An Introduction to Computational-Statistical Tradeoffs in High-Dimensional Statistics

Ilias Zadik

Draft monograph
June 2026

Preface

This draft monograph grew out of lecture notes for a graduate course I have taught twice at Yale University on *computational-statistical tradeoffs in high-dimensional statistics*, an exciting, emerging research area at the intersection of mathematical statistics, information theory, theoretical computer science, and mathematics. Because the draft version is based on classroom lectures, they may still contain **many typos** and still lack the ultimate polish of a textbook publication. Nevertheless, my hope is that they provide a helpful introduction to the fundamentals, allowing researchers new to the area to quickly bridge the gap to state-of-the-art research in the field. The only assumed prerequisites for this introductory monograph are a first-year graduate-level understanding of probability theory and mathematical statistics/information theory. Some maturity around algorithms, computational complexity and real analysis at an advanced undergraduate level is also beneficial.

The draft monograph begins with an introduction to the *information-theoretic limits* of high-dimensional statistical tasks, and survey old and new results on the topic. From there, we explore how to argue for the *computational barriers of detection problems*, showing how one can analyze the performance of low-degree polynomials, the statistical query framework, and average-case reductions to obtain rigorous predictions for them. Finally, we pivot to *the computational limits of estimation/recovery tasks*, discussing low-degree polynomials and Markov Chain Monte Carlo (MCMC) methods.

The collection of topics covered here is by no means exhaustive and it is *not* meant to be a survey of all relevant results or papers in the large and growing literature of the topic. For instance, due to the semester's time constraints, several highly impactful directions could not be included, such as Approximate Message Passing (AMP) analysis (see e.g., [FVR+22]), the Franz-Parisi potential (see e.g., [ZK16; BEH+22]), cryptographic connections (see e.g., [BRS+21; LZZ25]), the Sum-of-Squares (SoS) hierarchy (see e.g., [Kun22]) but also some thought-provoking counterexamples that shaped the development of the area [HW21; ZSW+22; DK22; BHJ+25; JV26] .

This draft monograph would not have been possible without the immense help of the CFIS-Polytechnic University of Catalonia undergraduate student **Aleix Artigas Moré**, who, while visiting my group at Yale, meticulously cleaned, organized, and improved the text. They are built upon the initial scribe notes taken by an exceptional group of Yale graduate students: Siyu Chen, Isay Katsman, Shuchen Li, Hugo Latourelle-Vigeant, Max Lovig, Katerina Mamali, Conor Sheehan, Abby Spears, Kostas Tsirkas, Haiyang Wang, Leda Wang, Ruixiao Wang, Xiuyuan Wang, Zijian Wang, and Peiyuan Zhang. I am deeply grateful to all of them for their dedication and hard work.

Contents

Part I

Statistical Foundations

Chapter 1

Statistical Models, Inference Tasks, and the Computational Perspective

1.1 Introduction

The central objective of high-dimensional inference is to recover, estimate, or merely detect hidden structure from noisy observations. In many modern problems, the ambient dimension is large and the space of possible parameters is itself combinatorially or geometrically complicated. As a consequence, two distinct questions arise naturally:

1. **Statistical question:** is the hidden structure in principle recoverable from the data?
2. **Computational question:** even if recovery is statistically possible, can it be achieved by a time-efficient algorithm?

These notes are devoted to the *tension* between these two questions. Before studying computational barriers, however, we must first formalize the statistical models, the inference tasks, and the criteria by which success is measured. The present chapter introduces this general framework.

The material in this chapter is introductory. We define the Bayesian model underlying much of the notes, discuss the main inference tasks —exact recovery, approximate recovery, and detection— and present two canonical examples: sparse regression and planted clique. Although the chapter is introductory, it already contains the seeds of the main themes of the notes.

In particular, it will become clear that exact recovery and detection are fundamentally different tasks and may exhibit distinct thresholds. Moreover, some problems are more naturally formulated as estimation tasks, while others are better understood through the lens of hypothesis testing. The choice of error metric is not merely a technical detail, but rather determines the type of phenomenon one can meaningfully capture. Finally, we will see that the statistically optimal procedure is often straightforward to define, yet computationally intractable.

1.2 A General Bayesian Model for Inference

We begin with the most basic formalization.

Definition 1.1 (Bayesian statistical model)**.** *Let $\mathcal{S}$ be a parameter space and let Ω be an observation space. A Bayesian statistical model consists of a prior distribution μ on $\mathcal{S}$, and for each $\theta \in \mathcal{S}$, a probability distribution $\mathbb{P}_\theta$ on Ω.*

A random parameter θ is first drawn according to

$$\theta \sim \mu,$$

and then the observation Y is drawn according to

$$Y \sim \mathbb{P}_\theta.$$

The hidden object θ will often be referred to as the *signal* or *parameter*. The observed data Y may be a vector, a matrix, a graph, a tensor, or a collection of i.i.d. samples, depending on the problem under consideration.

The Bayesian viewpoint is especially natural in the problems studied in these notes because in planted and random inference problems, the signal itself is often random by design. Moreover, the viewpoint leads naturally to canonical optimal procedures, such as the posterior mean estimator and the MAP estimator, which will later serve as information-theoretic benchmarks.

Remark 1.2. *Throughout this chapter, and often throughout the notes, we do not initially impose any computational restriction on the inference procedure. Thus, when we ask whether a task is possible, we mean possible by* some measurable procedure*, not necessarily an efficient one.*

Remark 1.3. *The choice of prior μ is part of the model. It encodes the structural assumptions one is willing to make about the signal. In particular, sparsity, low rank, and planted combinatorial structure are all common assumptions we will make on the prior.*

1.3 Estimators and Inference Procedures

An inference procedure is a rule that maps data to an output. The precise codomain depends on the task.

Definition 1.4 (Estimator)**.** *An* estimator *is a measurable map*

$$\mathcal{A} : \Omega \to \mathcal{S}.$$

Given an observation $Y \in \Omega$, the estimator outputs $\mathcal{A}(Y)$ as a guess for the hidden parameter θ.

In some problems the aim is not to recover the parameter itself, but to distinguish between two possible data-generating mechanisms. In such settings the relevant object is a statistical test.

Definition 1.5 (Test)**.** *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability distributions on Ω. A* test *is a measurable map*

$$\mathcal{T} : \Omega \to \{\mathbb{P}_1, \mathbb{P}_2\}.$$

Given an observation Y, the test outputs one of the two hypotheses.

The distinction between estimation and testing is basic but important. Estimation attempts to reconstruct a hidden object, whereas testing attempts only to determine which of two models better explains the data. In general, testing is the weaker task and is therefore often statistically easier.

1.4 The Main Inference Tasks

We now formalize the tasks that will appear throughout the notes.

1.4.1 Exact recovery

The strongest notion of successful inference is exact reconstruction of the hidden parameter.

Definition 1.6 (Exact recovery)**.** *An estimator $\mathcal{A} : \Omega \to \mathcal{S}$ achieves* exact recovery with high probability *if*

$$\mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}(\mathcal{A}(Y) = \theta) \to 1$$

as the problem dimension tends to infinity.

More generally, one may require

$$\mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}(\mathcal{A}(Y) = \theta) \geq 1 - \varepsilon$$

for arbitrarily small fixed $\varepsilon > 0$.

Exact recovery is appropriate when the parameter space is discrete or combinatorial. Typical examples include recovering a planted subset, a sparse support, or a hidden graph structure.

Intuitively, exact recovery requires enough information in the data to resolve the parameter uniquely among all candidates in the support of the prior. In many high-dimensional models, this is much stronger than merely producing a correlated estimate.

1.4.2 Approximate recovery

In many continuous models, or even in discrete models in regimes where exact recovery is impossible, it is natural to aim for partial reconstruction rather than exact.

Definition 1.7 (Approximate recovery)**.** *Suppose $\mathcal{S}$ is equipped with a metric $d(\cdot, \cdot)$. An estimator $\mathcal{A} : \Omega \to \mathcal{S}$ achieves* approximate recovery with high probability *at accuracy $\varepsilon > 0$ if*

$$\mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}(d(\mathcal{A}(Y), \theta) \leq \varepsilon) \to 1.$$

In Euclidean models, the most common metric is the ℓ_2 norm. One then often studies the mean-squared error instead of a high-probability criterion.

Definition 1.8 (Mean-squared error)**.** *For an estimator $\mathcal{A} : \Omega \to \mathbb{R}^d$, its* mean-squared error *is*

$$\mathrm{MSE}(\mathcal{A}) := \mathop{\mathbb{E}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}\left[\|\mathcal{A}(Y) - \theta\|_2^2\right].$$

Approximate recovery is particularly well-suited to problems where the signal lives in a Euclidean space and exact identification is either impossible or not the relevant notion of success.

At an intuitive level, exact recovery asks

"Can we identify the signal exactly?"

whereas approximate recovery asks

"Can we estimate the signal nontrivially well?"

The latter is often possible in regimes where the former is not.

1.4.3 Detection and hypothesis testing

A weaker but fundamental task is to determine whether the data contain signal at all.

Definition 1.9 (Detection problem)**.** *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two distributions on Ω. In the* detection problem*, one observes $Y \in \Omega$ sampled from either $\mathbb{P}_1$ or $\mathbb{P}_2$ and seeks to decide which of the two distributions generated the observation.*

Remark 1.10. *Typically $\mathbb{P}_1$ represents a* structured model *and $\mathbb{P}_2$ often represents* pure noise*.*

A test $\mathcal{T} : \Omega \to \{\mathbb{P}_1, \mathbb{P}_2\}$ incurs two kinds of error: it may label a sample from $\mathbb{P}_2$ as structured (type-I error), or label a sample from $\mathbb{P}_1$ as noise (type-II error).

Definition 1.11 (Two-sided error)**.** *The* two-sided error *of a test $\mathcal{T}$ is*

$$\mathit{err}(\mathcal{T}) := \mathop{\mathbb{P}}_{Y \sim \mathbb{P}_1}(\mathcal{T}(Y) = \mathbb{P}_2) + \mathop{\mathbb{P}}_{Y \sim \mathbb{P}_2}(\mathcal{T}(Y) = \mathbb{P}_1)\,.$$

This error criterion will be used repeatedly throughout the notes.

Definition 1.12 (Strong and weak detection)**.** *We say that detection is possible in the sense of*

- strong detection *if there exists a test $\mathcal{T}$ such that*

$$\mathrm{err}(\mathcal{T}) \to 0,$$

- weak detection *if there exists a test T such that*

$$\mathrm{err}(\mathcal{T}) < 1$$

for all sufficiently large dimensions.

The threshold $\mathrm{err}(\mathcal{T}) = 1$ is natural: a completely uninformed test that flips an independent fair coin has error exactly 1. Thus weak detection means doing strictly better than random guessing, while strong detection means asymptotically reliable distinction.

These three tasks will reappear throughout the notes, often with different thresholds and levels of difficulty.

1.5 Canonical Example I: Gaussian Sparse Regression

We now illustrate the general framework with a first example.

Let p be the ambient dimension and let k be a sparsity parameter. Consider the prior

$$\mu = \mathrm{Unif}\left(\{v \in \{0,1\}^p : \|v\|_0 = k\}\right),$$

where

$$\|v\|_0 := |\{i : v_i \neq 0\}|$$

denotes the number of nonzero entries of v. Thus, a parameter $\theta \sim \mu$ is a binary k-sparse vector in $\mathbb{R}^p$.

Given θ, we observe n i.i.d. samples of the form

$$(x_i, y_i), \qquad i = 1, 2, \ldots, n,$$

where

$$x_i \sim \mathcal{N}(0, I_p), \qquad y_i = \langle x_i, \theta \rangle + w_i, \qquad w_i \sim \mathcal{N}\left(0, \sigma^2\right),$$

and all randomness is independent across i.

Equivalently, the data are

$$Y = \{(x_i, y_i) : i = 1, \dots, n\} \sim \mathop{\mathbb{P}}_{\theta},$$

where

$$\mathop{\mathbb{P}}_{\theta} = \left(x_i \sim \mathcal{N}(0, I_p),\ y_i = \langle x_i, \theta \rangle + w_i \text{ where } w_i \sim \mathcal{N}\left(0, \sigma^2\right)\right)^{\otimes n}.$$

This is one of the most basic models in high-dimensional statistics. The signal θ is sparse, the covariates x_i are Gaussian, and the responses y_i are linear measurements of the signal corrupted by Gaussian noise.

The goal is to use the data Y to estimate the hidden vector θ. Depending on the regime of parameters and the notion of success, one may seek to recover the support of θ exactly, to approximate θ in Euclidean norm, or merely to determine whether any signal is present in the data.

Remark 1.13. *The recovery problem becomes easier as the number of samples n increases. Much of high-dimensional inference is therefore concerned with identifying the critical sample size or signal-to-noise ratio at which a qualitative change in behavior occurs.*

The same model also leads naturally to a detection problem. Under the structured hypothesis, the data are generated as above:

$$x_i \sim \mathcal{N}(0, I_p), \qquad y_i = \langle x_i, \theta \rangle + w_i.$$

Observe that here x_i and y_i are correlated through the hidden vector θ.

A natural null model is obtained by removing the signal and keeping only noise:

$$x_i \sim \mathcal{N}(0, I_p), \qquad y_i = w_i \sim \mathcal{N}\left(0, \sigma^2\right),$$

with x_i and y_i independent.

Thus the detection problem asks whether the responses carry any signal-dependent correlation with the covariates, or whether they are pure independent noise.

This example already exhibits a phenomenon that will recur throughout the notes: the same statistical model gives rise to multiple tasks of different difficulty. Recovery and detection are not interchangeable questions.

1.6 Canonical Example II: Planted Clique

Our second example is graph-theoretic, and it will also play a central role throughout the notes.

We work on the space of simple undirected graphs on n labeled vertices. Let k be a positive integer, and let

$$\mu = \text{Unif}\left(\{S \subseteq [n] : |S| = k\}\right).$$

A random set $\theta \sim \mu$ is called the *hidden clique* or *planted clique*.

Given S, the graph $G \sim \mathop{\mathbb{P}}_{\theta}$ is generated as follows:

- every pair of vertices in S is connected by an edge;

- every other pair of vertices is connected independently with probability $1/2$.

