.10 (Bernstein's Inequality for Subexponential Random Variables). *Let $X_1, \ldots, X_n$ be independent, mean-zero subexponential random variables with parameters $\nu_1, \ldots, \nu_n$. Then for every $t \geq 0$ we have*

$$\mathbb{P}\left(\Big|\sum_{i=1}^n X_i\Big| \geq t\right) \leq 2\exp\left(-\min\left(\frac{t^2}{4\sum_{i=1}^n \nu_i^2}, \frac{t}{2\max_{i\in[n]}|\nu_i|}\right)\right)$$

Exercise 3.11. *Prove Theorem 3.10 with the Chernoff bound you learned in class.*

4. Approximate Message Passing Algorithms

In class, you learned about Principal Component Analysis to recover the leading eigenvector of a matrix. For some statistical estimation problems, this is an efficient algorithm (meaning it succeeds in polynomial time), but not the best algorithm, i.e. there exist algorithms that achieve a better approximation to the solution of a statistical estimation problem than PCA while still running in polynomial time. We illustrate this with the problem of $\mathbb{Z}_2$-Synchronisation. This studies the following model:

$$Y = \lambda \frac{xx^\top}{n} + W, \quad x \in \{\pm 1\}^n, \quad W \sim \mathrm{GOE}(n),$$

where $\lambda \geq 0$ is the signal-to-noise ratio, $x \in \{\pm 1\}^n$ is the signal we want to recover and comes from the prior μ, and the GOE matrix W has independent (up to symmetry) Gaussian entries with variance $\frac{1}{n}$ off-diagonal, and $\frac{2}{n}$ on the diagonal.

We know from basic statistics that the Bayes Estimator is the best estimator for the respective error. If we take the ℓ_2-norm, this corresponds to the expectation under the posterior $\mathbb{P}(z \mid Y)$. Therefore, let us derive the posterior when μ is the uniform prior:

$$\begin{aligned}
\mathbb{P}(x = z \mid Y) &= f(Y \mid z)\mu(z)\frac{1}{f(Y)} \\
&\propto \exp\Big(-\frac{n}{4}\|Y - \lambda\tfrac{zz^\top}{n}\|_F^2\Big)\,\mu(z) \\
&\propto \exp\left(\frac{\lambda}{2} z^\top Y z\right),
\end{aligned}$$

where we used the Gaussian density of the GOE and dropped terms that do not depend on z. From this, we can derive the expectation of each coordinate x_i given the realisation of all remaining x_{-i}:

$$\begin{aligned}
\mathbb{P}(x_i = 1 \mid Y, x_{-i}) &= \frac{\exp(\lambda x_i \sum_{j\neq i} Y_{ij} x_j)}{\exp(\lambda \sum_{j\neq i} Y_{ij} x_j) + \exp(-\lambda \sum_{j\neq i} Y_{ij} x_j)} \\
&= \frac{\exp(\lambda Y_i x_{-i})}{\exp(\lambda Y_i x_{-i}) + \exp(-\lambda Y_i x_{-i})} \\
&= p_i^+,
\end{aligned}$$

where Y_i denotes the i-th row without entry i. This yields

$$\mathbb{E}[x_i \mid Y, x_{-i}] = p_i^+ - p_i^- = \tanh(\lambda Y_i x_{-i}). \tag{4.1}$$

Naively, to find $m_i = \mathbb{E}[x_i \mid Y]$, we could try to find the non-zero solutions of

$$m_i = \tanh(\lambda Y_i m_{-i}), \tag{4.2}$$

which is the solution to find stationary points of the *mean-field free energy*

$$F_{\mathrm{MF}}(m) = -\frac{\lambda}{2}\sum_{i\neq j} Y_{ij} m_i m_j + \sum_{i=1}^n \left[\frac{1+m_i}{2}\log\frac{1+m_i}{2} + \frac{1-m_i}{2}\log\frac{1-m_i}{2}\right].$$

The function $h(m_i) = \frac{1+m_i}{2}\log\frac{1+m_i}{2} + \frac{1-m_i}{2}\log\frac{1-m_i}{2}$ is the binary entropy function for $m \in [-1, +1]$. Note that we implicitly used Jensen's inequality to get from (4.1) to (4.2) and we assume that the system can be described through its *mean fields* only. We will see now that this simplification is not true in the case of a Gaussian noise matrix W. Instead, we observe the problem of backtracking, such that an algorithm finding the fixed points of (4.2) does not lead to the Bayes optimal solution. Instead, it will push high noise nodes

more and more. Let us consider $m^{(t-1)}$ at time step $t-1$. Note that node i influences all other nodes $j \neq i$ inside the tanh by $\lambda Y_{ji} m_i^{t-2}$. So, it *backtracks* in time step t to node i again by

$$m_i^t = \tanh(\sum_{j \neq i} \lambda Y_{ij} m_{j,-i}^{(t-1)} + \sum_{j \neq i} \lambda Y_{ij} \Delta m_j^{(t-1)})$$

where $m_{j,-i}^{(t-1)} = \tanh(\lambda Y_{jl} m_{-l}^{t-2} - \lambda Y_{ji} m_i^{t-2})$ denotes the update without backtracking and $\Delta m_j^{(t-1)} = m_j^{(t-1)} - m_{j,-i}^{(t-1)}$. A first order Taylor approximation yields

$$\begin{aligned} m_{j,-i}^{(t-1)} &= \tanh(\lambda Y_{jl} m_{-l}^{t-2}) - \lambda Y_{ji} m_i^{(t-2)} (1 - \tanh^2(\lambda Y_{jl} m_{-l}^{t-2})) \\ &= m_j^{(t-1)} - \lambda Y_{ji} m_i^{(t-2)} (1 - (m_j^{(t-1)})^2). \end{aligned}$$

This gives

$$m_i^t = \tanh(\sum_{j \neq i} \lambda Y_{ij} m_{j,-i}^{(t-1)} + \sum_{j \neq i} \lambda^2 Y_{ij}^2 (1 - (m_j^{(t-1)})^2) m_i^{(t-2)})$$

and the second term inside tanh is of constant order as the entries of Y are of order $\frac{1}{\sqrt{n}}$, Note that if Y was scaled with $\frac{1}{\sqrt{n}}$, the second term vanishes in the $n \to \infty$ limit and the naive approximation would lead to the optimal solution. We will see later in this course, that the optimal update is without backtracking, so we want to get rid off $\sum_{j \neq i} \lambda^2 Y_{ij}^2 (1 - (m_j^{(t-1)})^2)$. The Approximate Message Passing algorithm removes the backtracking part in every step by the following algorithm:

$$m_i^t = \tanh\left(\sum_j \lambda Y_{ij} m_j^{(t-1)} - \lambda^2 \left(1 - \frac{\|m^{(t-1)}\|_2^2}{n}\right) m_i^{(t-2)}\right). \tag{4.3}$$

Subtracting the second part is called *Onsager Correction*. This expression aims to minimize the *TAP free energy*, named after Thouless, Anderson, and Palmer:

$$F_{\text{TAP}}(m) = -\frac{\lambda}{2} \sum_{i \neq j} Y_{ij} m_i m_j + \sum_{i=1}^n h(m_i) - \frac{\lambda^2 n}{4} \left(1 - \frac{\|m\|_2^2}{n}\right)^2.$$

It was shown by Montanari and Venkataramanan [MV21], building on the work of Deshpande, Abbe, and Montanari [DAM16] that the AMP algorithm (4.3) converges to the Bayes optimal estimator in the sense that

$$\lim_{t \to \infty} \lim_{n \to \infty} \frac{1}{n^2} \|m^{(t)}(Y)(m^{(t)}(Y))^\top - \hat{X}_{\text{Bayes}} \hat{X}_{\text{Bayes}}^\top\|_F^2 \to 0,$$

when λ is above the algorithmic threshold and under suitable initialization $m^{(0)}$. Moreover, Celentano et al. [CFM23] prove (after Fan et al. [FMM21] proved it partially) the existence of stationary points of the TAP free energy corresponding to Bayes-optimal estimators of x in this model in the $n \to \infty$ limit and derive an efficient algorithm.

4.1. Plefka's Expansion and the TAP Free Energy

Another heuristic to motivate the TAP free energy (instead of the naive mean-field free energy) is the *Plefka expansion* [Ple82]. The idea is to approximate the Gibbs potential (i.e. the Legendre transform of the log-partition function) by a Taylor expansion around *infinite temperature* (equivalently, around zero coupling $\lambda = 0$). At infinite temperature the posterior distribution is the uniform measure, and the Gibbs potential reduces to the binary entropy function. Expanding around this tractable point and keeping terms up to second order in λ gives exactly the TAP correction term that we encountered earlier.

Consider a general Ising model with Hamiltonian

$$\mathbb{P}_\lambda(x) \propto \exp\Big(\frac{\lambda}{2} x^\top J x + h^\top x\Big), \qquad x \in \{\pm 1\}^n,$$

where J is a symmetric interaction matrix and h is an external field. Define the log-partition function

$$\mathcal{F}(\lambda, h) := \log\Big(\sum_{x\in\{\pm1\}^n} \exp\big(\tfrac{\lambda}{2} x^\top J x + h^\top x\big)\Big).$$

The *Gibbs potential* (or free energy expressed as a function of mean magnetizations $m \in (-1,1)^n$) is the Legendre transform

$$\Gamma_\lambda(m) := \sup_h \Big(h^\top m - \mathcal{F}(\lambda, h)\Big).$$

where the supremum is achieved at the field h that produces mean magnetization $m = \mathbb{E}_{\lambda,h}[x]$. By Legendre Duality, it suffices to approximate $\Gamma_\lambda(m)$ for all m to obtain $\mathcal{F}(\lambda, 0)$.

At $\lambda = 0$, the spins are independent with $\mathbb{P}(x_i = 1) = (1+m_i)/2$. In this case, it is easy to compute that

$$\Gamma_0(m) = \sum_{i=1}^n \Big[\tfrac{1+m_i}{2}\log\tfrac{1+m_i}{2} + \tfrac{1-m_i}{2}\log\tfrac{1-m_i}{2}\Big].$$

This is exactly the sum of binary entropy functions.

Expanding $\Gamma_\lambda(m)$ in λ around 0 gives (see [Ple82]):

$$\Gamma_\lambda(m) = \Gamma_0(m) - \frac{\lambda}{2} m^\top J m - \frac{\lambda^2}{4}\sum_{i,j} J_{ij}^2 (1-m_i^2)(1-m_j^2) + O(\lambda^3). \tag{4.4}$$

Minimizing this over m gives $-\mathcal{F}(\lambda, 0)$ directly by duality.

- The first correction $-\frac{\lambda}{2} m^\top J m$ is exactly the naive mean-field energy.
- The second correction $-\frac{\lambda^2}{4}\sum_{i,j} J_{ij}^2(1-m_i^2)(1-m_j^2)$ encodes the effect of pairwise fluctuations. This is the so-called *Onsager term.*

Exercise 4.1. *Let*

$$\mathcal{F}(\alpha, h) = \log \sum_{x\in\{\pm1\}^n} \exp\Big(\tfrac{\alpha}{2} x^\top J x + h^\top x\Big), \qquad \Gamma_\alpha(m) = \sup_h \big(h^\top m - \mathcal{F}(\alpha, h)\big),$$

where J is a symmetric matrix and $m \in (-1,1)^n$. Denote by $h(\alpha, m)$ the (unique) maximizer in the Legendre transform so that the Gibbs measure with parameters $(\alpha, h(\alpha, m))$ has mean m.

1. *Compute* $\left.\dfrac{\partial \Gamma_\alpha(m)}{\partial \alpha}\right|_{\alpha=0}$.
2. *Compute* $\left.\dfrac{\partial^2 \Gamma_\alpha(m)}{\partial \alpha^2}\right|_{\alpha=0}$.

Solution 4.2. *We will use the relation*

$$\Gamma_\alpha(m) = h(\alpha, m)^\top m - \mathcal{F}(\alpha, h(\alpha, m)),$$

and the fact that $h(\alpha, m)$ is chosen so that $\partial_h \mathcal{F}(\alpha, h(\alpha, m)) = m$. Throughout let $\langle\cdot\rangle_{\alpha,h}$ denote expectation under $\mathbb{P}_{\alpha,h}(x) \propto \exp\big(\frac{\alpha}{2} x^\top J x + h^\top x\big)$.

Note that

$$\frac{\partial \mathcal{F}}{\partial \alpha}(\alpha, h) = \frac{1}{2}\big\langle x^\top J x \big\rangle_{\alpha,h}, \qquad \frac{\partial \mathcal{F}}{\partial h_i}(\alpha, h) = \langle x_i \rangle_{\alpha,h}.$$

Also, for any observable $O(x)$,

$$\frac{\partial}{\partial \alpha}\langle O\rangle_{\alpha,h} = \mathrm{Cov}_{\alpha,h}\Big(O, \frac{1}{2}x^\top J x\Big) \quad \textit{and} \quad \frac{\partial}{\partial h_i}\langle O\rangle_{\alpha,h} = \mathrm{Cov}_{\alpha,h}(O, x_i).$$

Differentiate $\Gamma_\alpha(m) = h^\top m - \mathcal{F}(\alpha, h)$ at fixed m.

$$\frac{\partial \Gamma_\alpha(m)}{\partial \alpha} = \frac{\partial h^\top}{\partial \alpha} m - \frac{\partial \mathcal{F}}{\partial \alpha} - \big(\nabla_h \mathcal{F}\big)^\top \frac{\partial h}{\partial \alpha}.$$

But $\nabla_h \mathcal{F}(\alpha, h) = \langle x \rangle_{\alpha,h} = m$ for $h = h(\alpha, m)$, hence the terms involving $\partial h/\partial \alpha$ cancel, giving the simple identity (valid at the maximizer $h(\alpha, m)$):

$$\boxed{\frac{\partial \Gamma_\alpha(m)}{\partial \alpha} = -\frac{\partial \mathcal{F}}{\partial \alpha}\big(\alpha, h(\alpha, m)\big) = -\frac{1}{2}\big\langle x^\top J x \big\rangle_{\alpha, h(\alpha,m)}.}$$

Evaluate at $\alpha = 0$. At $\alpha = 0$ the spins are independent under $\mathbb{P}_{0,h(0,m)}$ and $h(0, m)$ is determined by $\mathrm{arctanh}(m_i) = h_i(0, m)$. *Independence implies $\langle x_i x_j \rangle_{0,h} = m_i m_j$ for $i \neq j$, while $\langle x_i^2 \rangle = 1$. Hence*

$$\big\langle x^\top J x \big\rangle_{0,h(0,m)} = \sum_{i,j} J_{ij} \langle x_i x_j \rangle_{0,h} = \sum_{i \neq j} J_{ij} m_i m_j + \sum_i J_{ii} \cdot 1.$$

If we assume J has zero diagonal, we arive at the desired term. Now differentiate the identity

$$\frac{\partial \Gamma_\alpha}{\partial \alpha} = -\frac{1}{2}\big\langle x^\top J x \big\rangle_{\alpha, h(\alpha,m)}$$

once more w.r.t. α, remembering that h depends on α (but m is held fixed). Using the covariance identity,

$$\frac{\partial}{\partial \alpha}\big\langle x^\top J x \big\rangle_{\alpha,h} = \frac{1}{2}\,\mathrm{Var}_{\alpha,h}(x^\top J x) \;+\; \sum_i \frac{\partial h_i}{\partial \alpha}\,\mathrm{Cov}_{\alpha,h}(x^\top J x, x_i).$$

Thus

$$\frac{\partial^2 \Gamma_\alpha}{\partial \alpha^2} = -\frac{1}{4}\,\mathrm{Var}_{\alpha,h}(x^\top J x) - \frac{1}{2}\sum_i \frac{\partial h_i}{\partial \alpha}\,\mathrm{Cov}_{\alpha,h}(x^\top J x, x_i).$$

We now evaluate at $\alpha = 0$. At $\alpha = 0$ the measure factorizes, which greatly simplifies covariances. A direct but elementary computation (expand variances and covariances into sums over indices and use factorization at $\alpha = 0$) yields the standard second-derivative Plefka result:

$$\boxed{\left.\frac{\partial^2 \Gamma_\alpha(m)}{\partial \alpha^2}\right|_{\alpha=0} = -\frac{1}{4}\sum_{i,j} J_{ij}^2 (1 - m_i^2)(1 - m_j^2).}$$

Specialization to $\mathbb{Z}_2$ Synchronization. For the $\mathbb{Z}_2$ synchronization model, we have $J = Y$ and $h = 0$. Since Y is drawn from a GOE matrix with variance $1/n$ off-diagonal, we have

$$\sum_{i,j} Y_{ij}^2 (1 - m_j^2)(1 - m_i^2) \approx \frac{1}{n}\sum_{i,j} (1 - m_j^2)(1 - m_i^2) = n\left(1 - \frac{\|m\|_2^2}{n}\right)^2.$$

Optimizing the approximation of Γ_λ recovers the TAP/AMP update

$$m_i^t \approx \tanh\Big(\lambda \sum_j Y_{ij} m_j^{(t-1)} - \lambda^2\Big(1 - \tfrac{\|m^{(t-1)}\|_2^2}{n}\Big) m_i^{(t-2)}\Big),$$

and the TAP free energy

$$F_{\mathrm{TAP}}(m) = \Gamma_0(m) - \frac{\lambda}{2} m^\top Y m - \frac{\lambda^2}{4}\Big(1 - \tfrac{\|m\|_2^2}{n}\Big)^2 + O(\lambda^3).$$

The Plefka expansion provides a systematic justification for the TAP correction. At infinite temperature, spins are independent. As the coupling increases, the expansion shows which fluctuation terms must be included to avoid overcounting correlations. This motivates why the TAP equations and the AMP algorithm achieve Bayes-optimal performance in spiked random matrix models, while naive mean-field fails.