Thus the induced subgraph on S is a complete graph, while the rest of the graph behaves as an Erdős-Rényi random graph with probability parameter $1/2$. The inference task is to recover S from the observed graph G.

Planted clique is a model of hidden dense structure inside random noise. Unlike sparse regression, where the signal is a vector in Euclidean space, here the signal is a subset of vertices. The model is discrete and combinatorial.

This problem will become one of the main running examples in the notes because it exhibits a striking gap between what is statistically possible and what is known algorithmically.

When it comes to the detection version, the following question arises:

Question 1.14. *What is a natural pure-noise distribution for the planted clique model?*

The answer is the Erdős-Rényi model

$$\mathbb{P}_2 = \mathcal{G}_{n,\frac{1}{2}},$$

namely the random graph obtained by including each edge independently with probability $1/2$. This is the appropriate null because it matches the background randomness of the planted model after removing the deterministic clique structure.

1.7 From Statistical Formulation to Phase Transitions

One of the striking features of high-dimensional inference is the emergence of sharp transitions. As a parameter such as the sample size, sparsity level, or signal-to-noise ratio changes, the behavior of the problem can shift abruptly from impossible to possible.

Already from the examples above, one should expect threshold phenomena of the following kind:

- below a critical regime, no method can recover or detect the signal;
- above that regime, an appropriate procedure succeeds with high probability.

These thresholds may differ depending on the task. Understanding this layered picture is one of the main goals of the text.

Chapter 2

Optimal Procedures for Exact Recovery, Detection, and Estimation

2.1 Introduction

In Chapter 1 we introduced the general Bayesian framework

$$\theta \sim \mu, \qquad Y \sim \mathop{\mathbb{P}}_{\theta},$$

together with the three main inference tasks that will be studied throughout these notes: exact recovery, approximate recovery and detection. The purpose of the present chapter is to identify the statistically optimal procedures associated with these tasks.

At this stage, as in the previous chapter, we impose no computational restriction on the inference procedure. Our goal is instead to determine what an optimal algorithm would do if computational resources were irrelevant. These optimal procedures will later serve as information-theoretic benchmarks against which efficient algorithms can be compared.

There are three canonical objects that arise in this way. For exact recovery, the optimal estimator is the maximum a posteriori estimator, or MAP estimator. For detection, the optimal test is given by the Neyman-Pearson rule, which compares the two likelihoods pointwise. For approximate recovery under mean-squared error, the optimal estimator is the posterior mean. The first two are developed in this chapter, while the third will play a central role in Chapter 4. We nevertheless state it here already, since it completes the parallel between the three inference tasks.

2.2 The posterior distribution

The notion of optimality in Bayesian inference is expressed through the posterior distribution of the hidden parameter given the data.

Definition 2.1 (Posterior distribution)**.** *Let $\theta \sim \mu$ and $Y \sim \mathbb{P}_\theta$. The* posterior distribution *of θ given Y is the conditional law of θ given the observation Y.*

Formally, if the model is discrete, then for each $\nu \in \mathcal{S}$ we have

$$\mathop{\mathbb{P}}_{\theta|Y}(\nu \mid Y) = \frac{\mathop{\mathbb{P}}_{\theta,Y}(\nu, Y)}{\mathbb{P}(Y)} = \frac{\mu(\nu)\,\mathop{\mathbb{P}}_{\nu}(Y)}{\mathbb{P}(Y)}.$$

Thus, up to the normalizing factor $\mathbb{P}(Y)$, the posterior mass assigned to a candidate ν is proportional to

$$\mu(\nu)\mathop{\mathbb{P}}_{\nu}(Y).$$

This expression is fundamental. It shows that the posterior combines two sources of information: the prior weight $\mu(\nu)$ assigned to the candidate ν, and the likelihood $\mathbb{P}_\nu(Y)$ of the observed data under that candidate. In other words, the posterior favors parameters that are both plausible *a priori* and statistically compatible with the data.

2.3 Exact recovery and the MAP estimator

We now return to the problem of exact recovery. Recall that the goal is to construct an algorithm

$$\mathcal{A} : \Omega \to \mathcal{S}$$

such that

$$\mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}\big(\mathcal{A}(Y) = \theta\big)$$

is as large as possible.

Definition 2.2 (MAP estimator)**.** *The* maximum a posteriori estimator*, denoted by $\mathcal{A}_{\mathrm{MAP}}$, is defined by*

$$\mathcal{A}_{\mathrm{MAP}}(Y) \in \arg\max_{\nu\in\mathcal{S}} \left\{ \mathop{\mathbb{P}}_{\theta|Y}(\nu \mid Y) \right\}.$$

Equivalently, by Bayes' rule,

$$\mathcal{A}_{\mathrm{MAP}}(Y) \in \arg\max_{\nu\in\mathcal{S}} \left\{ \mu(\nu)\, \mathop{\mathbb{P}}_{\nu}(Y) \right\}.$$

The definition is natural: after observing Y, the MAP estimator outputs a parameter value with maximal posterior probability. Thus, among all possible candidates, it chooses the one that is most likely to be the true signal in light of the data.

The key point is that this intuitive rule is in fact optimal for exact recovery.

Theorem 2.3. *The MAP estimator is the information-theoretically optimal estimator for exact recovery. More precisely, for every measurable algorithm $\mathcal{A} : \Omega \to \mathcal{S}$,*

$$\mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}\big(\mathcal{A}(Y) = \theta\big) \le \mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}\big(\mathcal{A}_{\mathrm{MAP}}(Y) = \theta\big).$$

Proof. Consider any algorithm $\mathcal{A}$. Its probability of success equals to

$$\begin{aligned}
\mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}(\mathcal{A}(Y) = \theta) &= \sum_{(\theta_0,Y_0)\,:\,\mathcal{A}(Y_0)=\theta_0} \mathop{\mathbb{P}}_{\theta,Y}(\theta_0, Y_0) \\
&= \sum_{Y_0} \mathop{\mathbb{P}}_{\theta,Y}(\mathcal{A}(Y_0), Y_0) \\
&= \sum_{Y_0} \mathop{\mathbb{P}}_{Y}(Y_0)\, \mathop{\mathbb{P}}_{\theta|Y}(\mathcal{A}(Y_0) \mid Y_0) \\
&\le \sum_{Y_0} \mathop{\mathbb{P}}_{Y}(Y_0) \max_{\nu}\, \mathop{\mathbb{P}}_{\theta|Y}(\nu \mid Y_0) \\
&= \sum_{Y_0} \mathop{\mathbb{P}}_{Y}(Y_0)\, \mathop{\mathbb{P}}_{\theta|Y}(\mathcal{A}_{\mathrm{MAP}}(Y_0) \mid Y_0) \\
&= \mathop{\mathbb{P}}_{\theta,Y}\left(\mathcal{A}_{\mathrm{MAP}}(Y) = \theta\right),
\end{aligned}$$

Equality holds if and only if for any Y_0, $\mathcal{A}(Y_0) \in \arg\max_{\nu} \mathbb{P}_{\theta|Y}(v \mid Y_0)$, i.e., $\mathcal{A}(Y) = \mathcal{A}_{\mathrm{MAP}}(Y)$. □

The theorem has an important conceptual consequence. If exact recovery is impossible for the MAP estimator, then it is impossible for every estimator. Thus the MAP rule is the correct benchmark for information-theoretic exact recovery.

2.4 Detection and the Neyman-Pearson test

We now turn to detection. Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability distributions on Ω, where $\mathbb{P}_1$ typically represents structured reality and $\mathbb{P}_2$ represents pure noise. The task is to construct a test

$$\mathcal{T} : \Omega \to \{\mathbb{P}_1, \mathbb{P}_2\}$$

that minimizes the two-sided error

$$\mathrm{err}(\mathcal{T}) := \mathbb{P}_{Y\sim\mathbb{P}_1}\big(\mathcal{T}(Y) = \mathbb{P}_2\big) + \mathbb{P}_{Y\sim\mathbb{P}_2}\big(\mathcal{T}(Y) = \mathbb{P}_1\big).$$

The optimal test is given by a pointwise comparison of the two likelihoods.

Theorem 2.4 (Neyman-Pearson)**.** *An optimal test is obtained by declaring $\mathbb{P}_1$ whenever*

$$\mathbb{P}_1(Y) \geq \mathbb{P}_2(Y),$$

and declaring $\mathbb{P}_2$ otherwise. Equivalently,

$$\mathcal{T}_{\mathrm{NP}}(Y) = \mathbb{P}_1 \quad \Longleftrightarrow \quad \mathbb{P}_1(Y) \geq \mathbb{P}_2(Y)$$

Proof. The proof will be given for the discrete setting for simplicity. It holds

$$\begin{aligned}
\mathrm{err}(\mathcal{A}) &= \mathbb{P}_{Y\sim\mathbb{P}_1}(\mathcal{A}(Y) = \mathbb{P}_2) + \mathbb{P}_{Y\sim\mathbb{P}_2}(\mathcal{A}(Y) = \mathbb{P}_1) \\
&= \mathbb{P}_{Y\sim\mathbb{P}_1}(\mathcal{A}(Y) = \mathbb{P}_2) + 1 - \mathbb{P}_{Y\sim\mathbb{P}_2}(\mathcal{A}(Y) = \mathbb{P}_2) \\
&= \sum_{Y\,:\,\mathcal{A}(Y)=\mathbb{P}_2} (\mathbb{P}_1(Y) - \mathbb{P}_2(Y)) + 1.
\end{aligned}$$

To minimize the last quantity it suffices to set $\mathcal{A}(Y) = \mathbb{P}_2 \Leftrightarrow \mathbb{P}_1(Y) - \mathbb{P}_2(Y) \leq 0$. Thus, our algorithm always gives the minimum possible error and the proof is complete. □

This test is the exact analogue, for detection, of the MAP estimator for exact recovery. Both are obtained by selecting the hypothesis that is most plausible after observing the data. In the detection setting, however, the decision is between two distributions rather than between many possible parameter values.

2.5 Posterior mean and the mean-squared error benchmark

Although the main focus of this chapter has been exact recovery and detection, it is useful to record already the corresponding optimal procedure for mean-squared error.

Suppose now that θ takes values in $\mathbb{R}^d$, and consider the problem of minimizing

$$\mathrm{MSE}(\mathcal{A}) = \mathbb{E}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}\left[\|\mathcal{A}(Y) - \theta\|_2^2\right].$$

Theorem 2.5. *The estimator minimizing the mean-squared error is the posterior mean:*

$$\mathcal{A}(Y) = \mathop{\mathbb{E}}_{\theta|Y}[\theta \mid Y].$$

Proof. For any algorithm $\mathcal{A}$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{A}(Y) - \theta\|_2^2\right] &= \mathbb{E}\left[\|\mathcal{A}(Y) - \mathbb{E}[\theta \mid Y] + \mathbb{E}[\theta \mid Y] - \theta\|_2^2\right] \\
&= \mathbb{E}\left[\| \underbrace{\mathcal{A}(Y) - \mathbb{E}[\theta \mid Y]}_{=:B(Y)} \|_2^2\right] + \mathbb{E}\left[\|\mathbb{E}[\theta \mid Y] - \theta\|_2^2\right] \\
&+ 2\mathbb{E}\left[\langle \mathcal{A}(Y) - \mathbb{E}[\theta \mid Y], \mathbb{E}[\theta \mid Y] - \theta \rangle\right] \\
&\geq \mathbb{E}\left[\|\mathbb{E}[\theta \mid Y] - \theta\|_2^2\right] + 2\mathbb{E}\left[\langle B(Y), \mathbb{E}[\theta \mid Y] - \theta \rangle\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[(\mathbb{E}[\theta_i \mid Y] - \theta_i)^2\right] + 2\sum_{i=1}^{d} \mathbb{E}[B(Y)_i(\mathbb{E}[\theta_i \mid Y] - \theta_i)] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[(\mathbb{E}[\theta_i \mid Y] - \theta_i)^2\right] + 2\sum_{i=1}^{d} \mathop{\mathbb{E}}_{Y}\left[B(Y)_i \mathop{\mathbb{E}}_{\theta_i|Y}[\mathbb{E}[\theta_i \mid Y] - \theta_i]\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[(\mathbb{E}[\theta_i \mid Y] - \theta_i)^2\right] + 2\sum_{i=1}^{d} \mathop{\mathbb{E}}_{Y}\left[B(Y)_i \mathop{\mathbb{E}}_{\theta_i|Y}[(\mathbb{E}[\theta_i \mid Y] - \theta_i) \mid Y]\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[(\mathbb{E}[\theta_i \mid Y] - \theta_i)^2\right] + 2\sum_{i=1}^{d} \mathop{\mathbb{E}}_{Y}[B(Y)_i(\mathbb{E}[\theta_i \mid Y] - \mathbb{E}[\theta_i \mid Y])] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[(\mathbb{E}[\theta_i \mid Y] - \theta_i)^2\right]
\end{aligned}$$

□

This theorem is the exact analogue, for approximate recovery under quadratic loss, of the optimality results proved above for exact recovery and detection. Together, the MAP estimator, the Neyman-Pearson test, and the posterior mean form the three canonical Bayesian benchmarks that will guide the rest of these notes.

2.6 Discussion

The main message of this chapter is that, once the statistical model is fixed, the correct information-theoretic benchmark is usually canonical. For exact recovery, one should study the MAP estimator. For detection, one should study the Neyman-Pearson test. For quadratic estimation, one should study the posterior mean.

The difficulty is therefore not conceptual but computational. These procedures are easy to define, yet in many high-dimensional problems they involve posterior maximization or averaging over a very large space of candidate signals. Later in the notes, this will be precisely the source of the gap between what is statistically possible and what is achievable by time-efficient algorithms.

In the next chapter, we turn from optimal procedures to statistical thresholds. Once the correct benchmark has been identified, the next question is to determine when it succeeds and when it fails.

Chapter 3

Information-Theoretic Thresholds for Exact Recovery and Detection

3.1 Introduction

In Chapter 2 we identified the canonical statistical procedures associated with the main inference tasks. For exact recovery, the relevant benchmark is the MAP estimator $\mathcal{A}_{\mathrm{MAP}}$; for detection, it is the Neyman-Pearson test $\mathcal{T}_{\mathrm{NP}}$. Once these optimal procedures have been identified, the next natural question is when they succeed.

There are two guiding principles throughout the chapter. The first is that, for exact recovery, one may restrict attention to the MAP estimator. The second is that, for detection, one may restrict attention to the Neyman-Pearson test, whose performance is governed by the total variation distance between the two competing distributions. In practice, total variation is often difficult to compute directly, so one uses more tractable divergences such as the Kullback-Leibler divergence and the χ^2-divergence.

We will illustrate these ideas on two basic examples. The first is a simple Gaussian signal-plus-noise model, where the exact recovery and detection thresholds can be computed explicitly. The second is the planted clique problem, where exact recovery is possible information-theoretically at clique size of order $\log n$.

3.2 A Gaussian signal-plus-noise model

We begin with the simplest nontrivial example. Let

$$\mu = \mathrm{Unif}\left(\{\mathbf{1}_d, -\mathbf{1}_d\}\right),$$

where $\mathbf{1}_d \in \mathbb{R}^d$ denotes the all-ones vector, and suppose that we observe

$$Y = \lambda\theta + Z, \qquad Z \sim \mathcal{N}(0, I_d).$$

The parameter $\lambda > 0$ plays the role of a signal-to-noise ratio, and may depend on d. The problem is to determine for which values of $\lambda = \lambda_d$ exact recovery is possible.

Computation of the posterior

Fix $\theta_0 \in \{\mathbf{1}_d, -\mathbf{1}_d\}$. Since the prior is uniform, Bayes' rule gives

$$\mathop{\mathbb{P}}_{\theta|Y}(\theta_0 \mid Y) \propto \mathop{\mathbb{P}}_{\theta_0}(Y).$$

Under $\theta = \theta_0$, the random vector Y has density proportional to

$$\exp\left(-\frac{1}{2}\|Y - \lambda\theta_0\|_2^2\right).$$

Expanding the square,

$$\|Y - \lambda\theta_0\|_2^2 = \|Y\|_2^2 + \lambda^2\|\theta_0\|_2^2 - 2\lambda\langle\theta_0, Y\rangle.$$

Since $\|\theta_0\|_2^2 = d$, we obtain

$$\mathop{\mathbb{P}}_{\theta|Y}(\theta_0 \mid Y) \propto \exp\left(-\frac{1}{2}\|Y\|_2^2 - \frac{\lambda^2 d}{2} + \lambda\langle\theta_0, Y\rangle\right) \propto \exp\big(\lambda\langle\theta_0, Y\rangle\big).$$

Therefore

$$\mathcal{A}_{\mathrm{MAP}}(Y) \in \arg\max_{\nu \in \{\mathbf{1}_d, -\mathbf{1}_d\}} \{\langle\nu, Y\rangle\}.$$

Thus the MAP estimator simply chooses the sign whose correlation with the observation is larger.

Exact recovery threshold

Suppose that the true signal is θ. Then

$$Y = \lambda\theta + Z,$$

so

$$\langle\theta, Y\rangle = \lambda\langle\theta, \theta\rangle + \langle\theta, Z\rangle = \lambda d + \langle\theta, Z\rangle.$$

Moreover,

$$\langle\theta, Z\rangle \overset{d}{=} \sqrt{d}X, \qquad X \sim \mathcal{N}(0, 1).$$

Since $\mathcal{A}_{\mathrm{MAP}}$ outputs θ if and only if

$$\langle\theta, Y\rangle > \langle-\theta, Y\rangle,$$

it follows that

$$\mathcal{A}_{\mathrm{MAP}}(Y) = \theta \quad \Longleftrightarrow \quad \langle\theta, Y\rangle > 0.$$

Hence

$$\mathop{\mathbb{P}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta}(\mathcal{A}_{\mathrm{MAP}}(Y) = \theta) = \mathop{\mathbb{P}}_{X\sim\mathcal{N}(0,1)}\left(X > -\lambda\sqrt{d}\right).$$

We have therefore proved the following.

Proposition 3.1. *In the Gaussian signal-plus-noise model*

$$Y = \lambda\theta + Z, \qquad \theta \sim \mathrm{Unif}\left(\{\mathbf{1}_d, -\mathbf{1}_d\}\right), \qquad Z \sim \mathcal{N}(0, I_d),$$

exact recovery with high probability is possible if and only if

$$\lambda\sqrt{d} \to \infty.$$

Equivalently, the information-theoretic exact recovery threshold is of order

$$\lambda_c \asymp \frac{1}{\sqrt{d}}.$$

This is a useful toy example. The posterior can be computed explicitly, the MAP estimator has a simple geometric interpretation, and the recovery threshold is determined by the competition between the deterministic term λd and the Gaussian fluctuation of size $\sqrt{d}$.

3.3 Planted clique and uniqueness of the posterior maximizer

We now turn to the planted clique model introduced in Chapter 1, where

$$\mu = \mathrm{Unif}\left(\{S \subseteq [n] : |S| = k\}\right),$$

and for $\theta \sim \mu$, the graph $G \sim \mathop{\mathbb{P}}_{\theta}$ is generated by planting a k-clique on the vertex set θ, while every other edge is present independently with probability $1/2$.

The exact recovery problem is to reconstruct the planted clique θ from the observed graph G.

Posterior structure

Fix a k-subset $C \subseteq [n]$. MAP gives

$$\begin{aligned}
\mathop{\mathbb{P}}_{PC|G}(PC = C|G) &\propto \mathop{\mathbb{P}}_{PC,G}(PC = C, G) \\
&= \mathop{\mathbb{P}}_{PC}(C) \mathop{\mathbb{P}}_{PC|G}(G \mid C) \\
&= \frac{1}{\binom{n}{k}} \mathop{\mathbb{P}}_{PC|G}(G \mid C) \\
&= \frac{1}{\binom{n}{k}} 2^{-\binom{n}{2}+\binom{k}{2}} \cdot \mathbf{1}_{\{C \text{ is a } k\text{-clique in } G\}} \\
&\propto \mathbf{1}_{\{C \text{ is a } k\text{-clique in } G\}}.
\end{aligned}$$

Thus, the posterior is uniform among all k-vertex cliques in G. If G has more than one k-clique we get

$$\mathbb{P}\left(\mathcal{A}(G) = PC|G\right) \leq 1/2,$$

and we cannot achieve exact estimation with high probability.

Consequently, exact recovery is possible if and only if the planted clique is the unique k-clique with high probability. Using this the following theorem can be relatively easily proven via a second moment method analysis (see [AS92] for a canonical introduction to the second moment method).

Theorem 3.2. *Let $G \sim \mathcal{G}_{n,1/2,k}$, and let $P \subseteq V(G)$ denote the planted clique. For every fixed $\varepsilon \in (0, 2)$, the following hold as $n \to \infty$:*

1. *If*
$$k \geq (2 + \varepsilon) \log_2 n,$$
then P is the unique k-clique in G with high probability.
2. *If*
$$k \leq (2 - \varepsilon) \log_2 n,$$
then P is not the unique k-clique in G with high probability.

The theorem, the proof of which can be found in Section A.1, shows that the information-theoretic threshold for exact recovery in planted clique is of logarithmic order. As we are going to see later, this is one of the first places where the difference between statistical feasibility and algorithmic feasibility becomes visible for such problems: exact recovery is possible information-theoretically as soon as k is slightly larger than $2 \log_2 n$, yet as we are going to discuss no polynomial-time algorithm is known in that regime.

3.4 Detection and total variation distance

We now move to detection. Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two distributions on the observation space Ω, where $\mathbb{P}_1$ represents structured reality and $\mathbb{P}_2$ represents pure noise. A test is a measurable map

$$\mathcal{T} : \Omega \to \{\mathbb{P}_1, \mathbb{P}_2\},$$

and its two-sided error is

$$\mathrm{err}(\mathcal{T}) = \mathop{\mathbb{P}}_{Y\sim\mathbb{P}_1}(\mathcal{T}(Y) = \mathbb{P}_2) + \mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(\mathcal{T}(Y) = \mathbb{P}_1).$$

Chapter 2 showed that the Neyman-Pearson test minimizes this quantity. In order to determine when detection is possible, we therefore need to compute the minimum possible error.

Definition 3.3. *The total variation distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ is*

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) = \frac{1}{2}\sum_{Y\in\Omega} |\mathbb{P}_1(Y) - \mathbb{P}_2(Y)|$$

in the discrete setting.

A standard calculation gives the exact relation between total variation distance and optimal detection error.

Proposition 3.4. *The minimum of $\mathrm{err}(\mathcal{T})$ over all tests $\mathcal{T}$ is*

$$1 - \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2).$$

Proof. Let $\mathcal{T}$ be any test. Then

$$\mathrm{err}(\mathcal{T}) = \mathop{\mathbb{P}}_{Y\sim\mathbb{P}_1}(\mathcal{T}(Y) = \mathbb{P}_2) + \mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(\mathcal{T}(Y) = \mathbb{P}_1).$$

Since

$$\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(\mathcal{T}(Y) = \mathbb{P}_1) = 1 - \mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(\mathcal{T}(Y) = \mathbb{P}_2),$$

we may rewrite

$$\mathrm{err}(\mathcal{T}) = 1 + \sum_{Y:\,\mathcal{T}(Y)=\mathbb{P}_2} (\mathbb{P}_1(Y) - \mathbb{P}_2(Y)).$$

This quantity is minimized by choosing $\mathcal{T}(Y) = \mathbb{P}_2$ exactly when

$$\mathbb{P}_1(Y) - \mathbb{P}_2(Y) \le 0,$$

that is, exactly when $\mathbb{P}_2(Y) \ge \mathbb{P}_1(Y)$. Therefore

$$\min_{\mathcal{T}}\{\mathrm{err}(\mathcal{T})\} = 1 - \sum_{Y:\,\mathbb{P}_1(Y)\ge\mathbb{P}_2(Y)} (\mathbb{P}_1(Y) - \mathbb{P}_2(Y)).$$

Since

$$\sum_{Y\in\Omega} (\mathbb{P}_1(Y) - \mathbb{P}_2(Y)) = 0,$$

the last sum equals

$$\frac{1}{2}\sum_{Y\in\Omega} |\mathbb{P}_1(Y) - \mathbb{P}_2(Y)| = \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2),$$

which proves the claim. □

This proposition shows that the total variation distance fully characterizes the information-theoretic difficulty of detection. In particular, if $\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) \to 0$, then every test has error tending to 1, and even weak detection is impossible.

3.5 KL divergence and χ^2-divergence

Although total variation is the right quantity conceptually, it is often difficult to compute directly. It is therefore useful to introduce other distances between probability distributions that can be easier to evaluate.

Assume that $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_2$. The Kullback-Leibler divergence is defined by

$$\mathrm{KL}(\mathbb{P}_1\|\mathbb{P}_2) := \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}\left[\log\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}\right],$$

and the χ^2-divergence is defined by

$$\chi^2(\mathbb{P}_1,\mathbb{P}_2) := \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[\left(\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}\right)^2\right] - 1.$$

The following inequalities relate these quantities.

Theorem 3.5. *For any two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$,*

$$\mathrm{TV}(\mathbb{P}_1,\mathbb{P}_2) \le \sqrt{\mathrm{KL}(\mathbb{P}_1\|\mathbb{P}_2)} \le \sqrt{\chi^2(\mathbb{P}_1,\mathbb{P}_2)}.$$

Proof. Let

$$L(Y) := \frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}$$

denote the likelihood ratio. Then

$$\mathrm{KL}(\mathbb{P}_1\|\mathbb{P}_2) = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}\left[\log\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}\right] = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[L(Y)\log L(Y)\right],$$

and

$$\chi^2(\mathbb{P}_1,\mathbb{P}_2) = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[L(Y)^2\right] - 1.$$

We first prove that

$$\mathrm{KL}(\mathbb{P}_1\|\mathbb{P}_2) \le \chi^2(\mathbb{P}_1,\mathbb{P}_2).$$

Using the elementary inequality

$$\log x \le x - 1, \qquad x > 0,$$

we obtain

$$L\log L \le L(L-1) = L^2 - L.$$

Taking expectation with respect to $\mathbb{P}_2$, and using

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[L(Y)] = 1,$$

we get

$$\mathrm{KL}(\mathbb{P}_1\|\mathbb{P}_2) = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[L\log L] \le \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[L^2 - L] = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[L^2] - 1 = \chi^2(\mathbb{P}_1,\mathbb{P}_2).$$

We now prove that

$$\mathrm{TV}(\mathbb{P}_1,\mathbb{P}_2) \le \sqrt{\mathrm{KL}(\mathbb{P}_1\|\mathbb{P}_2)}.$$

Recall that

$$\mathrm{TV}(\mathbb{P}_1,\mathbb{P}_2) = \frac{1}{2}\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[|L(Y)-1|\right].$$

We use the inequality

$$x \log x - x + 1 \geq \frac{(x-1)^2}{2(x+1)}, \qquad x > 0.$$

Therefore

$$\mathrm{KL}(\mathbb{P}_1 \| \mathbb{P}_2) = \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[L \log L] = \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[L \log L - L + 1] \geq \frac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}\left[\frac{(L-1)^2}{L+1}\right].$$

On the other hand, by Cauchy-Schwarz,

$$\left(\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[|L-1|]\right)^2 = \left(\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}\left[\frac{|L-1|}{\sqrt{L+1}}\sqrt{L+1}\right]\right)^2 \leq \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}\left[\frac{(L-1)^2}{L+1}\right] \cdot \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[L+1].$$

Since $\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[L] = 1$, we have

$$\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[L+1] = 2,$$

so

$$\left(\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[|L-1|]\right)^2 \leq 2 \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}\left[\frac{(L-1)^2}{L+1}\right].$$

Combining this with the previous bound gives

$$\left(\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[|L-1|]\right)^2 \leq 4\,\mathrm{KL}(\mathbb{P}_1 \| \mathbb{P}_2).$$

Multiplying by $1/4$, we conclude that

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2)^2 = \frac{1}{4}\left(\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2}[|L-1|]\right)^2 \leq \mathrm{KL}(\mathbb{P}_1 \| \mathbb{P}_2).$$

Hence

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) \leq \sqrt{\mathrm{KL}(\mathbb{P}_1 \| \mathbb{P}_2)}.$$

Putting the two inequalities together, we obtain

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) \leq \sqrt{\mathrm{KL}(\mathbb{P}_1 \| \mathbb{P}_2)} \leq \sqrt{\chi^2(\mathbb{P}_1, \mathbb{P}_2)}.$$

□

The role of these inequalities is straightforward. If one can show that the χ^2-divergence tends to zero, then the total variation distance also tends to zero, and detection is impossible.