4.2. State Evolution for AMP

The dynamics of the Approximate Message Passing (AMP) iteration (4.3) can be accurately tracked in the high-dimensional limit by a *scalar* recursion, the *state evolution* (SE). Below we state the SE equations for the $\mathbb{Z}_2$-synchronisation model

$$Y = \lambda \frac{xx^\top}{n} + W, \qquad x \in \{\pm 1\}^n,$$

and for the AMP iteration written in the generic form

$$u^{(t)} = Y m^{(t-1)} - b_{t-1}\, m^{(t-2)}, \qquad m^{(t)} = f(u^{(t)}),$$

where $f(\cdot) = \tanh(\cdot)$ in our previous derivation and the Onsager coefficient is $b_{t-1} = \lambda^2\big(1 - \frac{\|m^{(t-1)}\|_2^2}{n}\big)$. We will now remove the Onsager correction, but sample the noise independent in each update step. We hope that this removes the pairwise correlations. Write

$$Y m^{(t-1)} = \lambda \frac{xx^\top}{n} m^{(t-1)} + W m^{(t-1)} = \lambda\Big(\frac{x^\top m^{(t-1)}}{n}\Big) x + W m^{(t-1)}.$$

The first term is aligned with the signal x. The second term is approximately Gaussian with (coordinatewise) variance $\frac{1}{n}\|m^{(t-1)}\|_2^2$. Thus the field at coordinate i is, in the limit $n \to \infty$, approximated by

$$u_i^{(t)} \;\overset{\mathrm{d}}{\approx}\; \mu_t\, x_i \;+\; \sigma_t Z, \qquad Z \sim \mathcal{N}(0,1),$$

for scalar parameters $\mu_t, \sigma_t \geq 0$ that depend only on the previous iterates.

Define the scalar field variables (μ_t, σ_t) by the following recursion. Given (μ_t, σ_t) define the random field

$$H_t \;=\; \mu_t X + \sigma_t Z,$$

where $X \in \{\pm 1\}$ is with $\mathbb{P}(X = +1) = \mathbb{P}(X = -1) = 1/2$ and $Z \sim \mathcal{N}(0,1)$ is independent. We can derive that the state evolution updates under a uniform prior are

$$\mu_{t+1} \approx \lambda \mathbb{E}\big[\, X\, f(H_t)\,\big], \tag{4.5}$$

$$\sigma_{t+1}^2 \approx \lambda^2 \mathbb{E}\big[\, f(H_t)^2\,\big]. \tag{4.6}$$

A fixed point of SE satisfies

$$\mu = \lambda\, \mathbb{E}\big[X\, f(\mu X + \sigma Z)\big] = \mathbb{E}\big[\, f(\mu + \sigma Z)\big], \qquad \sigma^2 = \lambda^2 \mathbb{E}\big[f(\mu + \sigma Z)^2\big].$$

It is possible (see [DAM16]) to reduce the two-dimensional recursion (μ_t, σ_t) of state evolution to a *single* scalar recursion in terms of the effective signal-to-noise parameter

$$\gamma_t := \frac{\mu_t^2}{\sigma_t^2}.$$

One obtains

$$\gamma_{t+1} \;=\; \lambda^2\, \mathbb{E}\Big[\, \tanh\big(\gamma_t + \sqrt{\gamma_t}\, G\big)\,\Big], \qquad G \sim \mathcal{N}(0,1). \tag{4.7}$$

In particular:

- $\gamma = 0$ is always a fixed point, corresponding to zero overlap;
- For $\lambda > 1$ a nontrivial fixed point $\gamma^* > 0$ appears, yielding positive correlation with the signal and Bayes-optimal estimation (in the sense of asymptotic mean squared error).

Equation (4.7) is therefore a convenient tool to study thresholds and to visualize the algorithmic phase transition. The SE recursion determines whether AMP converges to the trivial fixed point $(\mu, \sigma) = (0, \lambda)$ (no correlation with the signal) or to a nontrivial fixed point with $\mu > 0$ (positive overlap). The transition is governed by the scalar recursion and depends on λ and the nonlinearity f (here tanh).

Practical exercises.

1. Implement the SE recursion numerically: at each step sample (X, Z) a large number of times (or approximate the Gaussian integrals by quadrature) and compute the expectations in (4.5)–(4.6).
2. Compare the SE prediction with empirical AMP runs on simulated spiked-Wigner matrices for several values of λ and initialisations.

5. The Semicircle Law

This section is based on [Tao12], Chapters 2.3 and 2.4. The *moment method* is a classical approach to studying the limiting spectral distribution of random matrices. For a sequence of random matrices X_n, one considers the moments of their empirical spectral measures and identifies their limiting behavior as $n \to \infty$. In particular, for *Wigner matrices*, the moment method leads to the celebrated **Wigner semicircle law**. Let $M_n = (\xi_{ij})_{1 \le i,j \le n}$ be a real symmetric random matrix with entries satisfying:

- $(\xi_{ij})_{i<j}$ are independent and identically distributed (i.i.d.) random variables with mean 0 and variance $1/n$.
- ξ_{ii} (diagonal entries) are real random variables with mean 0 and variance $2/n$ (in some conventions).

The *empirical spectral distribution* (ESD) is defined by

$$\mu_{M_n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(M_n)},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of M_n.

Remark 5.1. *Note that the ESD is a random probability measure.*

5.1. Convergence Types

We can define convergence of the empirical spectral distribution in several senses:

- **In expectation:** For every bounded continuous function f, we have
$$\mathbb{E}\left[\int f(x)\, d\mu_{M_n}(x)\right] \to \int f(x)\, d\mu_{sc}(x).$$
- **In probability:** For every bounded continuous f and every $\varepsilon > 0$,
$$\mathbb{P}\left(\left|\int f(x)\, d\mu_{M_n}(x) - \int f(x)\, d\mu_{sc}(x)\right| > \varepsilon\right) \to 0.$$
- **Almost surely:** With probability one, for all bounded continuous f,
$$\int f(x)\, d\mu_{M_n}(x) \to \int f(x)\, d\mu_{sc}(x).$$

Convergence in expectation is the weakest, while almost sure convergence is the strongest form. Since we are talking about *random* measures, we have three different notions of convergence for a distribution.

Let μ_{sc} be the measure with density

$$\mu_{sc}(x) = \frac{1}{2\pi} \sqrt{4 - x^2}\, \mathbf{1}_{|x| \le 2}.$$

Theorem 5.2 (Wigner Semicircle Law). *For the sequence of Wigner matrices (M_n) defined above, the empirical spectral distributions converge weakly, almost surely, to the semicircle distribution:*

$$\mu_{M_n} \xrightarrow{a.s.} \mu_{sc}.$$

We will first establish convergence *in expectation*. The stronger result, almost sure convergence, follows by refining the arguments using concentration inequalities and the Borel–Cantelli lemma.

5.2. The Moment Method

Moments play a central role in this approach. Under mild assumptions (e.g., subgaussian tails), convergence of all moments uniquely determines the limiting distribution by **Carleman's continuity theorem**.

In particular, the semicircle distribution satisfies Carleman's condition, ensuring that the moment sequence

$$\int x^k \, d\mu_{sc}(x)$$

uniquely identifies μ_{sc}. We will implicitly prove that the operator norm of M_n is bounded by bounding the trace moments (see Exercise 5.4). This guarantees that the subgaussian assumption is valid for our arguments. Moreover, we assume each entry satisfies $|\xi_{ij}| \leq K$. For Wigner matrices, $K = o_n(1)$ with high probability, and thus it vanishes in all our asymptotic bounds.

We study the expected k-th moment of the ESD:

$$\mathbb{E}\left[m_k(M_n)\right] = \mathbb{E}\left[\int x^k \, d\mu_{M_n}(x)\right] = \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(M_n^k)\right].$$

Exercise 5.3. *Show that for any integer $k \geq 1$ the k-th moment of the empirical spectral distribution equals the normalized trace:*

$$\int x^k \, d\mu_{M_n}(x) = \frac{1}{n}\mathrm{Tr}\big(M_n^k\big).$$

Conclude that convergence of the quantities $\frac{1}{n}\mathbb{E}[\mathrm{Tr}(M_n^k)]$ for all k is equivalent to convergence of the moments of μ_{M_n}.

Exercise 5.4. *Let M be an $n \times n$ real symmetric matrix and let $\|M\|$ denote its operator norm (largest absolute eigenvalue).*

1. *Prove that for any integer $r \geq 1$,*
$$\|M\| \leq \big(\mathrm{Tr}(M^{2r})\big)^{1/(2r)}.$$
2. *Deduce a lower bound and show that if the normalized trace moments $\frac{1}{n}\mathrm{Tr}(M^{2r})$ remain bounded (uniformly in n) for all r, then the operator norm $\|M\|$ can be controlled (uniformly in n).*

We will show that the expected trace moments of M_n correspond to the moments under μ_{sc}.

Lemma 5.5. *For k odd it holds that*

$$\int_{\mathbb{R}} x^k \, d\mu_{sc}(x) = 0$$

and for k even it holds that

$$\int_{\mathbb{R}} x^k \, d\mu_{sc}(x) = C_{\frac{k}{2}}$$

where $C_{\frac{k}{2}}$ denotes the $\frac{k}{2}$-th Catalan number.

Expanding the trace gives

$$\mathbb{E}\left[\mathrm{Tr}(M_n^k)\right] = \sum_{i_1,\ldots,i_k=1}^{n} \mathbb{E}\left[\xi_{i_1 i_2}\xi_{i_2 i_3}\cdots\xi_{i_k i_1}\right].$$

We do a warm-up by not directly computing tight bounds and bound all moments for $k \leq 6$. From this, we will see how to generalize and make the bound tighter by a counting argument later.

5.2.1. Odd Moments

It holds that

$$\mathbb{E}\left[\mathrm{Tr}(M_n)\right] = \mathbb{E}\left[\sum_{i=1}^{n} \xi_{ii}\right] = 0.$$

Moreover, convince yourself that $\mathbb{E}\left[\xi_{i_1 i_2}\xi_{i_2 i_3}\cdots\xi_{i_k i_1}\right] = 0$ whenever k is odd.

5.2.2. Even Moments

Exercise 5.6. *Show that* $\mathbb{E}\left[\mathrm{Tr}(M_n^2)\right] = O(n)$.

For the fourth moment, we make a case distinction. First, assume that all pairs $\{i_j, i_{j+1}\}$ are different. Then, assume that at least one edge $\{i_j, i_{j+1}\}$ is not repeated at least twice. Prove in both cases that the expectation of the product of the cycle is $\mathbb{E}\left[\prod \xi_{i_j,i_{j+1}}\right] = 0$. Then, it remains to count the cases where each edge is repeated at least twice and bound it.

Exercise 5.7. *Show that* $\mathbb{E}\left[\mathrm{Tr}(M_n^4)\right] = O(n + K^2 n)$.

In the case $k = 6$, we follow a similar approach

Exercise 5.8. *Show that* $\mathbb{E}\left[\mathrm{Tr}(M_n^6)\right] = O(n + K^2 n + K^4 n)$.

5.2.3. Tighter Bounds and general k even

The previous computations showed us that most cycles vanish in expectation. We only have to consider those where each edge occurs at least twice. Moreover, when $K = o_n(1)$, we saw that the terms disappear when $j < \frac{k}{2}$ edges span the cycle. We make this rigorous now for any even k:

- Suppose $1 \leq j \leq \frac{k}{2}$ distinct edges span the cycle, each occuring at least twice.
- Denote the multiplicity of each of the j edges by $a_1, \ldots, a_j \geq 2$ with $\sum_{i=1}^{j} a_i = k$.
- We can bound the number of cycles with j edges by $O(nj + 1)$ because they span at most $j+1$ vertices.
- Bound $\mathbb{E}[\xi_{i_1,i_2}\ldots\xi_{i_k,i_1}] \leq K^{a_1-2}\ldots K^{a_j-2}\frac{1}{n^j} = K^{k-2j}\frac{1}{n^j}$.
- The contribution per $1 \leq j < \frac{k}{2}$ is thus bounded by $O(nK^{k-2j}) = o(n)$ for $K = o_n(1)$.
- The number of such $1 \leq j < \frac{k}{2}$ is bounded by $\left(\frac{k}{2}\right)^k$ and independent of n.

Hence, the only non-vanishing influence comes from cycles where each edge is repeated exactly once. We call these *non-crossing cycles*. In this case it holds that

$$\mathbb{E}[\xi_{i_1,i_2}\ldots\xi_{i_k,i_1}] = \frac{1}{n^j}.$$

It remains to count the number of non-crossing cycles.

Lemma 5.9. *The number of non-crossing cycles of length k (even) in $\{1, \ldots, n\}$ is*

$$C_{\frac{k}{2}}\, n(n-1)\ldots(n-\frac{k}{2}),$$

where $C_{\frac{k}{2}} = \frac{1}{\frac{k}{2}+1}\binom{k}{\frac{k}{2}}$ is the $\frac{k}{2}$-th Catalan number.

We conclude:

$$\mathbb{E}\left[\mathrm{Tr}(M_n^k)\right] = C_{\frac{k}{2}}\, n(n-1)\ldots(n-\frac{k}{2})\frac{1}{n^{\frac{k}{2}}} = C_{\frac{k}{2}}(1+o(1))n.$$

Remark 5.10. *We can use the same analysis to prove* $\mathbb{E}\left[|\mathrm{Tr}(M_n^k)|^2\right] = n^2 C_{\frac{k}{2}}^2 + o(n^2)$. *This implies that the Variance is vanishing and we can establish convergence in probability. Moreover, by using Theorem 2.3.21 in [Tao12], we can establish almost sure convergence.*

6. Optimality and Suboptimality of PCA

This section is based on [PWBM18]. We investigate in the question whether PCA for the spiked Wigner model

$$Y = \lambda \frac{zz^\top}{n} + W$$

is an optimal algorithm for **detection**. We already have seen that for **recovery**, the algorithm is suboptimal in a Bayesian sense, but can the threshold for detection be lowered below $\lambda = 1$? This question is of statistical nature: Can we distinguish between two models with a statistical test?

Definition 6.1 (Distinguishablity). *Let (P_n, Q_n) be two sequences of probability measures on $(\Omega_n, \mathcal{F}_n)$ with $\Omega_n \subseteq \Omega$. A measurable function $f : \Omega \to \{0, 1\}$ strongly distinguishes P_n and Q_n if*

$$P_n(f^{-1}(\{1\})) \to 1, \quad n \to \infty,$$
$$Q_n(f^{-1}(\{0\})) \to 1, \quad n \to \infty.$$

P_n and Q_n are indistinguishable if no such function f exists.

It is simple to find a *sufficient* condition for distinguishability based on Le Cam's contiguity:

Definition 6.2 (Le Cam's Contiguity). *Q_n is contiguous with respect to P_n if for all sequences $(A_n)_{n\in\mathbb{N}}$ such that $A_n \in \mathcal{F}_n$*

$$\lim_{n\to\infty} P_n(A_n) \to 0 \implies \lim_{n\to\infty} Q_n(A_n) = 0.$$

Intuitively, this means that almost sure events under P_n have to occur almost surely under Q_n in the large n limit as well. Now, we can show that this implies statistical indistinguishability:

Lemma 6.3. *Assume Q_n is contiguous with respect to P_n or P_n is contiguous with respect to Q_n, then Q_n and P_n are statistically indistinguishable.*

Proof: Try to prove this by contradiction as a small exercise. ■

Now we aim to find a *sufficient* condition to show contiguity. Therefore we introduce the *Likelihood Ratio* between Q_n and P_n:

$$L_n := \frac{dP_n}{dQ_n}$$

which defines a random variable on $(\Omega_n, \mathcal{F}_n)$. This ratio helps us to show contiguity between two random variables.

Lemma 6.4 (Lemma 2.3 in [PWBM18]). *If the second moment of L_n remains bounded under Q_n, i.e.*

$$\mathbb{E}_{Q_n}\left[\left(\frac{dP_n}{dQ_n}\right)^2\right] \leq C$$

then P_n is contiguous w.r.t. Q_n.

Proof: Exercise. Hint: Use Cauchy-Schwartz.
Solution:

$$\mathbb{E}_{P_n}[A_n] = \int_{A_n} \frac{dP_n}{dQ_n}\, dQ_n \leq Q_n(A_n)^{\frac{1}{2}} \mathbb{E}_{Q_n}[L_n^2]^{\frac{1}{2}} \leq C Q_n(A_n)^{\frac{1}{2}}$$

implies that $Q_n(A_n) \to 0$ implies $P_n(A_n) \to 0$. ■

We want to derive the largest λ now for which we can show contiguity of

$$P_n : Y = \lambda z z^\top + \frac{1}{\sqrt{n}} W_n, \quad z \sim \pi_n$$
$$Q_n : Y = \frac{1}{\sqrt{n}} W_n$$

where W_n is a Gaussian Wigner matrix. The results will depend on the prior π_n. If we can show for a prior that P_n is contiguous w.r.t. Q_n for all $\lambda \leq 1$, we know that PCA is optimal for these cases!

For our results, we require π to be a *spike prior* such that its support concentrates around the sphere S^{n-1}: For every $\epsilon > 0$ it holds

$$\pi_n(1 - \epsilon \leq \|x\| \leq 1 + \epsilon) \to 0, \quad n \to \infty.$$

The BBP transition guarantees PCA to converge almost surely if $\|x\| = 1$, so for a spike prior, we get convergence in probability. Depending on the prior, we get the following bound on the likelihood ratio:

Lemma 6.5. *Let x, x' be two independent copies from π_n. Then,*

$$\mathbb{E}_{Q_n}\left[\left(\frac{dP_n}{dQ_n}\right)^2\right] = \mathbb{E}_{x,x'}\left[\exp(n\lambda^2 (x^\top x')^2)\right]$$

Proof: First, it holds that

$$dQ_n(Y) = \frac{1}{Z(n)} \exp(-\frac{n}{4} tr(Y^2))\, dY$$

for real symmetric $n \times n$-matrices Y and $Z(n)$ is the corresponding normalizing constants. Moreover, by construction it holds

$$dP_n(Y) = \frac{1}{Z(n)} \mathbb{E}_x[\exp(-\frac{tr((Y - \lambda x x^\top)^2)}{4})].$$

This implies

$$\begin{aligned}\frac{dP_n}{dQ_n}(Y) &= \mathbb{E}_x\left[\exp(-\frac{tr((Y - \lambda x x^\top)^2) - tr(Y^2)}{4})\right] \\ &= \mathbb{E}_x\left[\exp\left(-\frac{\lambda^2 \sum_{i,j} x_i^2 x_j^2 - 2\lambda \sum_{i,j} Y_{ij} x_i x_j}{4}\right)\right].\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{E}_{Q_n}\left[\left(\frac{dP_n}{dQ_n}(Y)\right)^2\right] \\ &= \mathbb{E}_{x,x'}\left[\mathbb{E}_{Q_n}\left[\exp\left(\frac{\lambda n}{2}\langle Y, xx^\top + x'(x')^\top\rangle - \frac{n\lambda^2}{4}\left(\langle xx^\top, xx^\top\rangle + \langle x'(x')^\top, x'(x')^\top\rangle\right)\right)\right]\right] \\ &= \mathbb{E}_{x,x'}\left[\exp\left(\frac{\lambda^2 n}{4}\|xx^\top + x'(x')^\top\|^2 - \frac{n\lambda^2}{4}\left(\langle xx^\top, xx^\top\rangle + \langle x'(x')^\top, x'(x')^\top\rangle\right)\right)\right] \\ &= \mathbb{E}_{x,x'}\left[\exp\left(\frac{\lambda^2 n}{2}\langle x, x'\rangle^2\right)\right]\end{aligned}$$

where the second step comes from the moment generating function of Y_{ij} and the independent entry structure (up to symmetry). ∎

Now, we want to know in which cases this expression is bounded. Based on the previous chapters, we can directly relate this to Subgaussian priors.