3.6 Detection in the Gaussian signal-plus-noise model

We now revisit the Gaussian signal-plus-noise model, this time as a detection problem. Consider

$$\mathbb{P}_1 : \quad Y = \lambda \nu + Z, \qquad \nu \sim \mathrm{Unif}\left(\{\mathbf{1}_d, -\mathbf{1}_d\}\right), \qquad Z \sim \mathcal{N}(0, I_d),$$

and

$$\mathbb{P}_2 : \quad Y = Z, \qquad Z \sim \mathcal{N}(0, I_d).$$

The question is to determine for which values of $\lambda = \lambda_d$ one can strongly detect between $\mathbb{P}_1$ and $\mathbb{P}_2$.

Computation of the likelihood ratio

Under $\mathbb{P}_2$, the density of Y is proportional to

$$\exp\left(-\frac{1}{2}\|Y\|_2^2\right).$$

Under $\mathbb{P}_1$, the distribution is the equally weighted mixture of the two shifted Gaussians centered at $\lambda \mathbf{1}_d$ and $-\lambda \mathbf{1}_d$. Therefore

$$\begin{aligned}
\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)} &= \frac{\frac{1}{2}\mathbb{P}(Y = \lambda\mathbf{1}_d + Z) + \frac{1}{2}\mathbb{P}(Y = -\lambda\mathbf{1}_d + Z)}{\exp\left(-\frac{1}{2}\|Y\|_2^2\right)} \\
&= \frac{1}{2}\frac{\exp\left(-\|Y - \lambda\mathbf{1}_d\|^2/2\right) + \exp\left(-\|Y + \lambda\mathbf{1}_d\|^2/2\right)}{\exp\left(-\frac{1}{2}\|Y\|_2^2\right)} \\
&= \frac{1}{2}\frac{\exp\left(-\frac{1}{2}\|Y\|_2^2 - \lambda^2 d/2 + \lambda\langle Y, \mathbf{1}_d\rangle\right) + \exp\left(-\frac{1}{2}\|Y\|_2^2 - \lambda^2 d/2 - \lambda\langle Y, \mathbf{1}_d\rangle\right)}{\exp\left(-\frac{1}{2}\|Y\|_2^2\right)} \\
&= \frac{1}{2}\left(e^{-\lambda^2 d/2 + \lambda\langle Y, \mathbf{1}_d\rangle} + e^{-\lambda^2 d/2 - \lambda\langle Y, \mathbf{1}_d\rangle}\right) \\
&= \frac{e^{-\lambda^2 d/2}}{2}\left(e^{\lambda\langle Y, \mathbf{1}_d\rangle} + e^{-\lambda\langle Y, \mathbf{1}_d\rangle}\right).
\end{aligned}$$

The χ^2-divergence

Squaring and taking expectation under $\mathbb{P}_2$, we obtain

$$\begin{aligned}
\chi^2\left(\mathbb{P}_1, \mathbb{P}_2\right) &= \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[\left(\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}\right)^2\right] - 1 \\
&= \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(0, I_d)}\left[\frac{e^{-\lambda^2 d}}{4}\left(e^{2\lambda\langle Y, \mathbf{1}_d\rangle} + e^{-2\lambda\langle Y, \mathbf{1}_d\rangle} + 2\right)\right] - 1 \\
&= \frac{e^{-\lambda^2 d}}{4}\left(2\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(0, I_d)}\left[e^{2\lambda\langle Y, \mathbf{1}_d\rangle}\right] + 2\right) - 1 \\
&= \frac{e^{-\lambda^2 d}}{4}\left(2e^{2\lambda^2 d} + 2\right) - 1 \\
&= \mathcal{O}\left(e^{\lambda^2 d}\right) - 1.
\end{aligned}$$

In particular, if

$$\lambda \ll \frac{1}{\sqrt{d}},$$

then

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) \to 0,$$

and therefore

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) \to 0.$$

Hence weak detection is impossible in that regime.

A matching upper bound

We now show that if

$$\lambda \gg \frac{1}{\sqrt{d}},$$

then strong detection is possible.

Consider the test $\mathcal{T}$ that declares $\mathbb{P}_1$ if either

$$\langle Y, \mathbf{1}_d \rangle \geq \frac{\lambda d}{2} \qquad \text{or} \qquad \langle Y, -\mathbf{1}_d \rangle \geq \frac{\lambda d}{2},$$

and declares $\mathbb{P}_2$ otherwise.

Under $\mathbb{P}_1$, we have $Y = \lambda\nu + Z$ for some $\nu \in \{\mathbf{1}_d, -\mathbf{1}_d\}$, so

$$\langle Y, \nu \rangle = \lambda d + \langle Z, \nu \rangle.$$

Since $\langle Z, \nu \rangle$ is Gaussian of variance d, it is of order $\sqrt{d}$ with high probability. If $\lambda d \gg \sqrt{d}$, equivalently $\lambda \gg d^{-1/2}$, then

$$\langle Y, \nu \rangle \geq \frac{\lambda d}{2}$$

with high probability. Thus the type-I error tends to zero.

Under $\mathbb{P}_2$, both $\langle Y, \mathbf{1}_d \rangle$ and $\langle Y, -\mathbf{1}_d \rangle$ are centered Gaussians with variance d, so both are of order $\sqrt{d}$ with high probability. If $\lambda d \gg \sqrt{d}$, then neither exceeds $\lambda d/2$ with high probability. Thus the type-II error also tends to zero.

We have therefore established the following threshold.

Proposition 3.6. *In the Gaussian signal-plus-noise model*

$$\mathbb{P}_1 : \quad Y = \lambda\nu + Z, \qquad \nu \sim \mathrm{Unif}\{\mathbf{1}_d, -\mathbf{1}_d\}, \qquad Z \sim \mathcal{N}(0, I_d),$$

versus

$$\mathbb{P}_2 : \quad Y = Z, \qquad Z \sim \mathcal{N}(0, I_d),$$

the information-theoretic threshold for detection is of order

$$\lambda_c \asymp \frac{1}{\sqrt{d}}.$$

More precisely, if $\lambda \ll d^{-1/2}$, then even weak detection is impossible, while if $\lambda \gg d^{-1/2}$, then strong detection is possible.

It is worth emphasizing that in this model the exact recovery and detection thresholds coincide, both being of order $d^{-1/2}$. This is not a universal phenomenon, but it provides a useful first benchmark.

3.7 Discussion

The results of this chapter establish the basic information-theoretic picture for exact recovery and detection.

For exact recovery, the relevant object is the posterior maximizer. In simple Gaussian models this leads to explicit threshold calculations, while in combinatorial models such as planted clique it leads to random graphs questions about uniqueness of the clique. For detection, the Neyman-Pearson test reduces the problem to comparing two distributions, and the total variation distance becomes the central quantity. Since total variation is often difficult to handle directly, the χ^2-divergence and the Kullback-Leibler divergence provide useful upper bounds.

Chapter 4

Information-Theoretic Thresholds for Approximate Recovery

4.1 Introduction

In the previous chapters, we studied exact recovery and detection. Both tasks are naturally governed by canonical statistical procedures: the MAP estimator in the first case and the Neyman-Pearson test in the second. Approximate recovery is different. Here the goal is not to identify the signal exactly, but rather to estimate it as accurately as possible under a quantitative loss function.

4.2 The mean-squared error and the posterior mean

Let

$$\theta \sim \mu, \qquad Y \sim \mathbb{P}_\theta,$$

where now θ takes values in $\mathbb{R}^d$. For an estimator

$$\mathcal{A} : \Omega \to \mathbb{R}^d,$$

we defined its mean-squared error by

$$\mathrm{MSE}(\mathcal{A}) := \mathop{\mathbb{E}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta} \left[\|\mathcal{A}(Y) - \theta\|_2^2\right]$$

and proved

$$\mathrm{MMSE} := \inf_{\mathcal{A}} \{\mathrm{MSE}(\mathcal{A})\}$$

is achieved by $\mathop{\mathbb{E}}_{\theta|Y}[\theta \mid Y]$, i.e.

$$\mathcal{A}_{\mathrm{PM}} := \mathop{\mathbb{E}}_{\theta|Y}[\theta \mid Y] \in \arg\inf_{\mathcal{A}} \{\mathrm{MSE}(\mathcal{A})\}$$

Besides the optimal estimator, it is useful to record a trivial benchmark. The estimator that ignores the data altogether and always returns the prior mean,

$$\mathcal{A}_{\mathrm{trivial}}(Y) := \mathop{\mathbb{E}}_{\theta\sim\mu}[\theta],$$

satisfies

$$\mathrm{MSE}(\mathcal{A}_{\mathrm{triv}}) = \mathop{\mathbb{E}}_{\theta\sim\mu}\left[\left\|\theta - \mathop{\mathbb{E}}_{\theta\sim\mu}[\theta]\right\|_2^2\right].$$

This is the error level one obtains from the prior alone. In many problems, the relevant question is not whether one can estimate the signal perfectly, but whether one can do substantially better than this trivial benchmark, this is why it is canonical to normalize with the performance of the trivial estimator $\mathcal{A}_{\mathrm{trivial}}$.

4.3 A two-point Gaussian model

We begin with a warm-up example in which the posterior mean can be computed explicitly. Let

$$\theta \sim \mathrm{Unif}\left(\{\mathbf{1}_d, -\mathbf{1}_d\}\right), \qquad Y = \lambda\theta + Z, \qquad Z \sim \mathcal{N}(0, I_d).$$

As in Chapter 3, $\lambda > 0$ plays the role of a signal-to-noise ratio.

Computation of the posterior mean

Since the prior is supported on the two points $\pm\mathbf{1}_d$, the posterior mean is

$$\mathop{\mathbb{E}}_{\theta|Y}[\theta \mid Y] = \mathbf{1}_d \mathop{\mathbb{P}}_{\theta|Y}(\mathbf{1}_d \mid Y) - \mathbf{1}_d \mathop{\mathbb{P}}_{\theta|Y}(-\mathbf{1}_d \mid Y).$$

Using the same calculation as in the exact recovery setting, one finds that

$$\mathop{\mathbb{P}}_{\theta|Y}(\pm\mathbf{1}_d \mid Y) = \frac{\exp\left(\pm\lambda\langle\mathbf{1}_d, Y\rangle\right)}{\exp\left(\lambda\langle\mathbf{1}_d, Y\rangle\right) + \exp\left(-\lambda\langle\mathbf{1}_d, Y\rangle\right)}.$$

Therefore

$$\mathcal{A}_{\mathrm{PM}}(Y) = \mathbf{1}_d \frac{\exp\left(\lambda\langle\mathbf{1}_d, Y\rangle\right) - \exp\left(-\lambda\langle\mathbf{1}_d, Y\rangle\right)}{\exp\left(\lambda\langle\mathbf{1}_d, Y\rangle\right) + \exp\left(-\lambda\langle\mathbf{1}_d, Y\rangle\right)} = \mathbf{1}_d \tanh\left(\lambda\langle\mathbf{1}_d, Y\rangle\right).$$

The corresponding MMSE

By symmetry, it is enough to condition on the event $\theta = \mathbf{1}_d$. In that case,

$$Y = \lambda\mathbf{1}_d + Z, \qquad Z \sim \mathcal{N}(0, I_d),$$

and hence

$$\lambda\langle\mathbf{1}_d, Y\rangle = \lambda^2 d + \lambda\langle\mathbf{1}_d, Z\rangle = \lambda^2 d + \lambda\sqrt{d}\,T, \qquad T \sim \mathcal{N}(0,1).$$

Therefore

$$\mathcal{A}_{\mathrm{PM}}(Y) = \mathbf{1}_d \tanh\left(\lambda^2 d + \lambda\sqrt{d}\,T\right),$$

and the minimum mean-squared error is

$$\begin{aligned}
\mathrm{MMSE} &= \frac{1}{2}\mathop{\mathbb{E}}_{Z}\left[\left\|\mathbf{1}_d - \mathbf{1}_d \frac{\exp\left(\lambda\langle\mathbf{1}_d, \lambda\mathbf{1}_d + Z\rangle\right) - \exp\left(-\lambda\langle\mathbf{1}_d, \lambda\mathbf{1}_d + Z\rangle\right)}{\exp\left(\lambda\langle\mathbf{1}_d, \lambda\mathbf{1}_d + Z\rangle\right) + \exp\left(-\lambda\langle\mathbf{1}_d, \lambda\mathbf{1}_d + Z\rangle\right)}\right\|_2^2 \,\middle|\, \theta = \mathbf{1}_d\right] \\
&\quad + \frac{1}{2}\mathop{\mathbb{E}}_{Z}\left[\left\|-\mathbf{1}_d - \mathbf{1}_d \frac{\exp\left(\lambda\langle\mathbf{1}_d, -\lambda\mathbf{1}_d + Z\rangle\right) - \exp\left(-\lambda\langle\mathbf{1}_d, -\lambda\mathbf{1}_d + Z\rangle\right)}{\exp\left(\lambda\langle\mathbf{1}_d, -\lambda\mathbf{1}_d + Z\rangle\right) + \exp\left(-\lambda\langle\mathbf{1}_d, -\lambda\mathbf{1}_d + Z\rangle\right)}\right\|_2^2 \,\middle|\, \theta = -\mathbf{1}_d\right] \\
&= d \mathop{\mathbb{E}}_{T\sim\mathcal{N}(0,1)}\left[\left(1 - \tanh\left(\lambda^2 d + \lambda\sqrt{d}\,T\right)\right)^2\right].
\end{aligned}$$

Since

$$\mathcal{A}_{\text{trivial}}(Y) \equiv 0, \qquad \mathrm{MSE}(\mathcal{A}_{\text{trivial}}) = d,$$

the MMSE expression shows explicitly how the nontrivial improvement over the trivial estimator depends on the effective parameter $\lambda\sqrt{d}$ (see Figure 4.1).