Lemma 6.6. *Assume $x^\top x'$ is a $\frac{\sigma^2}{n}$-Subgaussian random variable. Then, $\mathbb{E}_{Q_n}\left[\left(\frac{dP_n}{dQ_n}\right)^2\right]$ remains bounded if $\lambda < \frac{1}{\sigma}$.*

Proof: It holds that

$$\begin{aligned}
\mathbb{E}_{x,x'}\left[\exp\left(\frac{\lambda^2 n}{2}\langle x, x'\rangle^2\right)\right] &= \int_0^\infty \mathbb{P}\left(\exp\left(\frac{\lambda^2 n}{2}\langle x, x'\rangle^2\right) \geq t\right) dt \\
&= 1 + \int_1^\infty \mathbb{P}\left(|\langle x, x'\rangle| \geq \sqrt{\frac{2\log(t)}{n\lambda^2}}\right) dt \\
&\leq 1 + \int_1^\infty 2\exp\left(-\frac{2\log(t)}{n\lambda^2}\frac{n}{2\sigma^2}\right) dt \\
&= 1 + \int_1^\infty 2t^{\frac{-1}{\lambda^2\sigma^2}} dt
\end{aligned}$$

which is finite if and only if $\frac{1}{\lambda^2\sigma^2} > 1$, i.e. $\lambda < \frac{1}{\sigma}$. ■

Remark 6.7. *We can derive that $x^\top x'$ is $\frac{\sigma^2}{n}$-Subgaussian if π_n is only supported on the sphere and is $\frac{\sigma^2}{n}$-Subgaussian itself. We say that a $\mathbb{R}^n$-valued distribution π_n is $\frac{\sigma^2}{n}$-Subgaussian if for every $v \in \mathbb{R}^n$, $X \sim \pi_n$*

$$\mathbb{E}[\exp(v^\top X)] \leq \exp(\frac{\sigma^2\|v\|^2}{2}).$$

σ^2-Subgaussian i.i.d. priors can be transformed into $\frac{\sigma^2}{n}$-Subgaussian priors on $\mathbb{R}^n$ ($X_i \sim \frac{1}{\sqrt{n}}\mu$ i.i.d.). Hence, the uniform prior on the sphere which is 1-Subgaussian provides a contiguous model P_n to Q_n whenever $\lambda < \frac{1}{\sigma}$. By a conditioning argument, it is actually sufficient to be $\frac{\sigma^2}{n}$-Subgaussian also for every spike prior π_n, not only supported on the sphere, e.g. the Gaussian i.i.d. prior.

Are there priors that provide more information than uniform on the sphere or a Gaussian prior such that PCA is suboptimal and the recovery threshold is smaller than 1? We will try to answer this question by studying a sparse prior. Note that an unbounded likelihood ratio does not necessarily imply that two models are distinguishable! We will do the following analysis to get a first insight that $\lambda = 1$ *might* not be optimal for every prior, but we will also see that the likelihood method does not give sharp estimates for this. Consider the following i.i.d. prior for each coordinate: Let $\rho \in (0, 1)$ and

$$X_i = \begin{cases} 0, & \text{with probability } 1-\rho, \\ \frac{1}{\sqrt{\rho}}, & \text{with probability } \frac{\rho}{2}, \\ -\frac{1}{\sqrt{\rho}}, & \text{with probability } \frac{\rho}{2}. \end{cases}$$

Exercise 6.8. *Show that this sparse Rademacher prior is a spike prior.*

What is the Subgaussian parameter for this prior?

Lemma 6.9. *The sparse Rademacher i.i.d. prior with parameter $\rho \in (0, 1)$ is σ-Subgaussian with*

$$\sigma^2 = \sup_{t\in\mathbb{R}} \frac{2}{t^2}\log(1 - \rho + \rho\cosh(\frac{t}{\sqrt{\rho}})).$$

For $\rho \geq \frac{1}{3}$, it holds that $\sigma = 1$.

Proof: It holds that

$$\mathbb{E}\left[\exp(tX)\right] = \left(1 - \rho + \rho\cosh(\frac{t}{\sqrt{\rho}})\right)^n.$$

Hence, the entrywise prior is σ^2-Subgaussian if

$$F_\rho(t) = \log\left(1 - \rho + \rho\cosh(\frac{t}{\sqrt{\rho}})\right) \leq \sigma^2\frac{t^2}{2}$$

for all $t \in \mathbb{R}$. It holds that $F_\rho(0) = 0$, $F_\rho'(0) = 0$, $F_\rho''(0) = 1$, $F_\rho^{(3)}(0) = 0$, $F_\rho^{(4)}(0) = 3 - \frac{1}{3}$. Therefore, for $\rho < \frac{1}{3}$, it holds that the smallest σ to satisfy $F_\rho(t) \leq \sigma^2\frac{t^2}{2}$ for all $t \geq 0$ needs to be smaller than one. For $\rho \geq \frac{1}{3}$, it actually holds true that

$$\sup_{t\in\mathbb{R}} F_\rho''(t) = F_\rho''(0) = 1.$$

Therefore, $F_\rho(t) \leq \frac{t^2}{2}$ for all $t \in \mathbb{R}$. ■

Remark 6.10. *We know that PCA is optimal when $\rho \geq \frac{1}{3}$. What do we know for $\rho < \frac{1}{3}$? In the same paper [PWBM18] show that conitguity still holds for $\rho^* \geq 0.184$ by a conditioning method. This also shows that $\sigma \geq 1$ is not a necessary condition for contiguity here! The optimal threshold for which detection is possible for $\lambda < 1$ has been proven to be $\rho \approx 0.09$, although no polynomial time algorithm (including AMP) is known to solve this - a classical example of a* ***statistical-computational-gap****. This holds for any constant ρ. If ρ depends on n (decreases in n), we get extra information on the sparsity. If it is much smaller than constant, then there arises another gap: For ρ small enough there exists an algorithm that succeeds (diagonal thresholding), whereas for ρ large enough, it is conjectured that no polynomial time algorithm succeeds for $\lambda < 1$.*

7. Random Graphs and Finding Cliques

In this section we will deal with the following graphs: Given a vertex set V with $|V| = n$, connect two vertices and add it to the set of edges E with probability p independently. This is one version of the so called **Erdős–Rényi (ER) model** and we write it as $\mathcal{G}(n,p)$. A clique of size k is a complete subgraph of G, i.e. a subset of size k of V such that all vertices are connected.

Proposition 7.1. *The expected number of cliques of size k is $\mathbb{E}[X_k] = \binom{n}{k} p^{\binom{k}{2}}$.*

The intuition of this is that we count the number of possible cliques of size k and compute the probability for each clique to appear in the model.

Proof:

$$\mathbb{E}[X_k] = \mathbb{E}[\sum_{C:|C|=k} \mathbb{1}_{C\in G}] = \sum_{C:|C|=k} \mathbb{E}[\mathbb{1}_{C\in G}]$$

Therefore, we only have to compute the probabilities of each event and count the number (note that the events are not independent in general!!) The number of cliques of size k is given by $\binom{n}{k}$. The probability that one clique is contained in a random graph is the product of the probabilities that each edge is included. In a clique of size k there exist $\binom{k}{2}$ edges. Therefore,

$$\mathbb{E}[\mathbb{1}_{C\in G}] = \mathbb{P}(C \in G) = p^{\binom{k}{2}}.$$

Combining everything yields the result. ∎

We can see that for $k \gg \log(n)$, the expected value goes to 0 and therefore, a large clique should be unlikely. With the use of Markov's inequality, we can make this intuition rigorous and state the following for the ER model with $p = \frac{1}{2}$ for simplicity.

Corollary 7.2. *For $k = 2(1 + \frac{2}{\log_2(n)}) \log_2(n)$ and $G \sim \mathcal{G}(n, \frac{1}{2})$, it holds that $\mathbb{P}(X_k > 0) \to 0$ as $n \to \infty$.*

Proof: We use a bound on $\binom{n}{k}$:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n}{k}\frac{n-1}{k-1}\frac{n-2}{k-2}\cdots\frac{n-k+2}{2}\frac{n-k+1}{1} \leq n^k.$$

Therefore,

$$\mathbb{E}[X_k] \leq n^k (\frac{1}{2})^{\binom{k}{2}} = 2^{k\log_2(n)} 2^{-\frac{k(k-1)}{2}} = 2^{\frac{1}{2}k(2\log_2(n)-(k-1))}.$$

Hence, choosing $k = 2(1+\epsilon_n)\log_2(n)$, it holds

$$\mathbb{E}[X_k] \leq 2^{(log_2(n)+\epsilon_n)(\epsilon_n \log_2(n)+1)} = n^{-\epsilon_n \log_2(n)+1} 2^{\epsilon_n(\epsilon_n \log_2(n)+1))}$$

Hence, for $\epsilon_n = \frac{2}{\log(n)}$, $\mathbb{E}[X_k] \leq O(\frac{1}{n})$. Now, we can estimate the probability $\mathbb{P}(X_k > 0)$ with Markov's Inequality:

$$\mathbb{P}(X_k > 0) = \mathbb{P}(X_k \geq 1) \leq \mathbb{E}[X_k] \leq O(\frac{1}{n}) \to 0, \quad n \to \infty.$$ ∎

This proof strategy was based on a *first moment method*. In order to state a lower bound on the largest clique, we need a bound on the second moments as well. A combinatorial argument (see [AS16c], Theorem 4,5) gives us the following result:

Proposition 7.3. *Let $k(n) = 2\log_2(n)$ and $\varepsilon > 0$. It holds that*

$$\lim_{n\to\infty} \frac{\mathbb{E}[X^2_{(1-\varepsilon)k(n)}]}{\mathbb{E}[X_{(1-\varepsilon)k(n)}]^2} = 1.$$

To conclude that this implies a lower bound on the size of the largest clique that matches the upper bound, is left as a simple exercise.

Exercise 7.4. *Apply the Payley-Zygmund inequality to prove that with high probability a clique of size* $2(1+\varepsilon)\log_2(n)$ *exists.*

Now that we proved that the largest clique is of size $2\log_2(n)$, one important question is how to find large cliques in reasonable time.

Question 7.5. *Can you come up with an algorithm that can find cliques of size* $\approx \log(n)$ *with high probability?*

The very simple greedy algorithm finds cliques of size $\approx \log(n)$ with high probability:

Proposition 7.6. *With high probability, the greedy algorithm finds cliques of size* $\log(n) - \log(\log(n)) \leq k \leq \log(n) + \log(\log(n))$ *in the limit* $n \to \infty$ *for* $G \sim \mathcal{G}(n, \frac{1}{2})$.

Proof: Let us first show that $k \leq \log(n) + \log(\log(n))$ by showing that the algorithm has terminated after $\log(n) + \log(\log(n))$ steps with high probability. Assume the algorithm has run for k steps and produces the clique C_k. Suppose that N_k vertices have edges to all vertices in C_k, but have not been added yet. Then, in step $k+1$, one of these vertices will be added and the expectation of the number of vertices N_{k+1} given the value $N_k \geq 1$ is

$$\mathbb{E}[N_{k+1} \mid N_k] = \frac{N_k - 1}{2} \leq \frac{N_k}{2}.$$

If $N_k = 0$, the algorithm has terminated. We can now bound $\mathbb{E}[N_k]$ iteratively. For $k = 0$, the set of possible vertices is n. Hence,

$$\mathbb{E}[N_{k+1}] \leq \mathbb{E}[\frac{N_k}{2}] \leq \cdots \leq \frac{1}{2^k}\mathbb{E}[N_0] = \frac{n}{2^k}.$$

Therefore, for $k = \log_2(n) + \log_2(\log_2(n))$, it holds:

$$\mathbb{E}[N_k] \leq \frac{1}{\log_2(n)}$$

and again,

$$\mathbb{P}(N_k \geq 1) \leq \frac{1}{\log_2(n)}.$$

Hence, the algorithm has terminated with high probability. Let us now show the lower bound. At step k assume that the algorithm has terminated. The clique C that has been output is of size at most k. The remaining vertices do not have an edge to every vertex in C. The probability that one vertex does not have an edge to all vertices in C is at most $1 - \frac{1}{2^k}$. Since the probabilities are independent for the $\geq n - k$ vertices, the probability of terimating before k is bounded by

$$(1 - \frac{1}{2^k})^{n-k} \leq (\exp(-\frac{1}{2^k}))^{n-k}.$$

For $k \geq \log_2(n) - \log_2(\log_2(n))$, the probability of terminating is bounded by

$$(\exp(-\frac{1}{2^k}))^{n-k} \leq \exp(-2^{-k}\frac{n}{2}) = \exp(-2^{(-\log_2(n)+\log_2(\log_2(n)))\frac{n}{2}}) = \exp(-\frac{\log(n)}{2}) = \frac{1}{\sqrt{n}} \to 0, \quad n \to \infty.$$

■

Exercise 7.7. *What is the expected runtime of the greedy search algorithm until it terminates?*

We showed that one specific algorithm that terminates in polynomial time can find a clique of roughly half the size of the largest clique. An exhaustive search over all cliques would find the largest clique, but take longer than polynomial time. So, we will deal with the question whether there exists an *efficient* (polynomial time) algorithm that finds a larger clique than of size $\log_2(n)$. The failure of greedy search can have two reasons: On the one hand, it is very rigid: Once we add a *bad* vertex, we can not remove it to move to a *better* clique. On the other hand, there exist many more cliques of size $\log_2(n)$ than of $2\log_2(n)$. So, would an algorithm that allows to remove vertices actually find the larger clique in polynomial time or would it be stuck in the set of the many more smaller clique for a long time? We try to answer this question by studying Markov Chain Monte Carlo algorithms that aim to sample from the following distribution:

Definition 7.8. *Let $G = (V, E)$ be a graph and Ω_k the set of cliques of size $k \leq n$ that are contained in G. The distribution μ_λ on the set of all cliques contained in G, i.e. $\bigcup_{k=0}^{n} \Omega_k$ is given by*

$$\mu_\lambda(K) \propto \lambda^{|K|}.$$

Remark 7.9. *For $\lambda > 1$, μ_λ favors larger cliques over smaller ones. A sample from this distribution can therefore be expected to be useful for λ large enough. For constant λ, there is still an energy-entropy trade-off that favors smaller cliques, but when we increase λ, this effect should vanish. Indeed, for $\lambda = n$, only the largest cliques should dominate.*

The Idea of MCMC The main idea behind Markov Chain Monte Carlo (MCMC) methods is to generate samples from a target distribution π by constructing a Markov chain whose stationary distribution is precisely π. If we let P denote the transition kernel of the chain and μ_0 its initial distribution, then as $t \to \infty$, the distribution of the chain $\mu_0 P^t$ converges to π (under mild assumptions on P). In practice, we hope that for sufficiently large, but finite, t, the distribution $\mu_0 P^t$ is already close enough to π, or at least shares many of its important properties, so that the generated samples can be treated as approximately drawn from the target distribution.

Jerrum [Jer92] proved more than thirty years ago, that a simple Markov Chain with stationary distribution μ_λ will take longer than polynomial time to reach a clique of size $\geq (1+\varepsilon)\log_2(n)$ from a worst-case initialization independent of the choice of λ and hence, does not improve over the rigid greedy search algorithm. Define the *Metropolis Process* on the state space of cliques for $\lambda > 1$ by the following transition kernel:

$$P_\lambda(K, K') = \begin{cases} \frac{1}{n}, & \text{if } K \subset K', |K' \triangle K| = 1, \\ \frac{1}{\lambda n}, & \text{if } K' \subset K, |K' \triangle K| = 1, \\ 0, & \text{if } |K' \triangle K| \geq 2, \end{cases}, \qquad P_\lambda(K, K) = 1 - \sum_{K \neq K'} P_\lambda(K, K').$$

Exercise 7.10. *Check that P_λ indeed defines a transition kernel for a Markov Chain.*

In every step of the Markov Chain, one vertex is either added (with probability $\frac{1}{n}$), removed (with probability $\frac{1}{\lambda n}$), or it stays in K. This Markov Chain satisfies the detailed balance equations for $\pi = \mu_\lambda$:

$$\pi(K) P_\lambda(K, K') = \pi(K') P_\lambda(K', K).$$

Hence, P_λ is reversible and has stationary distribution μ_λ. In particular, in the $t \to \infty$ limit, the Markov Chain will converge to μ_λ, independent of the starting state. We are interested in the convergence rate and in particular, if the distribution is mostly supported on large cliques if $t = t(n)$ is chosen polynomially in n. Jerrum [Jer92] proved that this is not the case, independent of the choice of λ. To prove this, we introduce the notion of *m-gateways* which a Markov Chain has to pass in order to reach a large clique of size m.