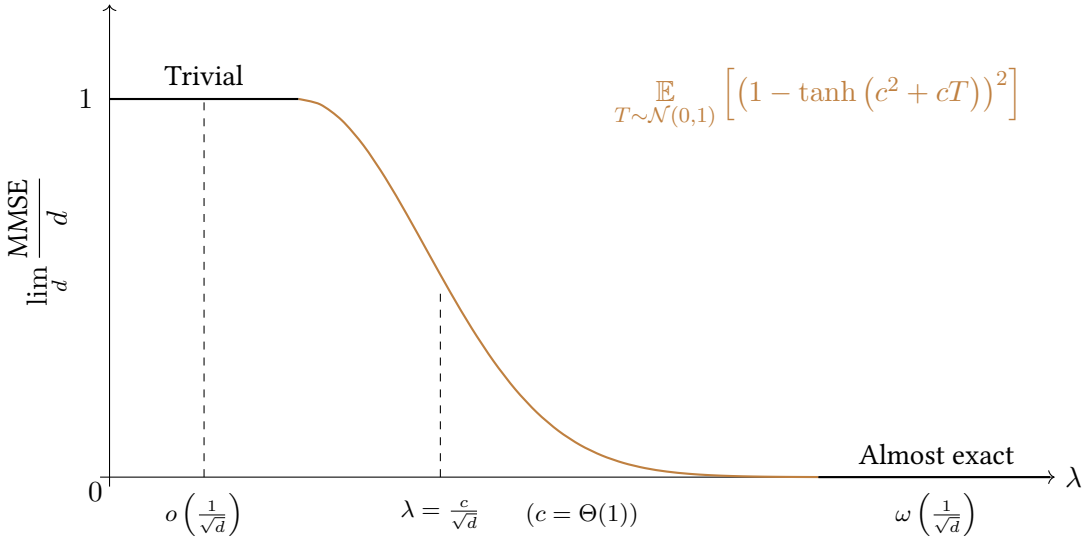

Figure 4.1: Normalized MMSE in the two-point Gaussian model

4.4 Reduction to a one-dimensional model

The preceding formula already suggests that the problem is effectively one-dimensional: only the scalar statistic $\langle \mathbf{1}_d, Y\rangle$ matters. This can be made exact.

Write

$$\theta = \theta_0 \mathbf{1}_d, \qquad \theta_0 \sim \mathrm{Unif}\left(\{-1,1\}\right),$$

and observe

$$Y_i = \lambda\theta_0 + Z_i, \qquad i = 1,\dots,d.$$

Define the normalized sum

$$y := \frac{1}{\sqrt{d}} \sum_{i=1}^{d} Y_i.$$

Then

$$y = \lambda\sqrt{d}\,\theta_0 + z, \qquad z \sim \mathcal{N}(0,1).$$

Thus the original d-dimensional model reduces to the scalar Gaussian channel

$$\theta_0 \sim \mathrm{Unif}\left(\{-1,1\}\right), \qquad y = \lambda\sqrt{d}\,\theta_0 + z.$$

The converse direction is also true: from this scalar channel one can reconstruct a full d-dimensional observation with the same statistical content.

Lemma 4.1 (Gaussian cloning lemma)**.** *Given an observation*

$$y = \lambda\sqrt{d}\,\theta_0 + z, \qquad z \sim \mathcal{N}(0,1),$$

one can generate random variables $Y_1,\dots,Y_d$ such that

$$Y_i = \lambda\theta_0 + Z_i, \qquad Z_1,\dots,Z_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Proof. Choose

$$t_2, \dots, t_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

and let

$$x = (y, t_2, \dots, t_d)^\top \in \mathbb{R}_d.$$

Then

$$x = \lambda\sqrt{d}\theta_0 e_1 + (z, t_2, \dots, t_d)^\top, \qquad (z, t_2, \dots, t_d)^\top \sim \mathcal{N}(0, I_d).$$

Let $u = \frac{1}{\sqrt{d}}\mathbf{1}_d$, and U unitary ($UU^\top = I_d$) with the first column of U being u. Then

$$\begin{aligned} Y = Ux &= \lambda\sqrt{d}\theta_0 U e_1 + U(z, t_2, \dots, t_d)^\top \\ &= \lambda\theta_0\sqrt{d}\frac{1}{\sqrt{d}}\mathbf{1}_d + (z_1, z_2, \dots, z_d)^\top \qquad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\ &= (\lambda\theta_0 + z_1, \dots, \lambda\theta_0 + z_d)^\top \end{aligned}$$

□

The Gaussian cloning lemma shows that the scalar model and the vector model are statistically equivalent. As a consequence, the MMSE in dimension d is just d times the scalar MMSE at effective signal-to-noise ratio $\lambda\sqrt{d}$.

Proposition 4.2. *Let* $\mathrm{MMSE}_d(\lambda)$ *denote the minimum mean-squared error in the model*

$$Y = \lambda\theta + Z, \qquad \theta \sim \mathrm{Unif}\{\mathbf{1}_d, -\mathbf{1}_d\}, \qquad Z \sim \mathcal{N}(0, I_d),$$

and let $\mathrm{MMSE}_1(\gamma)$ *denote the minimum mean-squared error in the scalar model*

$$y = \gamma\theta_0 + z, \qquad \theta_0 \sim \mathrm{Unif}\{-1, 1\}, \qquad z \sim \mathcal{N}(0,1).$$

Then

$$\mathrm{MMSE}_d(\lambda) = d\,\mathrm{MMSE}_1\left(\lambda\sqrt{d}\right).$$

Proof. By the Gaussian cloning lemma, $\exists T : \mathbb{R} \times \mathbb{R}^{d-1} \to \mathbb{R}^d$, $T(y, \xi) = (y_1, \dots, y_d)^\top$, where $\xi \sim \mathcal{N}(0, I_{d-1})$. Let $Y = T(y, \xi)$. Then

$$\begin{aligned} \mathrm{MMSE}_d(\lambda) &= \mathbb{E}\left[\|\theta_0\mathbf{1}_d - \mathbb{E}[\theta_0\mathbf{1}_d \mid Y]\|_2^2\right] \\ &= d\,\mathbb{E}\left[\left(\theta_0 - \mathbb{E}[\theta_0 \mid y] + \mathbb{E}[\theta_0 \mid y] - \mathbb{E}[\theta_0 \mid Y]\right)^2\right] \\ &= d\,\mathbb{E}\left[\left(\theta_0 - \mathbb{E}[\theta_0 \mid y]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}[\theta_0 \mid y] - \mathbb{E}[\theta_0 \mid Y]\right)^2\right] \\ &\quad - 2\mathbb{E}\left[\left(\theta_0 - \mathbb{E}[\theta_0 \mid y]\right)\left(\mathbb{E}[\theta_0 \mid y] - \mathbb{E}[\theta_0 \mid Y]\right)\right] \\ &= d\,\mathbb{E}\left[\left(\theta_0 - \mathbb{E}[\theta_0 \mid y]\right)^2\right] \\ &= d\,\mathrm{MMSE}_1\left(\lambda\sqrt{d}\right) \end{aligned}$$

where the last step follows from $Y = T(y, \xi)$ and the Tower Property. □

4.5 Spiked matrix estimation

We now turn to the first genuinely high-dimensional model of the chapter. Let $v \in \mathbb{R}^n$ be a random vector with independent entries

$$v_i \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathop{\mathbb{E}}_{x\sim P_0}[x] = 0, \qquad \mathop{\mathbb{E}}_{x\sim P_0}\left[x^2\right] = 1,$$

and consider the observation

$$Y = \lambda v v^\top + W,$$

where W is a symmetric Gaussian matrix, with upper-triangular entries distributed as independent $\mathcal{N}(0,1)$ random variables.

The object one wishes to estimate is the low-rank signal $vv^\top$, or equivalently the underlying spike v up to the natural symmetries of the model. Since the observation is a Gaussian additive perturbation of a rank-one matrix, this problem fits into the general framework of approximate recovery. At the same time, it exhibits much richer behavior than the two-point model.

It is customary to parametrize the signal strength as

$$\lambda = \sqrt{\frac{t}{n}},$$

where t is kept fixed as $n \to \infty$.

Motivation from Principal Component Analysis (PCA)

A natural first attempt is principal component analysis. Since

$$Y = \lambda v v^\top + W$$

is a rank-one perturbation of a random matrix, one may hope that the leading eigenvector of Y captures the signal direction.

Let

$$u_{\max}(Y) := \arg\max_{u\in\mathbb{S}^{n-1}} \left\{u^\top Y u\right\}$$

denote the top eigenvector of Y. Then a classical random matrix result, known as the BBP transition [BBP05], gives the asymptotic correlation between $u_{\max}(Y)$ and the signal.

Theorem 4.3 (BBP transition). *Assume*

$$Y = \sqrt{\frac{t}{n}}\, vv^\top + W, \qquad v_i \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathop{\mathbb{E}}_{x\sim P_0}[x] = 0, \qquad \mathop{\mathbb{E}}_{x\sim P_0}\left[x^2\right] = 1.$$

Then, with high probability,

$$\frac{|\langle u_{\max}(Y), v\rangle|}{\|v\|_2} \to \begin{cases} 0 & t < 1 \\ \sqrt{1-\frac{1}{t}}, & t > 1. \end{cases}$$

Moreover, for the natural rank-one estimator induced by PCA, one has

$$\lim_{n\to\infty} \frac{\mathrm{MSE}(\mathrm{PCA})}{\mathrm{MSE}(\mathcal{A}_{\text{trivial}})} = \lim_{n\to\infty} \frac{\mathrm{MSE}(\mathrm{PCA})}{\frac{n^2}{2}} = \begin{cases} 1 & t < 1 \\ \frac{1}{t}\left(2-\frac{1}{t}\right) & t > 1. \end{cases}$$

Remark 4.4. *This is the Wigner analogue of the BBP transition; see [FP07] for the largest eigenvalue of rank-one deformed matrices and [BN11] for eigenvalue/eigenvector limits of finite-rank perturbations.*

Thus PCA exhibits a phase transition at $t = 1$. Below that threshold, the leading eigenvector is asymptotically orthogonal to the spike and the corresponding estimator is no better than the trivial one. Above the threshold, PCA becomes positively correlated with the signal and yields nontrivial recovery.

A natural question now arises. Is PCA information-theoretically optimal? Equivalently, does the threshold $t = 1$ coincide with the MMSE threshold for approximate recovery? The answer depends on the prior P_0, and understanding this dependence leads to the replica-symmetric formula.

4.6 The replica-symmetric formula

To state the formula, one first introduces the scalar Gaussian channel

$$y = \sqrt{c}\, x + z, \qquad x \sim P_0, \qquad z \sim \mathcal{N}(0,1),$$

and defines

$$i(c) := I(x; y),$$

the mutual information between x and y.

A useful explicit representation is

$$\begin{aligned}
i(c) &= I(x;y) \\
&= \mathop{\mathbb{E}}_{x,y}\left[\log \frac{p(y \mid x)}{p(y)}\right] \\
&= \mathop{\mathbb{E}}_{x,y}\left[\log \frac{\frac{1}{\sqrt{2\pi}} e^{-y^2/2} \exp\left(\sqrt{c}xy - \frac{c}{2}x^2\right)}{\mathop{\mathbb{E}}_{x' \sim P_0}[p(y \mid x')]}\right] \\
&= \mathop{\mathbb{E}}_{x,y}\left[\log \frac{\frac{1}{\sqrt{2\pi}} e^{-y^2/2} \exp\left(\sqrt{c}xy - \frac{c}{2}x^2\right)}{\frac{1}{\sqrt{2\pi}} e^{-y^2/2} \mathop{\mathbb{E}}_{x' \sim P_0}\left[\exp\left(\sqrt{c}x'y - \frac{c}{2}(x')^2\right)\right]}\right] \\
&= \mathop{\mathbb{E}}_{x,y}\left[\sqrt{c}xy - \frac{c}{2}x^2\right] - \mathop{\mathbb{E}}_{y}\left[\log \mathop{\mathbb{E}}_{x \sim P_0}\left[\exp\left(\sqrt{c}\, xy - \frac{cx^2}{2}\right)\right]\right] \\
&= \left(c - \frac{c}{2}\right)\mathbb{E}\left[x^2\right] + \sqrt{c}\mathbb{E}[xz] - \mathop{\mathbb{E}}_{y}\left[\log \mathop{\mathbb{E}}_{x \sim P_0}\left[\exp\left(\sqrt{c}\, xy - \frac{cx^2}{2}\right)\right]\right] \\
&= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log \mathop{\mathbb{E}}_{x \sim P_0}\left[\exp\left(\sqrt{c}\, xy - \frac{cx^2}{2}\right)\right]\right]
\end{aligned}$$

The remarkable fact is that the asymptotic MMSE in the matrix model can be expressed entirely in terms of this one-dimensional quantity.

Theorem 4.5 (Replica-symmetric formula, [BDM+16; LM19])**.** *Assume*

$$Y = \sqrt{\frac{t}{n}}\, vv^\top + W, \qquad v_i \overset{\text{i.i.d.}}{\sim} P_0, \qquad \mathop{\mathbb{E}}_{x \sim P_0}[x] = 0, \qquad \mathop{\mathbb{E}}_{x \sim P_0}\left[x^2\right] = 1,$$

where $t > 0$ is fixed. Define

$$q^*(t) \in \arg\max_{q \geq 0} \left\{ \frac{tq}{2}\left(1 - \frac{q}{2}\right) - i(tq) \right\}.$$

Then

$$\lim_{n\to\infty} \frac{\mathrm{MMSE}}{\frac{n^2}{2}} = 1 - q^*(t)^2.$$

Remark 4.6. *This formula was first conjectured in the statistical physics literature and later proved rigorously (see [BDM+16], [LM19] and references therein). It provides a precise information-theoretic characterization of the spiked matrix model and shows that the asymptotic behavior of a high-dimensional matrix problem is controlled by a scalar optimization problem. The connection between mutual information and MMSE is governed by the I-MMSE identity of [GSV05].*

The formula is powerful because it allows one to compare algorithmic thresholds with information-theoretic thresholds in a way that depends explicitly on the prior P_0. We now discuss three important examples.

4.7 Applications of the replica-symmetric formula

Gaussian prior

Assume first that

$$P_0 = \mathcal{N}(0,1).$$

In this case one can compute explicitly

$$\begin{aligned}
i(c) &= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}\left[\exp\left(\sqrt{c}\,xy - \frac{cx^2}{2}\right)\right]\right] \\
&= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2} + \sqrt{c}\,xy - \frac{cx^2}{2}\right)\right] \\
&= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log\left(\exp\left(\frac{cy^2}{2(1+c)}\right)\int_{\mathbb{R}} \exp\left(-\frac{1+c}{2}\left(x - \frac{\sqrt{c}y}{1+c}\right)^2\right)\right)\right] \\
&= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log\left(\exp\left(\frac{cy^2}{2(1+c)}\right)\sqrt{\frac{2\pi}{1+c}}\right)\right] \\
&= \frac{c}{2} + \frac{1}{2}\log(1+c) - \frac{c}{2(1+c)}\mathop{\mathbb{E}}_{y}\left[y^2\right] \\
&= \frac{1}{2}\log(1+c)
\end{aligned}$$

and then one can also explicitly compute the replica-symmetric formula: let

$$\phi_t(q) := \frac{tq}{2}\left(1 - \frac{q}{2}\right) - \frac{1}{2}\log(1+tq) \quad \Longrightarrow \quad \phi_t'(q) = \frac{t}{2}\frac{q(t-1-tq)}{1+tq}$$

$$\Longrightarrow \begin{cases} \phi_t(q) \text{ is nonincreasing on } [0,\infty) \Longrightarrow q^*(t) = 0 & t \le 1, \\ \phi_t(q) \text{ is increasing on } [0, 1-\frac{1}{t}) \text{ and decreasing on } (1-\frac{1}{t}, \infty) \Longrightarrow q^*(t) = 1 - \frac{1}{t} & t > 1. \end{cases}$$

$$\Longrightarrow q^*(t) = \begin{cases} 0 & t \le 1 \\ 1 - \frac{1}{t} & t > 1. \end{cases}$$

Therefore

$$\lim_{n\to\infty} \frac{\mathrm{MMSE}}{\frac{n^2}{2}} = \begin{cases} 1 & t \le 1 \\ 1 - (1-\frac{1}{t})^2 = \frac{1}{t}(2 - \frac{1}{t}) & t > 1. \end{cases}$$

This coincides with the performance of PCA. Hence, for Gaussian priors, principal component analysis is information-theoretically optimal (see Figure 4.2).

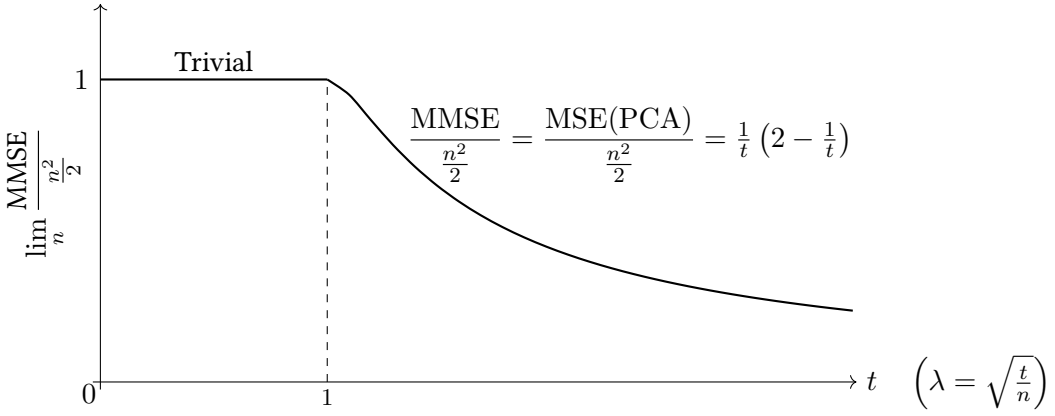

Figure 4.2: PCA Optimality for Gaussian prior

Rademacher prior

Assume now that

$$P_0 = \text{Unif}\left(\{-1, 1\}\right).$$

Then

$$\begin{aligned} i(c) &= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log \mathop{\mathbb{E}}_{x\sim P_0}\left[\exp\left(\sqrt{c}\,xy - \frac{cx^2}{2}\right)\right]\right] \\ &= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log\left(e^{-\frac{c}{2}}\cosh\left(\sqrt{c}\,y\right)\right)\right] \\ &= c - \mathop{\mathbb{E}}_{y}\left[\log\cosh\left(\sqrt{c}\,y\right)\right] \\ &= c - \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\log\cosh\left(c + \sqrt{c}\,Z\right)\right], \end{aligned}$$

where the last equality follows from the symmetries of $\cosh$ and P_0.

Then the replica-symmetric formula yields not a closed expression, but a fixed-point relation that allows to end up proving the non-optimality of PCA for $t > 1$ (see Section A.2 for the proof). PCA is only optimal for $0 \le t < 1$; although it succeeds above $t = 1$, it does not attain the information-theoretic minimum mean-squared error (see Figure 4.3).

Sparse prior with fixed sparsity level

Consider next the sparse prior

$$v_i = \begin{cases} 0 & \text{with probability } 1 - p \\ \frac{1}{\sqrt{p}} & \text{with probability } \frac{p}{2} \\ -\frac{1}{\sqrt{p}} & \text{with probability } \frac{p}{2} \end{cases}$$

where $p \in (0, 1)$ is a fixed constant. Then the expected number of nonzero coordinates is about pn, so the spike is sparse.

The replica-symmetric formula applies in this setting as well. One of its consequences is that if p is sufficiently small ($p < 0.09$) then there exist values of $t < 1$ for which $q^*(t) > 0$. In other words,

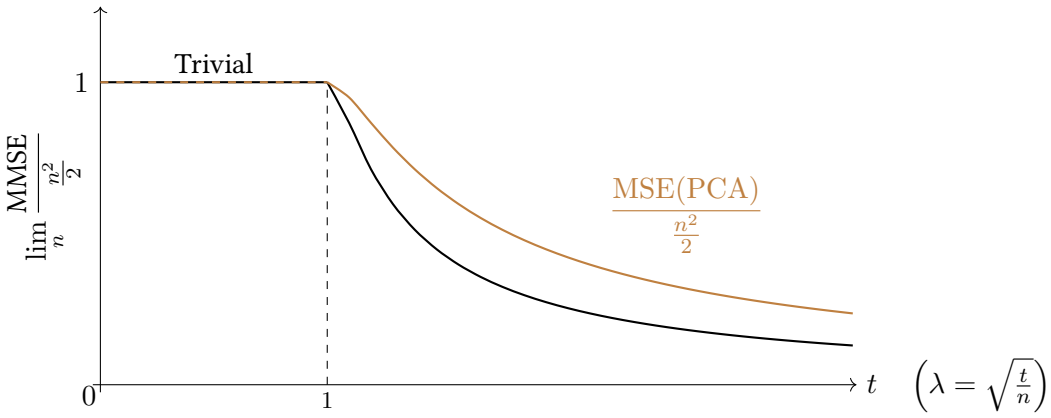

Figure 4.3: PCA Non-optimality for Rademacher prior

nontrivial information-theoretic recovery is already possible below the BBP threshold $t = 1$ (see [LL18] for more details). Thus PCA is suboptimal even at the level of the recovery threshold (see Figure 4.4).

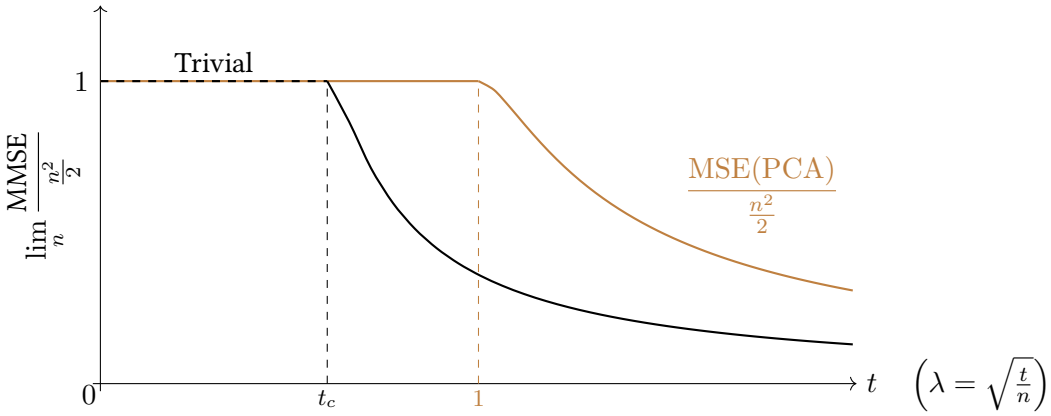

Figure 4.4: PCA suboptimality for fixed sparsity level prior ($p < 0.09$)

These examples illustrate a general principle: whether PCA is optimal depends not only on the signal-to-noise ratio, but also on the geometry of the prior. The Gaussian prior is a special case in which the PCA and information-theoretic thresholds coincide. As soon as one moves to more structured priors, they may separate.

4.8 The all-or-nothing phenomenon

We now let the sparsity level vanish with the dimension. Consider again the sparse prior

$$v_i = \begin{cases} 0 & \text{with probability } 1 - p_n \\ \frac{1}{\sqrt{p_n}} & \text{with probability } \frac{p_n}{2} \\ -\frac{1}{\sqrt{p_n}} & \text{with probability } \frac{p_n}{2} \end{cases} \qquad p_n \to 0,$$

and the model

$$Y = \lambda v v^\top + W.$$

In this regime the behavior changes qualitatively. Instead of a continuous transition as in the dense case, one observes an abrupt one: either the MMSE is essentially equal to the trivial error, or it is asymptotically negligible compared to it.

Theorem 4.7 (All-or-nothing phenomenon, [NZ20a]). *There exists a critical value λ_c, of order*

$$\lambda_c \asymp \sqrt{\frac{p_n}{n}},$$

such that for every fixed $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{\mathrm{MMSE}}{\mathrm{MSE}(\mathcal{A}_{\mathrm{trivial}})} = \begin{cases} 0 & \lambda > (1+\varepsilon)\lambda_c \\ 1 & \lambda < (1-\varepsilon)\lambda_c. \end{cases}$$

The theorem, which was proven in [NZ20a], says that there is essentially no intermediate regime (the first all-or-nothing phenomenon has been discovered in the context of sparse high dimensional linear regression [GZ22; NZ20b] and it has also been discovered in a plethora of models since). Below the critical threshold, the observations are useless from the MMSE point of view: one cannot do asymptotically better than the trivial estimator. Above the threshold, one can almost recover the spike, or equivalently the rank-one signal $vv^\top$, in a much stronger sense than in the dense case.

This is why the phenomenon is called all-or-nothing. The MMSE does not decrease gradually through the transition; it jumps from the trivial level to a near-perfect recovery regime (see Figure 4.5).

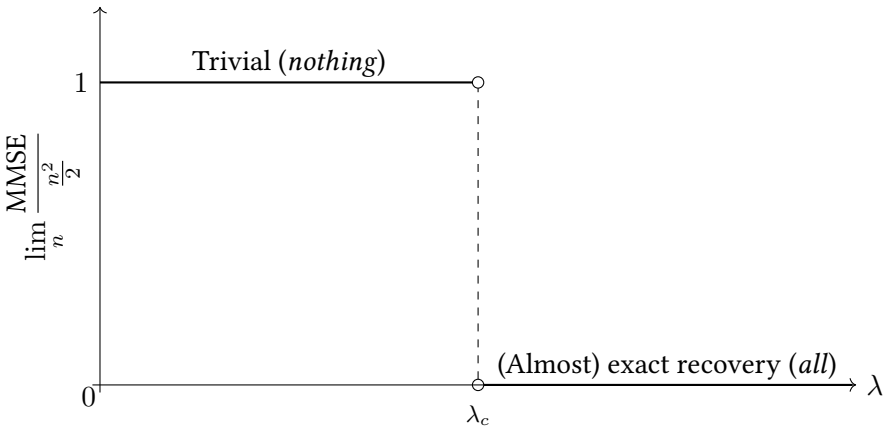

Figure 4.5: The all-or-nothing phase transition in sparse spiked matrix estimation

The contrast with the BBP transition is striking. In the classical dense setting, the correlation between the spectral estimator and the true spike increases continuously above threshold. In the sparse vanishing-density regime, by contrast, the information-theoretic behavior becomes discontinuous.

4.9 Discussion

The central message of this chapter is that approximate recovery is governed by the posterior mean and the MMSE, just as exact recovery and detection are governed by the MAP estimator and the Neyman-Pearson test. In simple Gaussian models, this leads to explicit formulas and effective low-dimensional reductions. In the spiked matrix model, however, the problem becomes much richer and reveals several layers of structure.

First, the BBP transition shows that PCA has a sharp algorithmic threshold at $t = 1$. Second, the replica-symmetric formula shows that the true information-theoretic threshold depends on the prior and may coincide with or differ from the spectral one. Third, in sparse models one encounters the all-or-nothing phenomenon, where approximate recovery undergoes a discontinuous transition.

Part II

Detection

Chapter 5

The Low-Degree Likelihood Ratio

5.1 The Computational Question for Detection

In the previous chapters, we studied the information-theoretic limits of exact recovery, detection, and approximate recovery. In particular, in the setting of detection, to which this part of the notes is devoted, we observe $Y \in \Omega$ drawn either from a structured distribution $\mathbb{P}_1$ or from a null distribution $\mathbb{P}_2$, and we seek to determine which of the two generated the data.

At the purely statistical level, this question is completely understood. By the Neyman-Pearson lemma, the optimal test declares $\mathbb{P}_1$ if and only if $\mathbb{P}_1(Y) \geq \mathbb{P}_2(Y)$, or equivalently if the likelihood ratio (assuming it is well-defined) $\mathbb{P}_1(Y)/\mathbb{P}_2(Y)$ is at least 1. However, this does not solve the computational problem. In many examples of interest, the likelihood ratio is, in principle, an intractable object.

Example 5.1. *This difficulty already appears in very simple Bayesian formulations. Suppose that $\theta \sim \mu$ is a random signal on $\mathbb{R}^d$, and conditionally on θ we observe $Y \sim \mathbb{P}_\theta$. Then the planted distribution is the mixture*

$$\mathbb{P}_1(Y) = \mathop{\mathbb{E}}_{\theta\sim\mu}[\mathbb{P}_\theta(Y)].$$

If the prior μ is supported on a very large set, this average may be very costly to compute. In many relevant problems, the support of μ grows exponentially with the ambient dimension. For instance, if the signal set is $\{-1,1\}^d$, then the mixture already involves 2^d terms.