Definition 7.11. *For $n \geq m > k$, an m-gateway of size k is a clique K of size k such that there exists set of cliques $K_0, \ldots, K_s$ with*

- $K_0 = K$,
- $|K_i \triangle K_{i+1}| = 1$, *for* $i = 0, \dots, s-1$
- $\{K_i\}_{i \geq 1}$ *does not pass a clique of size* k *anymore, i.e.* $|K_i| > k$ *for* $1 \leq i \leq s$,
- $|K_s| = m$.

We will see that the gateways are very rare among all cliques and therefore, intuitively, it will take very long to finally pass one of the gateways. The following Lemma proved by Jerrum [Jer92] states precisely that for a choice of $k \leq m = \lceil (1+\varepsilon)\log_2(n) \rceil$, the proportion of m-gateways of size k is less than polynomially small which make the m-cliques inaccesible for a large number of paths that the Markov Chain would choose.

Lemma 7.12 (Lemma 1 in [Jer92]). *Let* $\varepsilon > 0$, $k = \lceil (1+\frac{2}{3}\varepsilon)\log_2(n) \rceil$ *and* $m = \lceil (1+\varepsilon)\log_2(n) \rceil$. *Denote the set of* m*-gateways by* $C_{k,m}$. *Then, with high probability over the ER graph* $G(n, \frac{1}{2})$, *it holds*

$$\frac{|C_{k,m}|}{|\Omega_k|} \leq n^{-\Omega(\log(n))}. \tag{7.1}$$

Fix a graph G now that satisfies (7.1). With this Lemma we can show that the Markov Chain will not pass an m-gateway with high probability (over the Markov Chain) in $n^{\Omega(\log(n)}$ time:

Theorem 7.13 (Theorem 2 in [Jer92]). *Let* $\varepsilon > 0$, $k = \lceil (1+\frac{2}{3}\varepsilon)\log_2(n) \rceil$, $m = \lceil (1+\varepsilon)\log_2(n) \rceil$ *and* $\lambda > 1$. *There exists an initial state* K *with* $|K| \leq k$ *such that the expected time to reach a clique of size* m *exceeds* $n^{\Omega(\log(n)}$.

Proof: We define a partition through $C_{k,m}$ such that $\Omega_k \subseteq S_1$: Let S_1 denote the set of all cliques that are reachable from the empty set *without* passing **through** an m-gateway of size k (i.e. $C_{k,m} \subseteq S_1$). S_2 is then defined as $\Omega \backslash S_1$. In particular, all m-cliques lie in S_2. We will see that this partition creates a bottleneck between the two sets. Define

$$\Phi_S = \frac{\sum_{K \in S_1, K' \in S_2} \pi(K) P_\lambda(K, K')}{\sum_{K \in S_1} \pi(K)},$$

the probability of passing from S_1 to S_2 when starting in S_1. By definition it holds that $p(K, K') = 0$ if $K \notin C_{k,m}$. Therefore,

$$\phi_S \leq \frac{\pi(C_{k,m})}{\pi(\Omega_k)} = \frac{\lambda^{|K|}|C_{k,m}|}{\lambda^{|K|}|\Omega_k|} \leq n^{-\Omega(\log(n))}.$$

Why does this imply that the Markov Chain does not reach an m-clique in polynomial time with high probability when initialized from a specific state? Choose a starting state from S_1 with distribution $\mu_\lambda(\cdot \mid S_1)$ and stop the Markov Chain when it transitioned to S_2. The distribution after one step is the mixture of $\pi(\cdot \mid S_1)$ and the probability of having left after one step. The distribution after t steps can therefore be defined iteratively like this. Then the probability of leaving S_1 can only decrease over time and therefore is bounded by Φ_S. Define the hitting time T_{S_2} by the time the Markov chain moves to S_2 (where it terminates). By the argument before, it holds that $\mathbb{P}(T_S = t) \leq \phi(S)$ for every $t \geq 1$. For $M := \lceil \frac{1}{2\phi(S)} \rceil$, we have:

$$\begin{aligned}
\mathbb{E}[T_S] &= \sum_{t \geq 1} \mathbb{P}(T_S = t) t \\
&\geq \sum_{t=1}^{M} \phi_S t + M(1 - \phi_S M) \\
&\geq \phi_S \frac{M(M+1)}{2} + M(1 - \phi_S M) \\
&\geq \frac{1}{2\phi(S)}.
\end{aligned}$$

Hence, there exists a starting state in $K \in S_1$ such that $\mathbb{E}[T_S \mid X_0 = K] \geq n^{\Omega(\log(n))}$. ∎

7.1. Planting a Large Clique

Jerrum [Jer92] further posed the question of whether efficient algorithms can detect larger planted cliques in this model. Consider the following graph distribution over planted cliques of size k: Choose $K \subseteq V := [n]$ with $|K| = k$ uniformly at random and select edges (i, j) with probability 1, if $i, j \in K$ and with probability $\frac{1}{2}$ else. The first question in this model is about the number of large cliques in this model, i.e. for which k the large clique is unique. This is equivalent to the following question:

Question 7.14. *For which k is it information-theoretically impossible and for which k is it possible to recover the planted clique K?*

To answer this question, we have to consider the distribution over cliques of size k in the planted model to see whether the planted clique is the only one of this size! We will see a proof of this later for $k^* \leq (2+\varepsilon)\log_2(n)$.

Let us start to bound k^* by identifying algorithms that succeed in either recovering or in distinguishing.

7.1.1. Degree Test

The degree test is based on the idea that vertices in the planted clique should have higher degrees than all other vertices. The algorithm simply identifies the k vertices with the highest degree and outputs this clique $\hat{K}$ as a guess for the planted clique. We will show by simple concentration inequalities that this algorithm succeeds with high probability when $k \geq C\sqrt{n\log(n)}$.

Lemma 7.15. *The degree test satisfies*

$$\mathbb{P}(\hat{K} = K) \geq 1 - 2n\exp\left(-\frac{(k-1)^2}{8(n-1)}\right).$$

Therefore, it succeeds with high probability if $k \geq C\sqrt{n\log(n)}$.

Proof: Let $d(v)$ denote the degree of a vertex v.

First consider a vertex $v \in K$. Its degree can be written as

$$d(v) = (k-1) + X_v, \quad \text{where } X_v \sim \mathrm{Bin}(n-k, 1/2).$$

Hence

$$\mathbb{E}[d(v)] = k - 1 + \frac{n-k}{2} = \frac{n-1}{2} + \frac{k-1}{2}.$$

Now consider a vertex $u \notin K$. Its degree is

$$d(u) \sim \mathrm{Bin}(n-1, 1/2),$$

so

$$\mathbb{E}[d(u)] = \frac{n-1}{2}.$$

Set the threshold $T := \frac{n-1}{2} + \frac{k-1}{4}$ such that for $v \in K$,

$$\mathbb{E}[d(v)] - T = \frac{k-1}{2} - \frac{k-1}{4} = \frac{k-1}{4}.$$

Then,

$$\mathbb{P}(d(v) \leq T) = \mathbb{P}\big(X_v \leq \mathbb{E}[X_v] - (\mathbb{E}[d(v)] - T)\big) \leq \mathbb{P}\big(X_v \leq \mathbb{E}[X_v] - \frac{k-1}{4}\big).$$

By a standard Chernoff bound for binomial variables,

$$\mathbb{P}(X_v \leq \mathbb{E}[X_v] - t) \leq 2\exp\left(-\frac{2t^2}{n-k}\right).$$

Taking $t = \frac{k-1}{4}$, we obtain

$$\mathbb{P}(d(v) \leq T) \leq 2\exp\left(-\frac{(k-1)^2}{8(n-k)}\right).$$

Now, for $u \notin K$,

$$\mathbb{P}(d(u) \geq T) = \mathbb{P}\left(d(u) \geq \mathbb{E}[d(u)] + \frac{k-1}{4}\right).$$

Applying the Chernoff bound,

$$\mathbb{P}(d(u) \geq T) \leq 2\exp\left(-\frac{(k-1)^2}{8(n-1)}\right)$$

Let $\mathcal{E}$ be the event that all vertices in K have degree at least T and all vertices outside K have degree less than T. On $\mathcal{E}$, the k largest-degree vertices are exactly those in K, hence $\hat{K} = K$.

By the union bound,

$$\mathbb{P}(\mathcal{E}^c) \leq 2k\exp\left(-\frac{(k-1)^2}{8(n-k)}\right) + 2(n-k)\exp\left(-\frac{(k-1)^2}{8(n-1)}\right) \leq 2n\exp\left(-\frac{(k-1)^2}{8(n-1)}\right)$$

where we used that the second term dominates and adjusted constants.

Therefore,

$$\mathbb{P}(\hat{K} = K) \geq 1 - 2n\exp\left(-\frac{(k-1)^2}{8(n-1)}\right).$$

■

Remark 7.16. *Dekel et al. [DGGP14] extended the simple degree test to an iterative test that succeeds for $k \geq C\sqrt{n}$.*

7.1.2. Spectral Method

We will show now that we can distinguish the planted clique graph from $G(n, \frac{1}{2})$ for $k \geq C\sqrt{n}$ by the spectral method as well. The recovery problem has also been solved in this regime with spectral methods by Alon et al. [AKS98]. The idea is that for large enough k, the spectrum of the random adjacency matrix should have a spike compared to the spectrum of the simple Erdos-Renyi graph. Let

$$B := 2A - \mathbf{1}\mathbf{1}^\top.$$

We have the following result:

Proposition 7.17. *With high probability over $G(n, \frac{1}{2})$, it holds that*

$$\|B(G)\|_{op} \leq c\sqrt{n}.$$

In the planted clique model $G(n, \frac{1}{2}, k)$ with clique of size k it holds

$$\|B(G)\|_{op} \geq k.$$

Proof: Sketch: Recall the moment method for the first part. In particular, for bounded entries ($|B_{ij}| = 1$), and even integer r it holds:

$$\mathbb{E}[tr(M^r)] \leq (2 + o(1))^r n^{\frac{r}{2}+1}.$$

Now use Markov's Inequality and let $r = C\log(n)$ to deduce the result. For the second part, find a vector v with $v^\top Bv = k\|v\|_2^2$. ■

Exercise 7.18. *Complete the proof.*

To solve the recovery problem, Alon et al. [AKS98] prove that there exists an algorithm that extracts vertices from the second largest eigenvector of A and recovers the clique by ordering them to a base set and assigning them based on the number of neighbors in the base set.

7.1.3. MCMC

The natural question is now whether we can do better than n^β, $\beta = \frac{1}{2}$. This question has already been addressed by Jerrum [Jer92] where he found a negative answer in the case of the MCMC algorithm that we studied in the non-planted model before. First of all, the Theorem was stated in terms of another model instead of $G(n, \frac{1}{2}, k)$: Define

$$G'(n, p, k) := \{G = (V, E) : G \text{ contains a clique of size } k\}$$

and equip it with the uniform distribution over all such graphs. We will now see that the k-clique is unique when $k \geq C \log_2(n)$. In this case, we can prove the failure of MCMC using the model $G(n, \frac{1}{2}, k)$ and the results directly translate to $G'(n, \frac{1}{2}, k)$.

Lemma 7.19. *The expected number of cliques of size k is bounded above by*

$$\sum_{t=0}^{k} (nk2^{-\frac{1}{2}(k-1)})^t.$$

Jerrum [Jer92] shows in Lemma 3, that m-gateways of the same size as in Lemma 7.12 are unlikely to intersect with the largest clique on a large subset when $k \leq n^\beta$ and $\beta < \frac{1}{2}$. Hence, the ratio of cliques which are m-gateways is still of size $n^{-\Omega(\log(n))}$. The same argument as before proves that MCMC does not mix in polynomial time.

8. Dvoretzky's Theorem

In the lecture you studied dimension-reduction techniques with the Johnson-Lindenstrauss Lemma. It essentially states that the geometric structure of a set X is (approximately) preserved under random projections into lower-dimensional spaces as long as its dimension is sufficiently large (logarithmically in the size of X). But what happens when we project to very low dimensions — far below what JL requires? Rather than preserving structure, something surprising occurs: *regularity is created*. No matter how irregular a convex body may look, its low-dimensional sections will appear approximately spherical with high probability. We will study this phenomenon in the sequel.

8.1. The Theorem

Dvoretzky's Theorem essentially states that every high-dimensional normed vector space has a low-dimensional subspace which is almost Euclidean. It was first formulated by Dvoretzky [Dvo64] and later different proof strategies emerged to obtain better bounds for the dimensions for which the Theorem holds. We will prove Milman's version of the Dvoretzky Theorem which is a result from concentration of measure on the sphere and is based on a Chaining argument. Lets us first introduce some notation: The unique rotation-invariant measure on the sphere is denoted by σ. The set of orthogonal projection matrices on $\mathbb{R}^d$ is denoted by $O(d)$ and its unique rotation-invariant measure by ν.

Theorem 8.1 (Theorem 5.2.10 in [AAGM15]). *Let $X = \mathbb{R}^d$ and $\|\cdot\|$ be a norm in $\mathbb{R}^d$ such that $\|\cdot\| \leq L\|\cdot\|_2$ for every $x \in \mathbb{R}^d$. Define*

$$M := \int_{S^{d-1}} \|x\| \, d\sigma(x)$$

and let $\epsilon \in (0,1)$. For $k \leq c_2\epsilon^2 \frac{1}{\log(\frac{1}{\epsilon})} d\frac{M^2}{L^2}$ and a universal constant c_2 there exists a k-dimensional subspace F of X such that

$$M\frac{1}{1+\epsilon}\|x\|_2 \leq \|x\| \leq M(1+\epsilon)\|x\|_2 \tag{8.1}$$

for every $x \in F$. In particular, for any k-dimensional subspace F_0, choosing an orthogonal projection U w.r.t ν, it holds with high probability

$$M\frac{1}{1+\epsilon}\|Ux\|_2 \leq \|Ux\| \leq M(1+\epsilon)\|Ux\|_2$$

for every $x \in F_0$.

Remark 8.2.

- *The probability will be specified in Corollary 8.14 and it is upper bounded by $1 - \exp(-Cd\frac{M^2\epsilon^2}{L^2})$.*
- *We will show Milman's proof which is based on a Chaining argument. This Theorem includes an $\log(\frac{2}{\epsilon})$-dependency in the dimension bound. By using a different proof based on Gordon's Theorem, this term can be removed such that the bound on k is of order $\epsilon^2 d\frac{M^2}{L^2}$.*
- *There exists a Gaussian formulation of this problem as well where U is not a projection, but a Gaussian matrix.*

8.1.1. Examples

Consider the ℓ_1-ball on $\mathbb{R}^d$.

Exercise 8.3.

1. *Find the Lipschitz-constant L for $\|x\|_1 \leq L\|x\|_2$.*

2. *Compute M exactly. For large d, it holds $M \sim \sqrt{\frac{2d}{\pi}}$.*

Hint: You may use that $\int_{S^{d-1}} |x_1| \, d\sigma(x) = \frac{2\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d+1}{2})}$.

Let F be the k-dimensional subspace from Theorem 8.1. For all $x \in \partial B_{\|\cdot\|_1} \cap F$, we get that

$$\frac{1}{(1+\epsilon)}\sqrt{\frac{\pi}{2d}} \leq \|x\|_2 \leq (1+\epsilon)\sqrt{\frac{\pi}{2d}} \quad \forall x \in F \tag{8.2}$$

and in particular for $k \leq Cd$ and a constant $C > 0$, it holds

$$\frac{1}{2}\sqrt{\frac{\pi}{2d}} \leq \|x\|_2 \leq 2\sqrt{\frac{\pi}{2d}} \quad \forall x \in F. \tag{8.3}$$

Hence, the convex hull of $F \cap B_{\|\cdot\|_1}$ is almost a Euclidean Ball. Especially sparse vectors cannot be contained in F. Hence, random projections are likely to project sparse to dense vectors. Moreover, if F is randomly generated by UF_0, then (8.2) holds with probability at least $1 - \exp(-c_1 d\frac{\epsilon^2}{36\pi})$.

In fact, we can show that for all ℓ_p-balls the dimension k for which a random subspace is almost Euclidean with high probability is of order d.

Exercise 8.4. *Show that there exists a constant C such that for every $p \in [1, 2]$ Inequality (8.1) holds for all $k \leq Cd$ when the norm is $\|\cdot\|_p$.*

Exercise 8.5. *For $p = \infty$, argue that $M \geq C\sqrt{\frac{\log(d)}{d}}$ and $L = 1$.*

Therefore, the largest k for (8.1) to hold is of order $\log(d)$.

8.2. Proof

8.2.1. Concentration on the Sphere

Lemma 8.6 (Concentration on the Sphere). *Let $f : S^{d-1} \to \mathbb{R}$ be an L-Lipschitz function. Then there exists a universal constant c_1 such that*

$$\sigma(\{x \in S^{d-1} : |f(x) - \int_{S^{d-1}} f(x) \, d\sigma(x)| > t\}) \leq 4\exp(-c_1 dt^2/L^2)$$

Without loss of generality we will prove Theorem 8.1 for $x \in S^{d-1}$. Therefore we will show that we can find $m = \lfloor \frac{1}{4}\exp(\frac{c_1\kappa^2 d}{2}) \rfloor$ many points $x_1, \ldots, x_m$ on the sphere such that with high probability

$$\|Ux_i\| \in [M - L\kappa, M + L\kappa] \tag{8.4}$$

for all x_i and $U \sim \nu$. Hence, $\|Ux_i\|$ concentrates very much around its mean over the sphere.