This is the point at which the computational question enters. Rather than asking only whether detection is possible in principle, we now ask:

Question 5.2. *Can one find a polynomial-time computable test that achieves weak or strong detection?*

Of course, the most direct way to formalize the computational question would be to optimize over all polynomial-time tests, for instance over all tests computable in time at most n^{100}. However, this class is too broad and unstructured to analyze directly. It includes arbitrary algorithms with branching, adaptivity, randomness, memory, and model-dependent subroutines. Unlike a linear space of functions, it has no convenient basis, no inner product structure, and no projection theorem that identifies the best possible polynomial-time test.

The low-degree framework replaces this intractable class by a structured proxy. Instead of optimizing over all polynomial-time tests, we restrict attention to low-degree polynomial functions of the data. This class is still rich enough to capture many natural algorithmic statistics, but it is also a finite-dimensional linear space. As a result, one can use Hilbert-space geometry: the best low-degree approximation to

the likelihood ratio is its orthogonal projection onto the space of low-degree polynomials. This is the low-degree likelihood ratio.

The philosophy of the method is therefore not that every efficient algorithm is literally a low-degree polynomial. Rather, the low-degree method provides an analyzable proxy for efficient computation. Its predictions have been remarkably accurate in many high-dimensional inference problems; see, for example, [Hop18; KWB19].

5.2 Inner Products Between Functions

Before introducing low-degree polynomials, we first need a convenient functional-analytic language. Let $f, g : \Omega \to \mathbb{R}$ be two real-valued functions on the observation space Ω.

Definition 5.3. *The* inner product *between f and g with respect to the null distribution $\mathbb{P}_2$ is defined by*

$$\langle f, g \rangle_{\mathbb{P}_2} := \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} [f(Y) g(Y)].$$

This is the natural inner product on $L^2(\Omega, \mathbb{P}_2)$, the space of square-integrable functions with respect to the null model.

Lemma 5.4. *The map $\langle \cdot, \cdot \rangle_{\mathbb{P}_2}$ defines an inner product on $L^2(\Omega, \mathbb{P}_2)$.*

Proof. In order to show that $\langle \cdot, \cdot \rangle_{\mathbb{P}_2}$ is an inner product, we need to show that it satisfies the symmetry, linearity, and positive-definiteness properties of inner products.

1. *Symmetry*: For any functions $f, g : \Omega \mapsto \mathbb{R}$, we have
$$\langle f, g \rangle_{\mathbb{P}_2} = \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} [f(Y) g(Y)] = \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} [g(Y) f(Y)] = \langle g, f \rangle_{\mathbb{P}_2}.$$
2. *Linearity in the first argument*: For any functions $f, g, h : \Omega \mapsto \mathbb{R}$ and any scalars $\alpha, \beta \in \mathbb{R}$, it follows from the linearity of the expectation that
$$\begin{aligned} \langle \alpha f + \beta g, h \rangle_{\mathbb{P}_2} &= \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} [(\alpha f(Y) + \beta g(Y)) h(Y)] = \alpha \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} [f(Y) h(Y)] + \beta \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} [g(Y) h(Y)] \\ &= \alpha \langle f, h \rangle_{\mathbb{P}_2} + \beta \langle g, h \rangle_{\mathbb{P}_2}. \end{aligned}$$
3. *Positive-definiteness*: For any function $f : \Omega \mapsto \mathbb{R}$, we have
$$\langle f, f \rangle_{\mathbb{P}_2} = \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} \left[f(Y)^2 \right] \geq 0$$
by monotonicity of the expectation. Moreover, we have equality if and only if $f = 0$ almost surely under $\mathbb{P}_2$.

□

As usual, the associated norm is given by

$$\|f\|_{2, \mathbb{P}_2} := \sqrt{\langle f, f \rangle_{\mathbb{P}_2}} = \sqrt{\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} [f(Y)^2]}.$$

When no confusion is possible, we will simply write $\|f\|_2$.

The usefulness of this norm is that it allows us to compare candidate test functions under the null model. In the end, the likelihood ratio will appear as a distinguished vector in this Hilbert space.

Lemma 5.5 (Cauchy-Schwarz inequality)**.** *For any functions $f, g \in L^2(\Omega, \mathbb{P}_2)$,*

$$|\langle f, g\rangle_{\mathbb{P}_2}| \le \|f\|_2 \|g\|_2,$$

with equality if and only if $f = \lambda g$ for some scalar $\lambda \in \mathbb{R}$.

Proof. By definition of the inner product and the induced norm,

$$|\langle f, g\rangle_{\mathbb{P}_2}| = \left| \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[f(Y)g(Y)] \right| \le \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[|f(Y)g(Y)|] \le \sqrt{\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[f(Y)^2] \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[g(Y)^2]} = \|f\|_2\|g\|_2.$$

Here, the first inequality follows from the triangle inequality (or also Jensen's inequality), and the second inequality follows Cauchy-Schwarz inequality for expectations. Equality holds if and only if $f = \lambda g$ for some scalar $\lambda \in \mathbb{R}$ by the equivalent properties of the Cauchy-Schwarz inequality for expectations. □

5.3 A Closer Look at the Likelihood Ratio

Recall that the likelihood ratio is

$$L(Y) := \frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)},$$

whenever $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_2$.

The χ^2-divergence can be written directly in terms of L:

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[L(Y)^2\right] - 1.$$

In particular, the quantity $\|L\|_2$ is exactly the square root of $1 + \chi^2(\mathbb{P}_1, \mathbb{P}_2)$.

Claim 5.6. *The χ^2-divergence is nonnegative.*

Proof. Using the change-of-measure identity,

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}\right] = \sum_{Y\in\Omega} \frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}\mathbb{P}_2(Y) = \sum_{Y\in\Omega} \mathbb{P}_1(Y) = 1.$$

Applying the Jensen's inequality to the convex function $x \mapsto x^2$, we get

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[L(Y)^2\right] \ge \left(\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[L(Y)]\right)^2 = 1.$$

Hence

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[L(Y)^2\right] - 1 \ge 0.$$

□

The significance of this observation is that if the likelihood ratio has very small L^2 norm, then the planted and null distributions are close, and (weak) detection should be impossible. More precisely, the bounds from Chapter 3 imply that

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) \le \sqrt{\chi^2(\mathbb{P}_1, \mathbb{P}_2)} = \sqrt{\|L\|_2^2 - 1}.$$

Thus if $\|L\|_2 = 1 + o(1)$, then total variation tends to zero, and even weak detection is impossible.

A slightly weaker but still useful criterion is the following form of Le Cam's method.

Lemma 5.7 (Le Cam's method [KWB19, Lemma 1.13]). *Let* $(\mathbb{P}_{1,n})$ *and* $(\mathbb{P}_{2,n})$ *be two sequences of distributions. If*

$$\chi^2(\mathbb{P}_{1,n}, \mathbb{P}_{2,n}) = \mathcal{O}(1),$$

then strong detection is impossible.

Proof. Suppose, by contradiction, that there exists an algorithm $\mathcal{A}$ such that $\lim_{n\to\infty}(\mathbb{P}_{1,n}(\mathcal{A}(Y) = \mathbb{P}_2) + \mathbb{P}_{2,n}(\mathcal{A}(Y) = \mathbb{P}_1)) = 0$. But, by the change of measure trick and Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{P}_{1,n}(\mathcal{A}(Y) = \mathbb{P}_1) &= \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_{1,n}}[\mathbb{I}(\mathcal{A}(Y) = \mathbb{P}_1)] \\
&= \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_{2,n}}\left[\frac{\mathbb{P}_{1,n}(Y)}{\mathbb{P}_{2,n}(Y)}\mathbb{I}(\mathcal{A}(Y) = \mathbb{P}_1)\right] \\
&\le \sqrt{\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_{2,n}}\left[\left(\frac{\mathbb{P}_{1,n}(Y)}{\mathbb{P}_{2,n}(Y)}\right)^2\right]}\sqrt{\mathbb{P}_{2,n}(\mathcal{A}(Y) = \mathbb{P}_1)} \\
&= \sqrt{\chi^2(\mathbb{P}_{1,n}, \mathbb{P}_{2,n}) + 1}\sqrt{\mathbb{P}_{2,n}(\mathcal{A}(Y) = \mathbb{P}_1)}.
\end{aligned}$$

Since $\chi^2(\mathbb{P}_{1,n}, \mathbb{P}_{2,n}) = \mathcal{O}(1)$ and we are supposing $\mathbb{P}_{2,n}(\mathcal{A}(Y) = \mathbb{P}_{1,n}) \to 0$, then $\mathbb{P}_{1,n}(\mathcal{A}(Y) = \mathbb{P}_1) \to 0$. This contradicts the assumption that $\mathbb{P}_{1,n}(\mathcal{A}(Y) = \mathbb{P}_2) + \mathbb{P}_{2,n}(\mathcal{A}(Y) = \mathbb{P}_1) \to 0$, and the lemma follows. □

5.4 A Variational Formula for the χ^2-Divergence

We now explain why the likelihood ratio is the optimal test function from the $L^2(\mathbb{P}_2)$ point of view.

Theorem 5.8 (Variational formula). *For every square-integrable function* $f : \Omega \to \mathbb{R}$,

$$\frac{\left\langle f, \frac{\mathbb{P}_1}{\mathbb{P}_2}\right\rangle_{\mathbb{P}_2}}{\|f\|_2} \le \left\|\frac{\mathbb{P}_1}{\mathbb{P}_2}\right\|_2.$$

Moreover,

$$\left\|\frac{\mathbb{P}_1}{\mathbb{P}_2}\right\|_2 = \max_{f:\Omega\to\mathbb{R}} \frac{\left\langle f, \frac{\mathbb{P}_1}{\mathbb{P}_2}\right\rangle_{\mathbb{P}_2}}{\|f\|_2},$$

and equality holds if and only if f *is proportional to* $\mathbb{P}_1/\mathbb{P}_2$.

Proof. This is an immediate consequence of Cauchy-Schwarz in $L^2(\Omega, \mathbb{P}_2)$, applied to f and $\mathbb{P}_1/\mathbb{P}_2$. □

This theorem says that the likelihood ratio is the best possible test statistic with respect to this variational formula if one is allowed to optimize over all square-integrable functions. The computational issue is that this optimization takes place over far too large a class, including many potentially computationally intractable functions. The low-degree method will instead restrict attention to a much smaller space of functions.

5.5 The Low-Degree Likelihood Ratio

Suppose from now on that the observation Y is encoded as a vector in $\mathbb{R}^n$. For a multi-index $\alpha \in \mathbb{N}^n$, we write

$$Y^\alpha := \prod_{i=1}^n Y_i^{\alpha_i}, \qquad |\alpha| := \sum_{i=1}^n \alpha_i.$$

A polynomial $f(Y) = \sum_\alpha c_\alpha Y^\alpha$ is said to have degree at most D if $c_\alpha = 0$ whenever $|\alpha| > D$. We denote the space of such polynomials by $\mathbb{R}_{\le D}[Y]$.

A low-degree analogue of the variational formula is then

$$\max_{f \in \mathbb{R}_{\le D}[Y]} \frac{\left\langle f, \frac{\mathbb{P}_1}{\mathbb{P}_2} \right\rangle_{\mathbb{P}_2}}{\|f\|_2}.$$

By the projection theorem in finite-dimensional inner-product spaces, this maximum is attained by the orthogonal projection of the likelihood ratio onto $\mathbb{R}_{\le D}[Y]$.

Definition 5.9. *The orthogonal projection of $\mathbb{P}_1/\mathbb{P}_2$ onto $\mathbb{R}_{\le D}[Y]$ is denoted by*

$$\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D}$$

and is called the (degree-$\le D$) low-degree likelihood ratio.

Lemma 5.10.

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D} \right\|_2 = \max_{f \in \mathbb{R}_{\le D}[Y]} \frac{\left\langle f, \frac{\mathbb{P}_1}{\mathbb{P}_2} \right\rangle_{\mathbb{P}_2}}{\|f\|_2}.$$

Proof. By Cauchy-Schwarz inequality, $\langle v, v^* \rangle / \|v\| \le \|v^*\|$. In particular, the supremum over $v \in L$ is finite and, as every finite-dimensional subspace is closed, it is attained at some $v \in L$. The rest of the proof follows from the properties of the orthogonal projection. Write $v^* = \mathrm{Proj}_L(v^*) + (v^* - \mathrm{Proj}_L(v^*))$, and note $\langle v, v^* \rangle / \|v\| = \langle v, \mathrm{Proj}_L(v^*) \rangle / \|v\| \le \|\mathrm{Proj}_L(v^*)\|$ with equality for $v = \mathrm{Proj}_L(v^*)/\|\mathrm{Proj}_L(v^*)\|$. □

Thus the low-degree norm measures how much of the L_2-norm of the likelihood ratio can already be captured by a polynomial of degree at most D.

5.6 Why Low-Degree Polynomials?

We have motivated low-degree polynomials as a structured proxy for polynomial-time computation. We now explain why this proxy is plausible. The answer is not that all efficient algorithms are literally polynomial statistics. Rather, many natural statistics that arise in algorithm design are low-degree, or admit good low-degree approximations.

For example, the empirical mean is a degree-one statistic, and the empirical variance is a degree-two statistic. More interestingly, in matrix models the top eigenvalue can often be approximated by a low-degree polynomial. Suppose that the eigenvalues of a symmetric matrix Y satisfy

$$\lambda_n(Y) \le \cdots \le \lambda_2(Y) \le (1 - \varepsilon)\lambda_1(Y)$$

for some fixed $\varepsilon > 0$. Then

$$\begin{aligned}
\operatorname{tr}\left(Y^{\lceil c\log n\rceil}\right) &= \sum_{i=1}^{n} \lambda_i(Y)^{\lceil c\log n\rceil} \\
&= \lambda_1(Y)^{\lceil c\log n\rceil}\left(1+\sum_{i=2}^{n}\left(\frac{\lambda_i(Y)}{\lambda_1(Y)}\right)^{\lceil c\log n\rceil}\right) \\
&\le \lambda_1(Y)^{\lceil c\log n\rceil}\left(1+\sum_{i=2}^{n}(1-\varepsilon)^{\lceil c\log n\rceil}\right) \\
&\le \lambda_1(Y)^{\lceil c\log n\rceil}\left(1+\sum_{i=2}^{n} n^{1-c\log\frac{1}{1-\varepsilon}}\right).
\end{aligned}$$

If $c \ge -2/\log(1-\varepsilon)$ and $\lceil c\log n\rceil$ is even, then

$$\lambda_1(Y)^{\lceil c\log n\rceil} \le \operatorname{tr}\left(Y^{\lceil c\log n\rceil}\right) \le \lambda_1(Y)^{\lceil c\log n\rceil}(1+o(1)),$$

so the contribution of the smaller eigenvalues is negligible. Thus, this degree-$\mathcal{O}(\log n)$ statistic essentially recovers the top eigenvalue.