The main argument here is that for any $x \in S^{d-1}$

$$\nu(\{U \in O(d) : Ux_0 \in A\}) = \sigma(A). \tag{8.5}$$

Question 8.7. *Why does Equation (8.5) hold?*

Now define

$$B_i := \{U \in O(d) : \|Ux_i\| \in [M - L\kappa, M + L\kappa]\}$$

and note that independently of the choice of x_i

$$\begin{aligned}\nu(B_i) &= \sigma(\{x \in S^{d-1} : \|x\| \in [M - L\kappa, M + L\kappa]\}) \\ &= \sigma(\{x \in S^{d-1} : |f(x) - \int_{S^{d-1}} f(x)\, d\sigma(x)| \leq L\kappa\}) \\ &\geq 1 - 4\exp(-c_1 d\kappa^2)\end{aligned}$$

where we used Lemma 8.6 in the final inequality and the fact that $\|\cdot\|$ is an L-Lipschitz function with respect to the Euclidean distance. Hence, by defining

$$B := \bigcap_{i=1}^{m} B_i$$

it holds by union bound

$$\nu(B) = 1 - \nu(\bigcup_{i=1}^{m} B_i^c) \geq 1 - 4m\exp(-c_1 d\kappa^2).$$

Therefore, by taking $m \leq \frac{1}{4}\exp(c_1\kappa^2 d/2)$, we get

$$\nu(B) \geq 1 - \exp(-c_1 d\kappa^2/2). \tag{8.6}$$

8.2.2. Chaining

In the previous part, we have shown that our claim holds for finitely many points on the sphere. We will now extend this result to a subspace of the sphere by choosing the points such that they span a net over the sphere intersected with a subspace F_0 of dimension k. We will see that this dimension cannot be too large because the number of points to span a net increases in the dimension. Moreover, we restricted the number of points by m in the previous section.

Let us first introduce what we mean by a net:

Definition 8.8. *A δ-net on a subspace F_1 of X is a set $\mathcal{N} \subset F_1$ such that for every $y \in F$ there exists $x \in \mathcal{N}$ such that $\|y - x\|_2 < \delta$.*

With this, we will be able to show that Equation (8.4) holds on the whole sphere with high probability. First, let us bound the cardinality of the δ-net over a sphere S^{k-1}:

Exercise 8.9. *Prove that there exists a δ-net $\mathcal{N}(\delta)$ over the k-dimensional sphere it holds $|\mathcal{N}(\delta)| \leq (1 + \frac{2}{\delta})^k$.*

Note that this result depends exponentially on the dimension k. Therefore, our result will only hold for subsets $F \subseteq S^{d-1}$ which can be identified with S^{k-1}, i.e. a k-dimensional subset. This leads to the following Corollary:

Corollary 8.10. *Take k such that $(1 + \frac{2}{\delta})^k \leq \lfloor \frac{1}{4}\exp(c_1 d\kappa^2/2) \rfloor$. Then, for any k-dimensional subspace F_0 of X it holds with probability at least $1 - \exp(-c_1 d\kappa^2/2)$*

$$M - L\kappa \leq \|Ux\| \leq M + L\kappa \quad \forall x \in \mathcal{N}(\delta)_{F_0}$$

when $U \sim \nu$. In this case, $\mathcal{N}(\delta)_{F_0}$ is a δ-net over $F_0 \cap S^{d-1}$.

Proof: Combine Exercise 8.9 with Equation (8.6) and note that the bound over $|\mathcal{N}(\delta)_{F_0}|$ corresponds to the bound over the k-dimensional sphere. ∎

Remark 8.11.

- *We will denote the set $F_U := UF_0$ for simplicity.*
- *It holds that $dim(F_U) = k$ for every $U \in O(d)$.*
- *Note that for every $y \in F_U$ there exists $x \in F_0$ such that $y = Ux$. If $\mathcal{N}(\delta)_{F_0} = \{x_1, \dots, x_m\}$ is a δ-net over $S^{d-1} \cap F_0$, then $\mathcal{N}(\delta)_{F_U} := \{Ux_1, \dots, Ux_m\}$ is a δ-net over $F_U \cap S^{d-1}$ for every $U \in O(d)$ because U is an orthogonal projection in $\mathbb{R}^d$.*

Now we come to the most important part of chaining where Equation (8.4) will be proven for all $F_U \cap S^{d-1}$. This can be done by identifying every $y \in F_U$ through a sequence of close points from the δ-net.

Proposition 8.12. *Let F_U be the randomly generated subspace from F_0 and $U \in O(d)$. Then, with probability at least $1 - \exp(-c_1 d\kappa^2/2)$ over $U \sim \nu$, it holds*

$$\frac{1-2\delta}{1-\delta} M - \frac{L\kappa}{1-\delta} \leq \|y\| \leq \frac{M + L\kappa}{1-\delta}$$

for every $y \in F_U \cap S^{d-1}$ if $(1 + \frac{2}{\delta})^k \leq \lfloor \frac{1}{4} \exp(c_1 d\kappa^2/2) \rfloor$.

Remark 8.13. *Note that we did not choose δ yet. If we look carefully, δ describes a trade-off: If δ is close to zero, our bounds on $\|y\|$ become tighter. On the other hand, $(1 + \frac{2}{\delta})^k$ will be very large which imposes a restriction on the dimension. Therefore the following holds: the lower the dimension, the tighter the bound on $\|y\|$.*

Proof: Let $y \in F_U \cap S^{d-1}$. Then, there exists $y_0 \in \mathcal{N}(\delta)_{F_U}$ such that

$$\|y - y_0\|_2 =: \delta_0 < \delta.$$

Moreover $\frac{y-y_0}{\delta_0} \in F_U \cap S^{d-1}$. Hence, there exists $y_1 \in \mathcal{N}(\delta)_{F_U}$ with

$$\|\frac{y - y_0}{\delta_0} - y_1\|_2 =: \delta_1 < \delta.$$

Hence,

$$\|y - y_0 - \delta_0 y_1\|_2 = \delta_0 \delta_1 < \delta^2.$$

By repeating this procedure, we find inductively $y_2, y_3, ...y_n \in \mathcal{N}(\delta)_{F_U}$ such that

$$\|y - y_0 - \sum_{i=1}^{n} (\prod_{j=0}^{i-1} \delta_j) y_i\|_2 < \delta^{n+1}.$$

Since $\delta < 1$, we get the limit

$$y = y_0 + \sum_{i=1}^{\infty} (\prod_{j=0}^{i-1} \delta_j) y_i$$

in $\|\cdot\|_2$ and therefore

$$\|y\| = \|y_0 + \sum_{i=1}^{\infty} (\prod_{j=0}^{i-1} \delta_j) y_i\| \leq \|y_0\| + \sum_{i=1}^{\infty} \delta^i \|y_i\| \leq (M + L\kappa) \sum_{i=0}^{\infty} \delta^i = \frac{M + L\kappa}{1-\delta}$$

where the last equation follows from Corollary 8.10. We can take the limit in the triangle inequality because $\|\cdot\|_2$ and $\|\cdot\|$ are equivalent on $\mathbb{R}^d$, so the limit exists in $\|\cdot\|$. We obtain a lower bound by triangle inequality:

$$\|y\| \geq \|y_0\| - \|\sum_{i=1}^{\infty} (\prod_{j=0}^{i-1} \delta_j) y_i\| \geq M - L\kappa - \sum_{i=1}^{\infty} \delta^i \|y_i\| \geq M - L\kappa - \frac{\delta}{1-\delta}(M + L\kappa) = \frac{1-2\delta}{1-\delta} M - \frac{L\kappa}{1-\delta}.$$

Hence, the statement holds for every $y \in F_U \cap S^{k-1}$ with probability at least $1 - \exp(-c_1 d\kappa^2/2)$ over U. ■

It remains to choose δ and k. This is shown via the following Corollary:

Corollary 8.14. *Let* $\epsilon \in (0,1)$ *and*

$$k \leq c_2 \epsilon^2 \frac{1}{\log(\frac{2}{\epsilon})} d(\frac{M}{L})^2$$

and F_U *be the randomly generated subspace from* F_0 *and* $U \in O(d)$*. Then, with probability at least* $1 - \exp(-c_1 d \frac{M^2 \epsilon^2}{72 L^2})$ *over* $U \sim \nu$*, it holds*

$$\frac{M}{1+\epsilon} \leq \|y\| \leq M(1+\epsilon) \quad \forall y \in F_U \cap S^{d-1}. \tag{8.7}$$

Proof: In order for Proposition 8.12 to hold, we have to choose

$$k \leq c \frac{d\kappa^2/2}{\log(1+\frac{2}{\delta})}.$$

Choose $\delta = \frac{\epsilon}{6}$ and $\kappa = \frac{M\epsilon}{6L}$. Then, it suffices to choose

$$k \leq c_2 \epsilon^2 d \frac{M^2}{L^2} \frac{1}{\log(\frac{2}{\epsilon})}$$

for a specific constant c_2. Moreover,

$$\frac{1-2\delta}{1-\delta} M - \frac{L\kappa}{1-\delta} = M \frac{1-3\delta}{1-\delta} \geq M \frac{1}{1+\epsilon}.$$

Moreover,

$$\frac{M + L\kappa}{1-\delta} = M \frac{1+\delta}{1-\delta} = M(1 + \frac{2\delta}{1-\delta}) \leq M(1+\epsilon).$$

Combine the results with Proposition 8.12 and the Corollary follows. ■

It has been shown by Milman and Schechtman (Theorem 5.3.4 in [AAGM15]) that the largest dimension for (8.7) to hold with $\epsilon = \frac{1}{2}$ and probability $1 - \frac{k}{d+k}$ is upper bounded by $Cd\frac{M^2}{L^2}$. Therefore, the bound on k is tight up to constants when $\frac{M}{L} \geq C\sqrt{\frac{\log(d)}{d}}$.

9. Randomized Methods in Linear Algebra

In this section, we will show that randomness and therefore, structure can help to speed up algorithms that we usually consider in a deterministic setting, especially in Linear Algebra. We follow two chapters from the lecture noted [KT23]. For further reading these notes are highly recommended!

9.1. Warm-Up: Randomized Power Method

The first part of this chapter is to show that randomizing algorithms can provide convergence guarantees by exploiting the structure of randomness. Consider the Power Method for identifying the leading eigenvector and eigenvalue of a symmetric positive semidefinite matrix A. The power method also exists for general matrices for identifying leading singular values and the corresponding left and right singular vectors, but we will focus on the symmetric case here. Without loss of generality assume that $\lambda_1(A) = 1$

We expect that if a vector x_0 is not orthogonal to the leading eigenvector, that multiplying A with it, will keep its correlation with the leading eigenvector, but weaken its correlation with all eigenvectors with eignevalue smaller one. By repeating this, we hope to converge to λ_1 quickly.

For $t = 0, 1, 2, \ldots$ define

$$u_t := \frac{x_t}{\|x_t\|}, \qquad \widehat{\lambda}_{t+1} := u_t^\top A u_t, \qquad x_{t+1} := A u_t.$$

Optionally stop when $\|x_{t+1} - \widehat{\lambda}_{t+1} u_t\| < \varepsilon$ or after a maximum number of iterations.

Algorithm 1 Power iteration (estimate dominant eigenpair of A)

Require: $A \in \mathbb{R}^{n \times n}$, initial $x_0 \neq 0$, tolerance $\varepsilon > 0$, max iterations T

for $t = 0, 1, \ldots, T-1$ **do**
  $u_t \leftarrow \dfrac{x_t}{\|x_t\|}$ {normalize}
  $\widehat{\lambda}_{t+1} \leftarrow u_t^\top A\, u_t$ {Rayleigh quotient}
  $x_{t+1} \leftarrow A\, u_t$
  if $\|x_{t+1} - \widehat{\lambda}_{t+1} u_t\| < \varepsilon$ **then**
    return $(\widehat{\lambda}_{t+1},\, u_t)$
  end if
end for
return $(\widehat{\lambda}_T,\, u_{T-1})$

Note that by construction it holds that

$$x_t = \frac{A^t x_0}{\|A^t x_0\|}.$$

Let $v_1, \ldots, v_n$ be an orthonormal basis for A with eigenvalues $\lambda_1 \geq \ldots \lambda_n$. Then,

$$x_0 = \sum_{i=1}^{n} \omega_i v_i$$

with $\omega_i := x_0^\top v_i$. It holds:

$$\begin{aligned}
1 - \hat{\lambda}_t &= 1 - \frac{x_0^\top A^t A A^t x_0}{x_0^\top A^t A^t x_0} \\
&= \frac{x_0^\top (A^{2t} - A^{2t+1}) x_0}{x_0^\top A^t A^t x_0} \\
&= \frac{x_0^\top A^{2t}(I - A) x_0}{\|A^t x_0\|_2^2} \\
&= \frac{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t}(1 - \lambda_i)}{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t}}
\end{aligned}$$

We conclude:

Proposition 9.1. *If $x_0^\top v_i \neq 0$ for i with $\lambda_i = 1$, then $\hat{\lambda}_t \to \lambda_1$ as $t \to \infty$.*

Moreover, we can derive a convergence rate: Assume for simplicity that $\lambda_3 < \lambda_2 < \lambda_1$. The proof works analogously if the multplicities are higher. Moreover assume that the correlation of x_0 with v_2 is non-zero. Otherwise replace λ_2 by the smallest λ_i such that the correlation is nonzero, i.e. $\omega_i \neq 0$. The error at iteration t is denoted $\varepsilon_t := 1 - \hat{\lambda}_t$. Then,

$$\begin{aligned}
\frac{\varepsilon_{t+1}}{\varepsilon_t} &= \frac{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t+2}(1 - \lambda_i)}{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t}(1 - \lambda_i)} \frac{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t}}{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t+2}} \\
&= \frac{\omega_2^2 \lambda_2^{2t+2}(1 - \lambda_2) + \sum_{i>2}^n \omega_i^2 \lambda_i^{2t+2}(1 - \lambda_i)}{\omega_2^2 \lambda_2^{2t}(1 - \lambda_2) + \sum_{i>2}^n \omega_i^2 \lambda_i^{2t}(1 - \lambda_i)} \frac{\omega_1^2 + \sum_{i=1}^n \omega_i^2 \lambda_i^{2t}}{\omega_1^2 + \sum_{i>1}^n \omega_i^2 \lambda_i^{2t+2}} \\
&\to \lambda_2^2, \quad t \to \infty.
\end{aligned}$$

Hence, in the limit $t \to \infty$, we have the **asymptotic convergence rate**

$$\frac{\varepsilon_{t+1}}{\varepsilon_t} = \left(\frac{\lambda_2}{\lambda_1}\right)^2$$

when $\omega_1 \neq 0$. This implies exponential convergence in t, namely $\left(\frac{\lambda_2}{\lambda_1}\right)^{2t}$.

Note that if the multiplicity of λ_1 is greater one, x_t can converge to any linear combination of leading eigenvectors. Usually, we look at problems with unique leading eigenvalues and want to determine the convergence rate. This is then determined through the **spectral gap** $\lambda_1 - \lambda_2$. If the (relative) spectral gap is small, we have slower (asymptotic) convergence than when the spectral gap is larger. This is because in the latter case, the correlation with smaller eigenvectors diminishes more in each step and correlation with the leading eigenvector is favored more. Note that the method fails when $\omega_1 = 0$ because there is no way to construct correlation with v_1 by just multiplying A with the vector.

One way to establish $\omega_1 \neq 0$ is to choose $x_0 \sim \mathcal{N}(0, I)$. Note that by this choice it holds

$$\mathbb{P}(\omega_1 \neq 0) = \mathbb{P}(\omega_i \neq 0) = 1$$

for any $i \in \{1, \ldots, n\}$. Moreover, this random choice allows us to derive a non-asymptotic convergence rate with high probability. We state it as a bound on the expectation:

Proposition 9.2 ([Tro20]). *Assume $A \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite and let $x_0 \sim \mathcal{N}(0, I)$. Then, for any $t \geq 0$ it holds*

$$\mathbb{E}[\varepsilon_t] \leq \sqrt{2n} \left(\frac{\lambda_2}{\lambda_1}\right)^t. \tag{9.1}$$

A bound in probabilities was first established in [KW92] and later, Proposition 9.2 was proven by Joel Tropp [Tro20] and was simplified remarkably.

Note that the bound (9.1) is weaker than the asymptotic rate, i.e. $\frac{\lambda_2}{\lambda_1}$ instead of $\left(\frac{\lambda_2}{\lambda_1}\right)^2$ and we have a prefactor that could become important in high dimensions.

In the proof we will use the fact that $\mathbb{E}[\omega_i^2] = 1$ and so we can use a bound on ω_i^2 with high probability. Moreover, we use the following fact for Gaussian random variables: If $X \sim \mathcal{N}(0,1)$, then for any constant $c \geq 0$:

$$\mathbb{E}\left[\frac{c}{X^2 + c}\right] \leq \sqrt{\frac{c\pi}{2}}. \tag{9.2}$$

Proof: (of Proposition 9.2) Assume again without loss of generality that $\lambda_1 = 1$. Recall that

$$\begin{aligned}
\mathbb{E}[\varepsilon_t] &= \mathbb{E}\left[\frac{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t}(1-\lambda_i)}{\sum_{i=1}^n \omega_i^2 \lambda_i^{2t}}\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i>1}^n \omega_i^2 \lambda_i^{2t}}{\omega_1^2 + \sum_{i>1} \omega_i^2 \lambda_i^{2t}} \mid \omega_2, \ldots, \omega_n\right]\right] \\
&\leq \mathbb{E}\left[\sqrt{\frac{\pi}{2}\sum_{i>1} \omega_i^2 \lambda_i^{2t}}\right].
\end{aligned}$$

where we applied (9.2) in the second inequality. Now apply Jensen's inequality to obtain

$$\mathbb{E}\left[\sqrt{\frac{\pi}{2}\sum_{i>1} \omega_i^2 \lambda_i^{2t}}\right] \leq \sqrt{\frac{\pi}{2}\sum_{i>1} \mathbb{E}\left[\omega_i^2\right] \lambda_i^{2t}} \leq \sqrt{2(n-1)\lambda_2^2}$$

which concludes the proof. ∎

Convergence **independent of the spectral gap** was established in [KW92]. We can recover this result by using the Gaussian nature of x_0:

Exercise 9.3. *Adapt the proof strategy from Proposition 9.2 to prove that for any $0 < \beta < 1$ it holds*

$$\mathbb{E}[\varepsilon_t] \leq (1-\beta) + \sqrt{2n}\exp(-(1-\beta)t)$$

and optimize over β to see that for any $t \geq 1$:

$$\mathbb{E}[\varepsilon_t] \leq \frac{1 + \log(\sqrt{2n} + \log(t))}{t}.$$

For a proof, see the lecture notes [KT23].

9.2. Randomized SVD

The randomized SVD algorithm was formalized in [HMT11]. We will follow this work and the lecture notes [KT23] here.

9.2.1. Motivation

The singular value decomposition provides the best rank-k approximation of a matrix $A \in \mathbb{R}^{m \times n}$:

$$A \approx A_k = U_k \Sigma_k V_k^\top,$$

and no other rank-k matrix achieves a smaller approximation error in either the operator or Frobenius norm. However, computing the full SVD requires $O(mn^2)$ operations (for $m \geq n$), which is too expensive for large-scale problems.

A useful observation is that the leading singular directions of A typically lie in a low-dimensional subspace. Rather than computing the entire SVD, it therefore suffices to identify a subspace that captures most of the range of A. To identify these directions, one would have to compute the SVD beforehand and we would not gain anything. But if we want to succeed with high probability, random projections are an efficient way to do this: given a random test matrix $\Omega \in \mathbb{R}^{n\times\ell}$ (with ℓ only slightly larger than k), the sketch

$$Y = A\Omega$$

spans, with high probability, a subspace that contains almost all of the relevant information in A.

Once such a subspace Q has been obtained (e.g. by performing a QR decomposition of Y), one forms the smaller matrix

$$B = Q^\top A,$$

computes its SVD, and then lifts the result back to the original space. This yields a near-optimal rank-k approximation to A at a fraction of the cost of the full SVD.

Randomized SVD thus combines the approximation quality of the classical SVD with the efficiency of random projections.

9.2.2. Intuitive Approach

Draw a random vector $x \sim \mathcal{N}(0, I_n)$ in $\mathbb{R}^n$. Recall that

$$Ax = \sum_{i=1}^{r} \sigma_i \, u_i v_i^\top x, \qquad r := \min\{m, n\},$$

and define $\omega_i := v_i^\top x$. Because $\{v_i\}$ is orthonormal and x is Gaussian, the coefficients ω_i are i.i.d. $\mathcal{N}(0, 1)$, and therefore $\mathbb{E}[\omega_i^2] = 1$.

Hence, in expectation, Ax places larger weight on left singular vectors corresponding to larger singular values. If the singular values decay, Ax is typically dominated by the top few singular directions.

If we repeat this construction for $s = k + p$ (for a small oversampling parameter p) independent Gaussian vectors, then

$$Ax_1, \ldots, Ax_s$$

tend to span a subspace that captures almost all of the range of

$$P_k := U_k U_k^\top,$$

i.e. the subspace spanned by the leading k left singular vectors.

Form the random matrix

$$\Omega := [x_1, \ldots, x_s] \in \mathbb{R}^{n\times s}, \qquad Y := A\Omega \in \mathbb{R}^{m\times s}.$$

The idea is now to approximate the column space of A by the column space of Y. This is effective because if Y captures the dominant left-singular subspace, then projecting A onto range(Y) preserves nearly all of its "energy" (i.e. its action on vectors), while discarding directions associated with small singular values.

To obtain an orthonormal basis for range(Y), we compute a QR decomposition

$$Y = QR,$$

where $Q \in \mathbb{R}^{m\times s}$ has orthonormal columns and $R \in \mathbb{R}^{s\times s}$ is upper triangular.

We now compress A to the small matrix

$$B := Q^\top A \in \mathbb{R}^{s\times n}.$$

Optionally, we compute the SVD of B,

$$B = \hat{U}_0 \hat{\Sigma}_0 \hat{V}_0^\top,$$

Note that B is rectangular with $s \ll m$, so its SVD has the form

$$B = \hat{U}_0 \hat{\Sigma}_0 \hat{V}_0^\top,$$

where

$$\hat{U}_0 \in \mathbb{R}^{s \times s}, \qquad \hat{\Sigma}_0 \in \mathbb{R}^{s \times s}, \qquad \hat{V}_0 \in \mathbb{R}^{n \times s}.$$

Lifting this back gives the rank-s approximation

$$\hat{A}_s = QB = (Q\hat{U}_0)\hat{\Sigma}_0 \hat{V}_0^\top = \hat{U}\hat{\Sigma}_0 \hat{V}_0^\top.$$

9.2.3. Algorithm [HMT11]

Algorithm 2 Randomized SVD (basic version)

Require: Matrix $A \in \mathbb{R}^{m \times n}$, target rank k, oversampling parameter p
Ensure: Approximate rank-$(k+p)$ SVD of A

1: Draw a Gaussian test matrix $\Omega \sim \mathcal{N}(0,1)^{n \times (k+p)}$
2: Form the sample matrix $Y = A\Omega$
3: Compute a thin QR decomposition $Y = QR$
4: Form the compressed matrix $B = Q^\top A$
5: Compute the (economy-sized) SVD of B: $B = \hat{U}_0 \hat{\Sigma}_0 \hat{V}_0^\top$
6: Set $\hat{U} = Q\hat{U}_0$
7: **return** $\hat{U}$, $\hat{\Sigma}_0$, $\hat{V}_0$

9.2.4. Theoretical Guarantees

Let $A \in \mathbb{R}^{m \times n}$ have the singular value decomposition

$$A = U\Sigma V^\top = U_k \Sigma_k V_k^\top + U_\perp \Sigma_\perp V_\perp^\top,$$

where U_k, V_k correspond to the top k singular vectors, and $U_\perp, V_\perp$ correspond to the remaining singular directions. Let $\Omega \in \mathbb{R}^{n \times s}$ be a Gaussian random matrix with $s = k + p$ columns, and decompose it in the basis of V as

$$\Omega = V_k^\top \Omega + V_\perp^\top \Omega =: \Omega_k + \Omega_\perp.$$

Then one can show the following lemma (see [HMT11]):

Lemma 9.4 (Error decomposition for randomized SVD). *The Frobenius norm of the approximation error satisfies*

$$\|A - \hat{A}_s\|_F^2 \le \|\Sigma_\perp\|_F^2 + \|\Sigma_\perp \Omega_\perp \Omega_k^\dagger\|_F^2,$$

where $\hat{A}_s$ is the rank-s approximation obtained from the randomized procedure, and $\Omega_\perp^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\Omega_\perp$.

For a proof see the main lecture notes [BSS25]. The first term, $\|\Sigma_\perp\|_F^2$, is exactly the minimal error achievable by any rank-k approximation of A (the Eckart–Young theorem). Therefore, the main challenge is to control the second term, $\|\Sigma_k \Omega_k \Omega_\perp^\dagger\|_F^2$.

To bound this term, we first recall a standard fact about Gaussian matrices:

Lemma 9.5 (Expected Frobenius norm of a Gaussian pseudo-inverse). *Let $G \in \mathbb{R}^{m \times n}$ be a standard Gaussian matrix with $m > n + 1$. Then*

$$\mathbb{E}\big[\|G^\dagger\|_F^2\big] = \frac{m}{m - n - 1}.$$

Proof sketch. The matrix $GG^\top$ follows a Wishart distribution. The expected trace of its inverse gives the result.

Now, Ω_k and $\Omega_\perp$ are independent Gaussian matrices. Using the law of total expectation and factoring out the terms in the Frobenius norm, we can write

$$\mathbb{E}\big[\|\Sigma_\perp \Omega_\perp \Omega_k^\dagger\|_F^2\big] = \mathbb{E}\Big[\,\mathbb{E}\big[\|\Sigma_\perp \Omega_\perp \Omega_k^\dagger\|_F^2 \,\big|\, \Omega_k\big]\,\Big].$$

Conditioning on Ω_k, we can factor out Σ_k and apply the previous lemma to $\Omega_k^\dagger$, giving

$$\mathbb{E}\big[\|\Sigma_k \Omega_k \Omega_\perp^\dagger\|_F^2\big] \le \|\Sigma_\perp\|_F^2\, \mathbb{E}\big[\|\Omega_k^\dagger\|_F^2\big] = \sum_{j>k} \sigma_j^2 \,\frac{k}{p-1}.$$

Here $p = s - k$ is the oversampling parameter, and we assumed $p > 1$.

Theorem (Expected error of randomized SVD, [HMT11]). Let $\hat{A}_s$ be the randomized rank-s approximation of A constructed with $s = k + p$ Gaussian test vectors. Then

$$\mathbb{E}\big[\|A - \hat{A}_s\|_F^2\big] \le \underbrace{\sum_{j>k} \sigma_j^2}_{\text{optimal rank-}k\text{ error}} + \underbrace{\sum_{j>k} \sigma_j^2 \frac{k}{p-1}}_{\text{oversampling term}}.$$

9.2.5. Computational Cost

Let $A \in \mathbb{R}^{m\times n}$ and suppose we want a rank-k approximation with $s = k + p$ Gaussian test vectors. The main steps of randomized SVD and their costs are as follows:

1. **Forming the sketch $Y = A\Omega$:** Computing the matrix-matrix product with $\Omega \in \mathbb{R}^{n\times s}$ requires

   $$O(mns)$$

   operations, assuming Ω is dense (e.g., Gaussian) and A is unstructured.

2. **QR decomposition of Y:** Computing a QR decomposition $Y = QR$, where $Q \in \mathbb{R}^{m\times s}$, costs

   $$O(ms^2).$$

3. **Forming the small matrix $B = Q^\top A$:** This matrix-matrix multiplication requires

   $$O(mns)$$

   operations.

4. **SVD of the small matrix $B \in \mathbb{R}^{s\times n}$:** Computing the (economy-sized) SVD of B costs

   $$O(s^2 n).$$

Therefore, the total computational cost is dominated by the two $O(mns) = O(mn(k+p))$ terms. This is significantly cheaper than the $O(mn^2)$ cost of the classical full SVD when $k \ll n$, which makes randomized SVD very attractive for large-scale matrices.

9.2.6. Further Ideas

A natural way to improve the accuracy of randomized SVD is to use *subspace iteration* (or power iteration). The idea is to amplify the contribution of the leading singular directions relative to the smaller ones.

Given the Gaussian sketch $Y = A\Omega$, we can replace it by

$$Y = (AA^\top)^q A\Omega,$$

where $q \geq 1$ is a small integer. If the singular value decomposition of A is $A = U\Sigma V^\top$, then

$$(AA^\top)^q A\Omega = U\Sigma^{2q+1}V^\top\Omega.$$

The factor Σ^{2q+1} raises each singular value to a high power. After forming Y, we can proceed as usual: compute a thin QR $Y = QR$ and then $B = Q^\top A$.

Remark 9.6. *When singular values decay very slowly, the oversampling parameter p becomes more important: increasing p gives Y a larger dimension, which increases the chance that the dominant subspace is captured accurately.*

10. Weak Recovery in the Stochastic Block Model

Recall the Stochastic block model with multiple communities:

Definition 10.1. *A graph $G = (V, E)$ consisting of n nodes with community memberships $c \in [k]^n$ is drawn from the stochastic block model if each edge $(i, j) \in E$ is drawn independently with probability Q_{c_i,c_j}.*

We will consider the probabilities $Q_{lk} = \frac{a}{n}$ if $l = k$ and $Q_{lk} = \frac{b}{n}$ if $l \neq k$ and $l, k \in [k]$ with a and b constant. The vector c is unknown and usually also randomly generated, following a distribution p. Therefore, we define $(c, G) \sim SBM(n, p, Q)$ as a random draw from the Stochastic Block Model.

Now suppose we observe the edges E but not the community labels of a graph G. Assume the labels follow a prior distribution denoted by P. The posterior of $\sigma \in [k]^n$ is then given by

$$P(\sigma \mid E) = \frac{1}{Z} P(E \mid \sigma) P(\sigma) = \frac{1}{Z} \prod_{i \sim j} (\frac{a}{n} \mathbb{1}_{\sigma_i = \sigma_j} + \frac{b}{n} \mathbb{1}_{\sigma_i \neq \sigma_j}) \prod_{i \not\sim j} (1 - \frac{a}{n} \mathbb{1}_{\sigma_i = \sigma_j} - \frac{b}{n} \mathbb{1}_{\sigma_i \neq \sigma_j}) P(\sigma)$$

where Z is a normalizing constant not depending on σ and the product is over all edges and over all non-edges.

If we have two communities and choose the labels to be in $\{-1, +1\}$, this can be simplified to

$$\begin{aligned} P(\sigma \mid E) &= \frac{1}{Z} \exp(\sum_{i \sim j} (\log(\frac{a}{n}) \mathbb{1}_{\sigma_i = \sigma_j} + \log(\frac{b}{n}) \mathbb{1}_{\sigma_i \neq \sigma_j}) + \sum_{i \not\sim j} (\log(1 - \frac{a}{n}) \mathbb{1}_{\sigma_i = \sigma_j} \log(1 - \frac{b}{n}) \mathbb{1}_{\sigma_i \neq \sigma_j})) P(\sigma) \\ &= \frac{1}{\tilde{Z}} \exp(\sum_{i \sim j} \frac{\log(a) - \log(b)}{2} \sigma_i \sigma_j + \sum_{i \not\sim j} \frac{\log(1 - \frac{a}{n}) - \log(1 - \frac{b}{n})}{2} \sigma_i \sigma_j) P(\sigma) \\ &= \frac{1}{\tilde{Z}} \exp(\sum_{i \sim j} \theta_+ \sigma_i \sigma_j + \sum_{i \not\sim j} \theta_- \sigma_i \sigma_j) P(\sigma) \end{aligned}$$

with $\theta_+ := \frac{1}{2} \log(\frac{a}{b}) > 0$ and $\theta_- := \frac{1}{2} \log(\frac{1 - \frac{a}{n}}{1 - \frac{b}{n}}) < 0$.

10.1. Zero-Temperature Belief Propagation

If we have prior beliefs about σ, we want to update them step by step. First assume that the prior is deterministic in all nodes but one, denoted by j, and $P(\sigma_j = 1) = \frac{1}{2}$. Then, the posterior becomes

$$P(\sigma_j = 1 \mid E, \{\sigma_i\}_{i \neq j}) = \frac{1}{Z} \exp(\sum_{i \sim j} \theta_+ \sigma_i + \sum_{i \not\sim j} \theta_- \sigma_i).$$

If the posterior should again be deterministic about the updated belief for σ_j, we add an inverse temperature parameter $\lambda > 0$ and define

$$P_\lambda(\sigma_j := 1 \mid E, \{\sigma_i\}_{i \neq j}) = \frac{1}{Z} \exp(\lambda(\sum_{i \sim j} \theta_+ \sigma_i + \sum_{i \not\sim j} \theta_- \sigma_i)).$$

When $\lambda \to \infty$, σ_j is chosen to maximize the a-posteriori probability. Assume that the prior is roughly balanced, i.e. $|\sum_{i \not\sim j} \sigma_i| \leq C$ for a constant $C > 0$, since θ_- is of order $\frac{1}{n}$, the a-posteriori probability is maximal for

$$\sigma_j = sign(\sum_{i \sim j} \sigma_i)$$

which is the majority of its neighbors.

10.2. Notions of Recovery

We aim to recover the community vector c by an algorithm. Therefore we define a metric that measures the closeness between two assignments. Note that permutations of $[k]$ do not matter for detecting communities and therefore we take the maximum over permutations:

$$A(\sigma, c) := \frac{1}{n} \max_{\pi \in S_k} \sum_{i=1}^{n} \mathbb{1}_{c_i = \pi(\sigma_i)}$$

In the optimal case we would like to fully recover our signal. Define exact recovery as follows:

Definition 10.2 (Exact Recovery). *Let $(c, G) \sim SBM(n, p, Q)$. An algorithm achieves exact recovery if it outputs a guess σ for which it holds*

$$\Pr(A(\sigma, c) = 1) \to 1, \quad n \to \infty.$$

If we aim to achieve strong recovery, it is in general useful to maximize the a-posteriori likelihood because this minimizes the probability of making a mistake. Note that in the regime with fixed a and b and two communities, strong recovery is impossible and therefore, we will aim for weak recovery in general. Assume that the probabilities for belonging to one community under p is $\frac{1}{k}$. Then, weak recovery holds if an algorithm is better than random guessing:

Definition 10.3 (Weak Recovery). *Let $(c, G) \sim SBM(n, p, Q)$. If there exists an $\epsilon > 0$ such that the algorithm outputs a guess σ for which it holds*

$$\Pr(A(\sigma, c) \geq \frac{1}{k} + \epsilon) \to 1, \quad n \to \infty,$$

it achieves weak recovery.

If only weak recovery is required, maximizing the a-posteriori likelihood is not necessarily the optimal way and instead we want to update our beliefs of σ according to the posterior distribution.

In the following section we will derive an algorithm which updates the beliefs about the true labels iteratively. In contrast to 0-temperature belief propagation, the updates are made according to the posterior and therefore, we do not get deterministic values, but probability distributions.

10.3. Belief Propagation on Trees

We will introduce the algorithm on tree models because prior beliefs are easy to propagate here. In general, we will assume that the prior beliefs affect the nodes independently in belief propagation. At first glance this assumption seems to be wrong in general graph models. But in the case of the Stochastic Block model, we will see that subgraphs of it are close to a family of trees.

How does Belief Propagation work on a tree? Suppose you are given a tree where each vertex v has a value in $\{-1, +1\}$. The children of v are generated independently and have the same value as v with probability α and the flipped value with probability $1 - \alpha$.

1. Fix beliefs at depth d of the tree. We assign a probability p_j^d to each vertex σ_j^d of being 1. The probabilities will be small pertubations of $\frac{1}{2}$.

2. Update the belief of each parent. Therefore, assume a vertex σ_i^{d-1} has k children $\{\sigma_j^d\}_{j=1}^k$:

$$\begin{aligned}
\Pr(\sigma_i^{d-1} = 1 \mid \{p_j^d\}_{j=1}^k) &= \sum_{\tau \in \{-1,1\}^k} \Pr(\sigma_i^{d-1} = 1, \sigma_j^d = \tau_j \quad \forall j = 1, \dots, k \mid \{p_j^d\}_{j=1}^k) \\
&= \sum_{\tau \in \{-1,1\}^k} P_\alpha(\sigma_i^{d-1} = 1 \mid \sigma_j^d = \tau_j \quad \forall j = 1, \dots, k) \prod_{j=1}^k p_j^d(\tau_j) \\
&= \frac{1}{Z} \sum_{\tau \in \{-1,1\}^k} P_\alpha(\sigma_j^d = \tau_j \quad \forall j = 1, \dots, k \mid \sigma_i^{d-1} = 1) \prod_{j=1}^k p_j^d(\tau_j) \\
&= \frac{1}{Z} \sum_{\tau \in \{-1,1\}^k} \prod_{j=1}^k (\frac{1}{2} + (\alpha - \frac{1}{2})\tau_j) p_j^d(\tau_j) \\
&= \frac{1}{Z} \prod_{j=1}^k (\alpha p_j^d + (1-\alpha)(1-p_j^d))
\end{aligned}$$

where Z is again a normalizing constant and P_α is the distribution of the tree, parameterized by α.

3. Repeat this for $d-2, d-1, \dots, 1, 0$.

If the beliefs about σ^0, the root vertex, are $p^0 = \frac{1}{2}$, the algorithm does not perform better than a random guess. In this case, $p_j^l = \frac{1}{2}$ for all j, l is a fixed point of the algorithm and therefore we will not use this as an initialization of the beliefs, but small pertubations of it.

10.4. Belief Propagation for the Stochastic Block Model

We analyze belief propagation for the Stochastic Block Model for $k = 2$. Recall that we derived the posterior

$$P(\sigma \mid E) = \frac{1}{Z} \exp(\sum_{i \sim j} \theta_+ \sigma_i \sigma_j + \sum_{i \nsim j} \theta_- \sigma_i \sigma_j) P(\sigma)$$

If we want to do belief propagation as in the tree model, assume you want to update the belief of σ_i and you have beliefs about all other vertices, independently affecting σ_i:

$$P(\sigma_j = \tau_j \quad \forall j \neq i) := \prod_{j \neq i} P_i(\tau_j).$$

The posterior for σ_i becomes

$$P(\sigma_i = 1 \mid E, \{P_i(\sigma_j)_{j \neq i}\}) = \frac{1}{Z} \prod_{j \sim i} \sum_{\tau_j \in \{-1,+1\}} \exp(\theta_+ \tau_j) P_i(\tau_j) \prod_{j \nsim i} \sum_{\tau_j \in \{-1,+1\}} \exp(\theta_- \tau_j) P_i(\tau_j)$$

With this we can recursively define the updated beliefs. But we have to be careful about one detail: The beliefs passed from i to k should not depend on the current belief of σ_k. Therefore, we define the updated beliefs as:

$$P_k(\sigma_i = 1) = \frac{1}{Z} \prod_{j \sim i, j \neq k} \sum_{\tau_j \in \{-1,+1\}} \exp(\theta_+ \tau_j) P_i(\tau_j) \prod_{j \nsim i, j \neq k} \sum_{\tau_j \in \{-1,+1\}} \exp(\theta_- \tau_j) P_i(\tau_j)$$

Let us analyze the convergence of the iterated beliefs. First, define the updated expected value of σ_i which is passed to σ_k:

$$m_{ik} := \mathbb{E}_k[\sigma_i \mid E, (P_i(\sigma_j))_{j \neq i,k}] = P_k(\sigma_i = 1) - P_k(\sigma_i = -1).$$

This simplifies to

$$m_{ik} = \frac{\alpha - \beta}{\alpha + \beta} = \tanh(\frac{1}{2}\log(\frac{\alpha}{\beta}))$$

where we define

$$\alpha := \prod_{j \sim i, j \neq k} (\exp(\theta_+)\frac{1+m_{ji}}{2} + \exp(-\theta_+)\frac{1-m_{ji}}{2}) \prod_{j \nsim i, j \neq k} (\exp(\theta_-)\frac{1+m_{ji}}{2} + \exp(-\theta_-)\frac{1-m_{ji}}{2})$$

and

$$\beta := \prod_{j \sim i, j \neq k} (\exp(-\theta_+)\frac{1+m_{ji}}{2} + \exp(\theta_+)\frac{1-m_{ji}}{2}) \prod_{j \nsim i, j \neq k} (\exp(-\theta_-)\frac{1+m_{ji}}{2} + \exp(\theta_-)\frac{1-m_{ji}}{2})$$

and the denominator ensures that the probabilities sum up to 1.

Moreover, $\frac{1}{2}\log(\frac{\alpha}{\beta})$ can be simplified to

$$\frac{1}{2}\log(\frac{\alpha}{\beta}) = \sum_{j \sim i, j \neq k} \operatorname{atanh}(\frac{a-b}{a+b} m_{ji}) + \sum_{j \nsim i, j \neq k} \operatorname{atanh}(\frac{b-a}{2n-b-a} m_{ji})$$

Note that for large n and a roughly balanced prior the influence of non-neighbors is neglectable, and therefore we approximate

$$m_{ik} \approx \tanh(\sum_{j \sim i, j \neq k} \operatorname{atanh}(\frac{a-b}{a+b} m_{ji}))$$

This approximation only depends on the neighborhood of i now. Moreover, we can justify the factorization of probabilities because at a fixed vertex, SBM behaves like a tree with high probability. In particular, it was shown in [MNS12] that the radius-d-neighborhood of a SBM is distributionally close to Galton-Watson-Tree, when d is logarithmic in n.

Definition 10.4 (Galton Watson Tree). *In a Galton-Watson-Tree, parametrized by (α, k), each vertex is assigned to a spin from $\{-1, +1\}$ and gives birth to $Pois((1-\alpha)k)$ children of the same spin and $Pois(\alpha k)$children of the opposite spin.*

Question 10.5. *Show that the number of neighbors in a SBM is asymptotically Poisson. What are α and k if we want to argue that the SBM is close to a Galton-Watson-Tree?*

We will study the Belief Propagation in the SBM now through Belief propagation on a Galton-Watson-Tree: Suppose you initialize the beliefs with a small mean μ and variance σ^2. Since *tanh* and its inverse linearize for small values, the messages sent to the parent nodes are approximately

$$m_v = \sum_u \frac{a-b}{a+b} m_u = \sum_u (1-2\alpha) m_u$$

where u are the children and $\alpha := \frac{b}{a+b}$. Note that the number of children is a random variable. Then it holds:

$$\mathbb{E}[m_v] = (1-2\alpha)\mathbb{E}[\sum_u m_u] = (1-2\alpha)k((1-\alpha)\mu - \alpha\mu) = (1-2\alpha)^2 k\mu$$

and

$$Var(m_v) = (1-2\alpha)^2 \mathbb{E}[\sum_u \sigma^2 = (1-2\alpha^2)k\sigma^2.$$

Suppose that

$$k(1-2\alpha)^2 = \frac{(a-b)^2}{2(a+b)} < 1,$$

then small pertubations of the trivial fixpoint with beliefs initialized close to 0 are stable and the algorithm can not succeed. This threshold is called the **Kesten-Stigum-Threshold**. Moreover, given the true values in a radius-d-neighborhood of a vertex, we are not able to recover the true spin or make a better guess than random, which suggests that weak recovery below the threshold is impossible in general. Above the threshold there might exist other stable fixpoints which are not-trivial like 0 which corresponds to the prior of having probability $\frac{1}{2}$ for every label.

Remark 10.6. *Weak recovery in the Stochastic Block Model for $k = 2$ is indeed impossible below the Kesten-Stigum threshold which can be shown by a general reduction of the SBM on trees (for a survey see [Abb18], for a proof see [MNS15]).*

Remark 10.7. *The success of weak recovery above the Kesten-Stigum Threshold for $k = 2$ was shown in [KMM+13]. This algorithm aims to find the second eigenvector of a non-backtracking matrix. In contrast to studying the adjacency matrix, the non-backtracking matrix prevents loops and therefore is not biased towards high-degree vertices.*

We considered the model for $k = 2$ only. The Kesten-Stigum Threshold for $k \geq 2$ is

$$\frac{(a-b)^2}{k(a+(k-1)b)} > 1.$$

Remark 10.8. *Interestingly, it was shown [AS16b] that weak recovery is information theoretically possible in a regime below this threshold for $k \geq 4$, but no polynomial time algorithm is known.*

Remark 10.9. *Similarly to the case for $k = 2$, for any $k \geq 2$ there exist algorithms which succeed above the KS-threshold which was shown by Abbe and Sandon [AS16a]. The algorithm is a linearized version of belief propagation - called Approximate Belief Propagation.*

10.5. A Spectral Method for Weak Recovery in the Semi-Dense Regime

Before we discuss a spectral algorithm that succeeds up to the KS-threshold, let us study a *naive* spectral method and identify the regimes where this succeeds. Let A be the (random) adjacency matrix of the random graph G generated by the SBM. If we want to cut the graph into two communities we want to solve the *Balanced Cut*:

$$\min_{z \in \{\pm 1\}^n} z^\top A z \quad \text{s.t.} \quad z^\top \mathbf{1} = 0. \tag{10.1}$$

When the adjacency matrix is perturbed by noise, spectral algorithms can only succeed under specific conditions. First, we consider the SBM when the matrix A is still dense. Then, under specific conditions, the noise can be controlled, if the signal is stronger because the spectrum of the noise matrix is small compared to the leading eigenvalues that come from the block structure. We derived the following matrix to study:

Define $M := \mathbb{E}[A] + pI$ and without loss of generality let the first rows of A belong to V_1 and the last to V_2. Then,

$$M\mathbf{1} = (\frac{n}{2}p + \frac{n}{2}q)\mathbf{1}.$$

and

$$M(\mathbf{1}_{V_1} - \mathbf{1}_{V_2}) = \frac{n}{2}(p-q)(\mathbf{1}_{V_1} - \mathbf{1}_{V_2}).$$

Since M has rank two, we can write it as

$$M = \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top + \frac{p-q}{2}(\mathbf{1}_{V_1} - \mathbf{1}_{V_2})(\mathbf{1}_{V_1} - \mathbf{1}_{V_2})^\top.$$

Since we are interested in the leading eigenvector orthogonal to the all-ones vector, we actually want to study

$$\tilde{M} := M - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top = \mathbb{E}[A] + pI - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top.$$

Note that the leading eigenvector of this matrix does not necessarily have to correspond to the leading eigenvector of the matrix of interest, which is

$$B := A + pI - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top,$$

especially if $A - \mathbb{E}[A]$ has large eigenvalues. But for a specific range of p and q, one can show that these matrices are close with high probability. Moreover, the eigenvectors are close and a partitioning algorithm with respect to the leading eigenvector works with high probability with making only small mistakes. Let us first state the algorithm:

1. Compute the leading eigenvector v of B.
2. Set $V_1 = \{i : v_i \geq 0\}$ and $V_2 = \{i : v_i < 0\}$.

The following can be guaranteed when $p - q$ is of order at least $\sqrt{\frac{p+q}{n}}$:

Theorem 10.10. *Let G be generated by the stochastic block model with parameters n, p, q. If*

$$p - q \geq \frac{C}{\alpha}\sqrt{\frac{p+q}{n}}$$

for a specific constant $C > 0$, and $p+q \geq C\frac{\log(n)^4}{n}$, then with high probability, the spectral algorithm classifies at least $(1-\alpha)n$ vertices correctly.

Remark 10.11. *Note that this algorithm guarantees only recovery of a large part. Actually, exact recovery is possible in an even tighter regime by using a different algorithm, explained in [ASB23].*

We can prove the theorem in four steps: First, we characterize values of p and q such that the variance of each entry of $M - B$ is bounded. Then, we can bound the spectral norm with high probability. Then, we relate the gap in leading eigenvectors of both M and B to the spectral norm of $M - B$. In the end, we will show that the number of errors is bounded by the gap in eigenvectors. First, we use a Theorem from Vu [Vu05] without proof:

Theorem 10.12. *Let $R \in \mathbb{R}^{n\times n}$ be a random matrix with*

1. *$|R| \leq 1$ entrywise,*
2. *$\mathbb{E}[R] = 0$,*
3. *$\mathrm{Var}(R) = \sigma^2$ entrywise.*

Then, with high probability it holds for constants C,C':

$$\|R\| \leq 2\sigma\sqrt{n} + C\sqrt{\sigma}n^{\frac{1}{4}}\log(n)$$

if $\sigma^2 \geq C'\frac{\log(n)^4}{n}$.

We can use this Theorem to bound $\|\tilde{M} - B\|$:

Exercise 10.13. *Show that there exists a constant $C > 0$ such that*

$$\|\tilde{M} - B\| \leq C\sqrt{n(p+q)}$$

with high probability for n sufficiently large and $p + q \geq C\frac{\log(n)^4}{n}$

The second part of the proof relates the eigenvectors to the eigenvalues through the angle between the vectors. We use the celebrated Davis and Kahan theorem for this:

Theorem 10.14 (Davis and Kahan). *Let $v_1, \dots, v_n$ be the eigenvectors of M with eigenvalues $\lambda_1, \dots \lambda_n$ and let w_1 be the leading eigenvector of B with eigenvalue μ_1. Let θ be the angle between w_1 and v_1. Then it holds:*

$$\sin(\theta) \leq \frac{\|M - B\|}{\min_{j \neq 1} |\lambda_1 - \lambda_j|}.$$

Since we know the eigenvalues of $\tilde{M}$ in contrast to B, we can use this theorem easily:

Exercise 10.15. *Let w_1,v_1 be the normalized leading eigenvector of $\tilde{M}$ and B respectively. Show that there is a constant such that*

$$\min\{\|v_1 - w_1\|_2^2, \| - v_1 - w_1\|_2^2\} \leq 4C \frac{\sqrt{n(p+q)}}{n(p-q)}$$

with high probability.

It remains to show that $\min\{\|v_1 - w_1\|_2^2, \| - v_1 - w_1\|_2^2\}$ can be related to the number of errors that are made in the algorithm. Therefore, think about the entries of w_1 and try to come up with a bound on the number of mistakes using $\min\{\|v_1 - w_1\|_2^2, \| - v_1 - w_1\|_2^2\}$:

Exercise 10.16. *Show that the number K of mistakes (incorrectly classified) is bounded by*

$$n \min\{\|v_1 - w_1\|_2^2, \| - v_1 - w_1\|_2^2\}$$

and conclude that

$$K \leq \alpha n$$

when $p - q > \frac{4C}{\alpha}\sqrt{\frac{p+q}{n}}$ and $p + q \geq C\frac{\log(n)^4}{n}$.

Note that this guarantees success for constant p and q and up to order $\frac{\log(n)^4}{n}$ for which you have seen exact recovery in class. But when the expected number of edges is constant, i.e. $p = \frac{a}{n}$ and $q = \frac{b}{n}$, this algorithm will actually not work. The reason for this is essentially that the matrix A is sparse and the noise cannot be controlled *on average*. Especially if there are high-degree vertices, there will be large eigenvalues not necessarily correlated with the communities. We will introduce another matrix now where most of the eigenvalues can be controlled even in the sparse regime.

10.6. A Spectral Method for Weak Recovery

The following section is based on the arguments introduced in [KMM+13] and serves as a heuristic argument for the success of the algorithm. A rigurous analysis of the conjecture made in [KMM+13] can be found in [BLM15]. Define the *Non-Backtracking Matrix* of the random graph as follows: Let E be the (random) set of edges with $|E| = m$ and define $\bar{E}$ as the set of directed edges. Note that $|\bar{E}| = 2m$. For $i, j \in \{2m\}^2$, and i corresponding to $u \to v \in \bar{E}$, j corresponding to $w \to x \in \bar{E}$, define

$$B_{ij} = \begin{cases} 1, & \text{if } v = w, u \neq x, \\ 0, & \text{otherwise.} \end{cases}$$

The powers of this matrix tell the number of non-backtracking paths:

Exercise 10.17. *Show that B^r_{ij} is the number of paths $(v_1, \dots, v_n)$ with $v_i \neq v_{i+1}$, $n = r+1$, $(v_1, v_2) = (u, v)$, $(v_{n-1}, v_n) = (w, x)$. Moreover, show that $(B^r(B^r)^\top)_{ii}$ corresponds to the number of vertices that can be reached in exactly r steps from v via non-backtracking paths.*

We want to analyze the spectrum of B instead the spectrum of A. But we need to be careful:

Question 10.18. *Is B symmetric?*

Therefore, we might have to deal with complex eigenvalues.

Exercise 10.19. *Let $\mu_1, \dots, \mu_{2m}$ be the (possibly complex) eigenvalues of B. Then,*

$$\sum_{i=1}^{2m} |\mu_i|^{2r} \leq \mathrm{tr}((B^r)(B^r)^\top).$$

You may use the decomposition $B = QRQ^\top$ with orthogonal matrices Q and an upper triangular matrix R with the eigenvalues on the diagonal.

In expectation, the trace of $(B^r)(B^r)^\top$ corresponds to the number of directed edges times the expected number of neighbors at exact distance r (without backtracking) which is bounded by $(\frac{a+b}{2})^r$. Hence, if we define a distribution randomly over the eigenvalues, it holds that

$$\mathbb{E}[|\mu|^{2r}] \leq \left(\frac{a+b}{2}\right)^r.$$

Since this holds for any r, most of the eigenvalues should be bounded by $(\frac{a+b}{2})^{\frac{1}{2}}$. Now, if we can find an eigenvalue $> (\frac{a+b}{2})^{\frac{1}{2}}$ such that the corresponding eigenvector is correlated with the truth, it is easy to distinguish it from the remaining ones.

The construction of this eigenvector comes from the matrix A again: Define

$$f^{(r)} \in \mathbb{R}^{|V|}, \quad f_v^{(r)} := \frac{1}{\mu^r} \sum_{u:d(u,v)=r} \sigma_u$$

where $\sigma_u \in \{\pm 1\}$ denotes the community membership.

Exercise 10.20. *Verify that*

$$(Af^{(r)})_v = \mu f_v^{(r+1)} + (d_v - 1) f_v^{(r-1)} \frac{1}{\mu}.$$

Moreover, the expectation of $f_v^{(r)}$ is correlated with the community membership regardless of r if $\mu = \frac{a-b}{2}$:

Exercise 10.21. *Show that $\mathbb{E}[f_v^{(r)}] = \frac{1}{\mu^r}(\frac{a-b}{2})^r \sigma_v$.*

The variance of the sum in the definition of $f_v^{(r)}$ is of order $(\frac{a+b}{2})^r$ and therefore, $f^{(r)}$ is stabilizing for large r if

$$\frac{a+b}{2} < (\frac{a-b}{2})^2.$$

Then, $f := f^{(r)} \approx f^{(r\pm 1)}$ and

$$Af \approx \mu f + \frac{1}{\mu}(D - I)f.$$

Note that D is not a multiple of I in the sparse case in general! Now we can show that f can be obtained through an eigenvector of B instead of A to the eigenvalue $\frac{a-b}{2}$ which is by assumption larger than $(\frac{a+b}{2})^{\frac{1}{2}}$ and easy to detect.

For every $i \in [2m]$ associated with $(u, v) \in \bar{E}$, define

$$g_i^{(r)} := \frac{1}{\mu^r} \sum_{j \in \bar{E}: d(i,j)=r} \sigma_x.$$

Exercise 10.22. *Show that $Bg^{(r)} = \mu g^{(r+1)}$.*

Moreover, when $\mu = \frac{a-b}{2}$ as before, $g^{(r)}$ becomes stable and we define g as the stabilized version of $g^{(r)}$. Then g is an eigenvector of B with eigenvalue μ. By the analysis above, this eigenvalue pops out of the spectrum related to randomness and g can be found via spectral analysis. Moreover, f can be restored from g and thus g can be used to find a classification correlated with the truth:

$$\sum_{u \in N(v)} g^{(r)}_{(u,v)} = f^{(r)}_v.$$

Hence, by summing over all incoming vertices of v, we can recover f_v.

11. Functional Inequalities and Isoperimetry

11.1. Isoperimetry

Given a metric space (X, d) and a measure μ and $A \subseteq X$, define

$$A_t := \{x \in X \mid d(x, A) \leq t\}.$$

The question we deal with today is the following: If for two sets A and B it holds that $\mu(A) = \mu(B)$, for which sets can we say $\mu(A_t) \leq \mu(B_t)$ and in particular, when is the boundary measure of A smaller than the one of B, i.e.

$$\mu^+(A) := \lim_{t \to 0} \frac{\mu(A_t) - \mu(A)}{t} \leq \mu^+(B)?$$

Moreover, will the sets A be independent of the dimension of the metric space X? We will answer this question affirmatively for Standard n-dimensional Gaussian Random Variables and end with a conjecture about the validity for a whole function class.

We start with an example for an isoperimetric inequality which is very intuitive and simple to prove: Let $X = \mathbb{R}^n$, $d = \|\cdot\|_2$ and μ the Lebesgue measure. If A is a ball with $\mathrm{vol}(A) = \mathrm{vol}(B)$ and B any set, then it holds

$$\mathrm{vol}(B_t) \geq \mathrm{vol}(A_t) \quad \text{and} \quad \mu^+(B) \geq \mu^+(A). \tag{11.1}$$

Exercise 11.1. *Show (11.1) using the Brunn-Minkowski inequality*

$$\mathrm{vol}(A + B)^{\frac{1}{n}} \geq \mathrm{vol}(A)^{\frac{1}{n}} + \mathrm{vol}(B)^{\frac{1}{n}}.$$

We will now turn to the setting $X = \mathbb{R}^n$, $d = \|\cdot\|_2$ and γ_n denoting the n-dimensional standard Gaussian measure. We will prove the following: Let $I : [0, 1] \to [0, \frac{1}{\sqrt{2\pi}}]$ and

$$I(x) := \varphi(\Phi^{-1}(x))$$

where Φ denotes the standard one-dimensional Gaussian cumulative distribution function and φ the corresponding density. Then, for every $A \subseteq \mathbb{R}^n$, it holds

$$\gamma_n(A_t) \geq \Phi(\Phi^{-1}(\gamma_n(A)) + t) \quad \text{and} \quad \gamma_n^+(A) \geq I(\gamma_n(A)). \tag{11.2}$$

This implies directly that halfspaces $H := \{x : x_i \leq s\}$ for some $i \in [n]$ and $s \in \mathbb{R}$ minimize the boundary given the volume:

Exercise 11.2. *Prove that $\gamma_n(H_t) = \Phi(s + t)$ and conclude that if $\gamma_n(A) = \gamma_n(H)$, it holds*

$$\gamma_n(A_t) \geq \gamma_n(H_t).$$

This allows us to derive functional inequalities and concentration of Lipschitz functions which we will discuss later. A proof of the Gaussian isoperimetric inequality (11.2) was first derived from the isoperimetric inequality on the sphere by [ST78]. It was later reproven through Semigroups [BL96] and with Functional Inequalities in [Bob97]. We will discuss the latter proof because it allows us to derive several functional inequalities as well.

11.2. Functional Inequalities

The main result of this section is Bobkov's inequality [Bob97]. We start by the following two-point inequality due to Bobkov [Bob97]: For $a, b \in [0, 1]$ it holds:

$$I(\frac{a+b}{2}) \leq \frac{1}{2}\sqrt{I(a)^2 + \left|\frac{a-b}{2}\right|^2} + \frac{1}{2}\sqrt{I(b)^2 + \left|\frac{a-b}{2}\right|^2}.$$

In fact, Bobkov proved that I is the maximal function satisfying this inequality for which also $I(0) = I(1) = 0$. With the following three steps we will arrive at a functional inequality for the Gaussian measure:

Step 1: Let $f : \{\pm 1\} \to [0,1]$ and $X \sim \mu := \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$. Then:

$$I(\mathbb{E}[f(X)]) \leq \mathbb{E}[\sqrt{I(f(X))^2 + |\nabla f(X)|^2}],$$

where $|\nabla f|$ denotes the discrete gradient $|\frac{f(1)-f(-1)}{2}|$.

Step 2: Induction over n Let $f : \{\pm 1\}^n \to [0,1]$ and $X \sim \mu_n := \left(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}\right)^{\otimes n}$. By induction one can show:

$$I(\mathbb{E}[f(X)]) \leq \mathbb{E}[\sqrt{I(f)^2 + |\nabla f|^2}],$$

where $|\nabla f|$ denotes the discrete gradient

$$|\nabla f(x)|^2 := \frac{1}{4}\sum_{i=1}^{n} |f(x) - f(x_{-i})|^2$$

and x_{-i} corresponds to x with flipped coordinate i.

Step 3: Apply the Central Limit Theorem Let $f_N := f(\frac{1}{\sqrt{N}}\sum_{k=1}^{N} x_k)$ for $x_1, \ldots, x_N$ independently from μ_n. If f is twice continuously differentiable, taking $N \to \infty$ and applying the Central Limit Theorem, we have for $Y \sim \gamma_n$

$$I(\mathbb{E}[f(Y)]) \leq \mathbb{E}[\sqrt{I(f(Y))^2 + |\nabla f(Y)|^2}]. \tag{11.3}$$

We can extend this for every locally Lipschitz function f.

Bobkov's inequality has a flavor like the Poincaré inequality that you learned in the lecture: The fluctuation of f can be controlled by its gradient. Indeed, we can derive other functional inequalities from (11.3) for f having sufficiently bounded moments:

Exercise 11.3. *Derive the log-Sobolev inequality*

$$\int f^2 \log(f^2)\, d\gamma_n - \int f^2\, d\gamma_n \log(\int f^2\, d\gamma_n) \leq \int |\nabla f|^2\, d\gamma_n,$$

which bounds the entropy by the norm of the gradient. Hint: Apply Bobkov's inequality for εf^2 and send $\varepsilon \to 0$.

Exercise 11.4. *Derive the Poincaré inequality*

$$\mathrm{Var}(f) \leq \int |\nabla f|^2\, d\gamma_n,$$

which bounds the variance by the norm of the gradient. Hint: Apply log-Sobolev inequality for $1 + \varepsilon f$ and send $\varepsilon \to 0$.

11.3. Proving Gaussian Isoperimetry

We will prove (11.2) by a simple application of Bobkov's inequality: Define

$$f_t(x) := \begin{cases} 1, & x \in A, \\ 1 - \frac{d(x,A)}{t}, & x \in A_t \backslash A, \\ 0, & x \in \mathbb{R}^n \backslash A_t. \end{cases}$$

Question 11.5. *What is the Lipschitz constant of f_t?*

Note that f_t is not differentiable. To make it so, in order to apply Bobkov's inequality, we smooth it and define

$$f_{t,\sigma}(x) := \mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2 I)}[f_t(X)]$$

Note that $\lim_{\sigma\to 0} f_{t,\sigma}(x) = f_t(x)$, i.e. we have pointwise convergence.

Now, we can show that $f_{t,\sigma}$ is $\frac{1}{t}$-Lipschitz and we can apply Bobkov's inequality for every $f_{\sigma,t}$. By dominated convergence, we have

$$I(\int f_{t,\sigma}\, d\gamma_n) \to I(\gamma_n(A)), \quad t \to 0.$$

Moreover,

$$\lim_{t\to 0} \int \sqrt{I(f_{t,\sigma})^2 + \|\nabla f_{t,\sigma}\|_2^2}\, d\gamma_n \le \gamma_n^+(A).$$

Hence we can derive the second part of (11.2). For the first part, define

$$h(t) := \Phi^{-1}(\gamma_n(A_t)).$$

Note that

$$h'(t) = \frac{\gamma_n^+(A_t)}{I(\gamma_n(A_t))} \ge 1$$

for all $t \ge 0$ by the previous analysis. Therefore,

$$h(t) \ge h(0) + t$$

and

$$\gamma_n(A_t) = \Phi(h(t)) \ge \Phi(h(0)+t) = \Phi(\Phi^{-1}(\gamma_n(A)) + t))$$

which proves the first part as well.

11.4. Lipschitz Concentration

As an application of Gaussian Isoperimetry, we can prove **dimension-independent** concentration of Lipschitz functions for Gaussian vectors: Let $f : \mathbb{R}^n \to \mathbb{R}$ be σ-Lipschitz continuous and $X \sim \gamma_n$, then there exist constants $c, C > 0$ such that

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \ge t) \le C\exp(-c\frac{t^2}{2\sigma^2}).$$

To prove this, apply the isoperimetric inequalities to the sets

$$A_1 := \{f \ge \mathrm{med}(f)\}, \quad A_2 := \{f \le \mathrm{med}(f)\}.$$

Note that by Lipschitz-continuity

$$A_{1,t/\sigma} \subseteq \{f \ge \mathrm{med}(f) - t\}, \quad A_{2,t/\sigma} \subseteq \{f \le \mathrm{med}(f) + t\}.$$

Exercise 11.6. *By comparing with half-spaces, conclude that*

$$\mathbb{P}(|f(X) - \mathrm{med}(f)| \ge t) = \gamma_n(A_{1,t/\sigma}^C \cup A_{2,t/\sigma}^C) \le 2\exp(-c_1\frac{t^2}{\sigma^2}).$$

Now we want to derive the inequality for $\mathbb{E}[f(X)]$ instead of $\text{med}(f)$:

Exercise 11.7. *Show*

$$\mathbb{P}(|f(X) - f(\tilde{X})| \geq t) \leq 4\exp(-c_1\frac{t^2}{\sigma^2})$$

for an independent copy $\tilde{X}$. Then, prove that

$$\mathbb{E}[\exp(\lambda^2(f(X) - f(\tilde{X}))^2)] \leq 2\lambda^2 \int_0^\infty t\exp((\lambda^2 - \frac{c_1}{\sigma^2})t^2)\,dt.$$

Conclude by Markov and Jensen and choosing the right λ that

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq C\exp(-c_1\frac{t^2}{2\sigma^2}).$$

11.5. Beyond Gaussian Distributions

We have seen that the Isoperimetric Inequality is pretty strong: We can derive several other functional inequalities independent of the dimension and we can derive concentration in high dimensions! Therefore it would be a pretty strong statement if we can derive a dimension-independent isoperimetric inequality for a whole function class. The KLS conjecture ([KLS95]) is about the isoperimetric inequality for all log-concave probability measures:

Conjecture 11.8 ([KLS95]). *For every log-concave probability measure it holds*

$$\inf_{S\subseteq\mathbb{R}^d} \frac{\mu^+(S)}{\min\{\mu(S),\mu(S^c)\}} \geq c \cdot \inf_{H\subseteq\mathbb{R}^d, H \text{ is halfspace}} \frac{\mu^+(H)}{\min\{\mu(H),\mu(H^c)\}}$$

where c is a universal, dimension-free constant. In particular,

$$\inf_{S\subseteq\mathbb{R}^d} \frac{\mu^+(S)}{\min\{\mu(S),\mu(S^c)\}} \geq \frac{c}{\sqrt{\|A\|_2}}$$

where $A = \text{Cov}_\mu(X)$.

Significant progress on resolving this conjecture has been made by Yuansi Chen [Che21], improving over previously known polynomial bounds on c to a *almost-constant* lower bound. It was shown by using a stochastic localization process and tilt the density p to

$$p_t(x) \propto \exp(-\frac{t}{2}x^\top A^{-1}x + y_t^\top x)p(x),$$

where y_t is a stochastic process that ensures p_t to define a martingale over measures. Then, p_t is more log-concave than a Gaussian with Covariance matrix $\frac{1}{t}A$.

This allows to study

$$p^+(S) = \mathbb{E}[p_t^+(S)] \geq \mathbb{E}[\frac{1}{2}\|(tA^{-1})^{-1}\|^{-\frac{1}{2}}\min\{p_t(S), p_t(S^c)\}]$$

and hence, if $\|A\|_2 = 1$ and $p(S) = \frac{1}{2}$, control $\mathbb{P}(\frac{1}{4} \leq p_t(S) \leq \frac{3}{4})$ for t as large as possible.

Boaz Klartag [Kla23] improved over this to prove that $c \geq \frac{1}{C\sqrt{\log(d)}}$ which is the currently best-known bound. While related problems like the Thin-Shell Conjecture and Bourgain's Slicing Conjecture have been recently proven ([KL25b], [KL25a]), the KLS conjecture still remains open.

References

[AAGM15] Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D Milman. *Asymptotic geometric analysis, Part I*, volume 202. American Mathematical Society, 2015.

[AAMM24] Abdulmajeed Alqasem, Heshan Aravinda, Arnaud Marsiglietti, and James Melbourne. On a conjecture of feige for discrete log-concave distributions. *SIAM Journal on Discrete Mathematics*, 38(1):93–102, 2024.

[Abb18] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13(3-4):457–466, 1998.

[AS16a] Emmanuel Abbe and Colin Sandon. Achieving the ks threshold in the general stochastic block model with linearized acyclic belief propagation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[AS16b] Emmanuel Abbe and Colin Sandon. Crossing the ks threshold in the stochastic block model with information theory. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 840–844, 2016.

[AS16c] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[ASB23] T. Strohmer A. S. Bandeira, A. Singer. Mathematics of data science, 2023. Accessed: 2025-03-27.

[BJ25] Nikhil Bansal and Haotian Jiang. Decoupling via affine spectral-independence: Beck-fiala and komlós bounds beyond banaszczyk, 2025.

[BL96] D. Bakry and M. Ledoux. Lévy–gromov's isoperimetric inequality for an infinite dimensional diffusion generator. *Inventiones mathematicae*, 123(1):259–281, 1996.

[BLM15] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

[Bob97] S. G. Bobkov. An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in gauss space. *Annals of Probability*, 25(1):206–214, 1997.

[BSS25] Afonso S Bandeira, Amit Singer, and Thomas Strohmer. Topics in mathematics of data science. *Preprint available online: https://people. math. ethz. ch/ abandeira//BandeiraSingerStrohmer-MDS-draft. pdf*, 2025.

[CFM23] Michael Celentano, Zhou Fan, and Song Mei. Local convexity of the tap free energy and amp convergence for z 2-synchronization. *The Annals of Statistics*, 51(2):519–546, 2023.

[Che21] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.

[DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the binary stochastic block model. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 185–189. IEEE, 2016.

[DGGP14] Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. *Combinatorics, Probability and Computing*, 23(1):29–49, 2014.

[Dvo64] Aryeh Dvoretzky. Some results on convex bodies and banach spaces. *Matematika*, 8(1):73–102, 1964.

[Fei06] Uriel Feige. On sums of independent random variables with unbounded variance and estimating the average degree in a graph. *SIAM Journal on Computing*, 35(4):964–984, 2006.

[FMM21] Zhou Fan, Song Mei, and Andrea Montanari. Tap free energy, spin glasses and variational inference. 2021.

[GHLL20] Jiayi Guo, Simai He, Zi Ling, and Yicheng Liu. Bounding probability of small deviation on sum of independent random variables: Combination of moment approach and berry-esseen theorem. *arXiv preprint arXiv:2003.03197*, 2020.

[Grü60] Branko Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. 1960.

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[HZZ10] Simai He, Jiawei Zhang, and Shuzhong Zhang. Bounding probability of small deviation: A fourth moment approach. *Mathematics of Operations Research*, 35(1):208–232, 2010.

[Jer92] Mark Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[KL25a] Boaz Klartag and Joseph Lehec. Affirmative resolution of bourgain's slicing problem using guan's bound. *Geometric and Functional Analysis*, pages 1–22, 2025.

[KL25b] Boaz Klartag and Joseph Lehec. Thin-shell bounds via parallel coupling. *arXiv preprint arXiv:2507.15495*, 2025.

[Kla23] Bo'az Klartag. Logarithmic bounds for isoperimetry and slices of convex sets. *Ars Inveniendi Analytica*, 2023.

[KLS95] Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3):541–559, 1995.

[KMM+13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, November 2013.

[KT23] Anastasia Kireeva and Joel A. Tropp. Randomized matrix computations: themes and variations. 2023.

[KW92] J. Kuczyński and H. Woźniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992.

[MNS12] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction, 2012.

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162:431–461, 2015.

[MV21] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321 – 345, 2021.

[Ple82] Timm Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, 15(6):1971, 1982.

[PWBM18] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.

[Spe85] Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985.

[ST78] V. N. Sudakov and B. S. Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *Journal of Mathematical Sciences*, 9(1):9–18, 1978.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[Tro20] Joel A Tropp. Randomized algorithms for matrix computations, 2020.

[Ver09] Roman Vershynin. High-dimensional probability, 2009.

[Vu05] Van H Vu. Spectral norm of random matrices. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 423–430, 2005.