This example is one of the main motivations for the low-degree method: even nonlinear spectral statistics can be well-approximate by low-degree polynomials.

5.7 The Low-Degree Conjecture

The previous discussion leads to the following guiding principle, discussed in [Hop18, Conjecture 2.2.4] and [KWB19, Conjecture 1.16], among others.

Conjecture 5.11 (Low-degree conjecture)**.** *For* sufficiently nice *detection problems (see [KWB19, Section 4.2.4]), if*

$$\left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D}\right\|_2 = 1+o(1)$$

for some degree $D = (\log n)^{1+\varepsilon}$, $\varepsilon > 0$, then weak detection should be impossible for polynomial-time algorithms. Similarly, if

$$\left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D}\right\|_2 = \mathcal{O}(1),$$

then strong detection should be impossible for polynomial-time algorithms.

While this conjecture remains open, it has been remarkably successful as a predictive principle across a wide range of high-dimensional inference problems. In these notes, we will not explicitly define what constitutes a "nice" detection problem; instead, we intentionally assume that the conjecture applies to most of the tasks presented here. Delineating the exact family of tasks to which the low-degree conjecture applies remains an active area of research, shaped by both informative counterexamples [HW21; ZSW+22; DK22; BHJ+25; JV26] and partial proofs in specific settings [HKK+26]. Our primary focus in these notes is to explore the implications of this conjecture, *assuming* it holds for our models of interest.

5.8 Strong Separability

To support the conjectural picture, it is useful to formalize a notion of statistical separation discussed in [BEH+22] that can be ruled out by bounded low-degree norm.

Definition 5.12 ([BEH+22, Definition 1.8]). *A function $f : \Omega \to \mathbb{R}$ strongly separates $\mathbb{P}_1$ and $\mathbb{P}_2$ if*

$$\left| \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}[f(Y)] - \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[f(Y)] \right| = \omega\left(\sqrt{\max\left\{ \mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_1}(f(Y)), \mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_2}(f(Y)) \right\}} \right).$$

The intuition is that if a statistic strongly separates the planted and null models, then its typical values under the two distributions are far enough apart that thresholding yields strong detection. We prove this fact in the following lemma (stated in [BEH+22, Section 1.2.]).

Lemma 5.13. *If f strongly separates $\mathbb{P}_1$ and $\mathbb{P}_2$, then f also strongly detects.*

Proof. Set

$$m_i := \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_i}[f(Y)], \qquad \sigma_i := \sqrt{\mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_i}(f(Y))}.$$

Since f strongly separates $\mathbb{P}_1$ and $\mathbb{P}_2$, we have

$$\frac{|m_1 - m_2|}{\max\{\sigma_1, \sigma_2\}} \to \infty.$$

Without loss of generality, assume that $m_1 > m_2$.

Then

$$\frac{m_1 - m_2}{\sigma_1 + \sigma_2} \to \infty,$$

because $\sigma_1 + \sigma_2 \le 2\max\{\sigma_1, \sigma_2\}$. Hence, for all sufficiently large dimensions and for every fixed $C > 0$, we have

$$m_1 - m_2 \ge C(\sigma_1 + \sigma_2).$$

In particular, we may choose a threshold T such that

$$m_1 - C\sigma_1 \ge T \ge m_2 + C\sigma_2.$$

Now define the test $\mathcal{T}$ by

$$\mathcal{T}(Y) = \begin{cases} \mathbb{P}_1 & \text{if } f(Y) \ge T, \\ \mathbb{P}_2 & \text{if } f(Y) < T. \end{cases}$$

We bound its two error probabilities using Chebyshev's inequality.

Under $\mathbb{P}_1$, we have

$$\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_1}(\mathcal{T}(Y) = \mathbb{P}_2) = \mathop{\mathbb{P}}_{Y\sim\mathbb{P}_1}(f(Y) < T).$$

Since $T \le m_1 - C\sigma_1$, the event $\{f(Y) < T\}$ implies

$$|f(Y) - m_1| \ge m_1 - T \ge C\sigma_1.$$

Therefore, by Chebyshev's inequality,

$$\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_1}(\mathcal{T}(Y) = \mathbb{P}_2) \le \mathop{\mathbb{P}}_{Y\sim\mathbb{P}_1}\left(|f(Y) - m_1| \ge C\sigma_1\right) \le \frac{\sigma_1^2}{C^2\sigma_1^2} = \frac{1}{C^2}.$$

Similarly, under $\mathbb{P}_2$,

$$\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(\mathcal{T}(Y)=\mathbb{P}_1)=\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(f(Y)\geq T).$$

Since $T\geq m_2+C\sigma_2$, the event $\{f(Y)\geq T\}$ implies

$$|f(Y)-m_2|\geq T-m_2\geq C\sigma_2.$$

Hence, again by Chebyshev's inequality,

$$\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(\mathcal{T}(Y)=\mathbb{P}_1)\leq\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}\big(|f(Y)-m_2|\geq C\sigma_2\big)\leq\frac{\sigma_2^2}{C^2\sigma_2^2}=\frac{1}{C^2}.$$

Thus

$$\mathrm{err}(\mathcal{T})=\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_1}(\mathcal{T}(Y)=\mathbb{P}_2)+\mathop{\mathbb{P}}_{Y\sim\mathbb{P}_2}(\mathcal{T}(Y)=\mathbb{P}_1)\leq\frac{2}{C^2}.$$

Since $C>0$ is arbitrary and the strong separation assumption guarantees that such a threshold exists for arbitrarily large C when the dimension is large enough, it follows that $\mathrm{err}(\mathcal{T})\to 0$. Therefore f strongly detects. □

We illustrate the concept of strong separability in Figure 5.1.

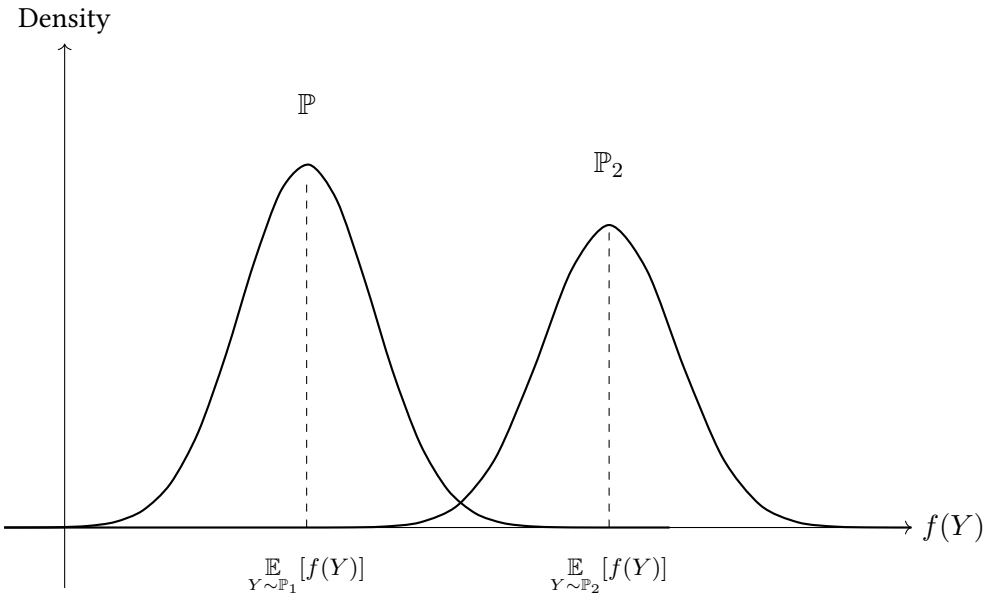

Figure 5.1: Illustration of strong separability.

The following theorem is the main rigorous statement of the chapter.

Theorem 5.14 ([BEH+22, Proposition 6.2.]). *If*

$$\left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\leq D}\right\|_2=\mathcal{O}(1),$$

then no polynomial of degree at most D strongly separates $\mathbb{P}_1$ and $\mathbb{P}_2$.

Proof of Theorem 5.14. Assume by contradiction that there exists a d-degree ($d \le D$) polynomial p that strongly separates $\mathbb{P}_1$ and $\mathbb{P}_2$. Then, by definition of strong separability,

$$\frac{\left|\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}[p(Y)] - \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[p(Y)]\right|}{\sqrt{\max\left\{\mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_1}(p(Y)), \mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_2}(p(Y))\right\}}} \to \infty.$$

Consider $f(Y) := p(Y) - \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[p(Y)]$. Then, $\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}[f(Y)] = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}[p(Y)] - \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[p(Y)]$ and

$$\max\left\{\mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_1}(p(Y)), \mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_2}(p(Y))\right\} \ge \mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_2}(p(Y)) = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[f(Y)^2\right].$$

Hence, by changing the measure and applying Lemma 5.10,

$$\frac{\left|\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}[p(Y)] - \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[p(Y)]\right|}{\sqrt{\max\left\{\mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_1}(p(Y)), \mathop{\mathrm{Var}}_{Y\sim\mathbb{P}_2}(p(Y))\right\}}} \le \frac{\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_1}[f(Y)]}{\|f\|_2} = \frac{\langle f, \mathbb{P}_1/\mathbb{P}_2\rangle_{\mathbb{P}_2}}{\|f\|_2} \le \left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)^{\le d}\right\|_2.$$

The right hand side is $\mathcal{O}(1)$ by assumption, but the left hand diverges to infinity by strong separability. This is a contradiction, and the theorem follows. □

Hence, this theorem shows that a bounded low-degree norm implies a natural and concrete family of "low-degree" tests fails.

Chapter 6

Detection for Gaussian Additive Models

6.1 Introduction

Recall from the previous Chapter 5 that the quantity

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2$$

plays a central role in the low-degree approach to detection. Under the low-degree Conjecture 5.11, if this norm remains bounded for a *sufficiently nice* problem, then polynomial-time detection should be impossible.

The next natural question is whether one can actually compute low-degree norms in interesting models. This chapter addresses that question in the setting of Gaussian additive models. These models are important for two reasons. First, they include several central examples in high-dimensional inference. Second, their Gaussian structure makes it possible to write explicit formulas both for the full likelihood ratio and for its low-degree truncation.

The workflow of the chapter is the following. We begin with sparse PCA as a motivating example. This serves to illustrate, before any general formalism is introduced, the kind of computational-statistical gap that we want to understand. We then derive general formulas for Gaussian additive models, first for the full likelihood ratio and its L^2 norm, and then for the low-degree likelihood ratio. Finally, we return to sparse PCA and combine these formulas with an overlap estimate to obtain a low-degree lower bound.

Thus the chapter has a double role. It both motivates the Gaussian additive framework and shows, on a concrete model, how the low-degree method is actually used. The proofs of the low-degree lower bounds for this Chapter are often following similar or identical results from the survey [KWB19].

6.2 Gaussian Additive Models

Let μ be a prior on $\mathbb{R}^N$. A Gaussian additive model is a detection problem of the form

$$\mathbb{P}_1 : \quad Y = \lambda X + Z,$$

where $X \sim \mu$, $Z \sim \mathcal{N}(0, I_N)$, and $\lambda > 0$ is a signal-to-noise ratio parameter, versus the null model

$$\mathbb{P}_2 : \quad Y = Z,$$

where again $Z \sim \mathcal{N}(0, I_N)$.

Many examples of interest fit into this framework. In particular, spiked matrix and tensor models can often be rewritten as Gaussian additive models after vectorization or flattening. The first example we treat is sparse PCA in detection form.

6.3 Sparse PCA as a motivating example

Consider the detection version of sparse PCA. We observe a symmetric matrix $Y \in \mathbb{R}^{n \times n}$, and we want to distinguish between

$$\mathbb{P}_1 : \quad Y = \frac{\lambda}{\sqrt{2}} \theta\theta^\top + W,$$

where W is a symmetric Gaussian noise matrix and the spike $\theta \in \mathbb{R}^n$ has i.i.d. coordinates distributed as

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

versus

$$\mathbb{P}_2 : \quad Y = W.$$

Here $k = \rho n$ is the sparsity level, and we should think of ρ as a small constant. Since the coordinates of θ are i.i.d., we have

$$\|\theta\|_0 \sim \mathrm{Bin}\left(n, \frac{k}{n}\right),$$

so $\|\theta\|_0 \approx k \pm \sqrt{k}$, while

$$\|\theta\|_2^2 \approx 1.$$

The parameter λ controls the signal strength. The basic question is to understand for which values of λ one can distinguish the planted matrix from pure noise.

6.3.1 The spectral benchmark

A first natural idea is to apply principal component analysis and look at the top eigenvalue or top eigenvector of Y. As discussed earlier in the notes, when

$$\lambda = t\sqrt{n},$$

with t constant, one has the BBP-type transition Theorem 4.3

$$\lim_{n\to\infty} \left\langle v_{\max}(Y), \frac{\theta}{\|\theta\|_2} \right\rangle^2 = \begin{cases} 0 & t < 1, \\ 1 - \dfrac{1}{t^2} & t > 1, \end{cases}$$

with high probability, where $v_{\max}(Y)$ denotes the leading eigenvector of Y.

Similarly, for the leading eigenvalue one has, with high probability,

$$\lim_{n\to\infty} \frac{\lambda_{\max}(Y)}{\sqrt{n}} = \begin{cases} 2 & t < 1, \\ t + \dfrac{1}{t} & t > 1, \end{cases}$$

under the planted model (see [FP07, Theorem 1.1.]); hence, taking $t = 0$, under the null model

$$\lim_{n\to\infty} \frac{\lambda_{\max}(Y)}{\sqrt{n}} = 2.$$

Thus PCA gives a polynomial-time test for sparse PCA detection once $t > 1$, that is, once λ is larger than $\sqrt{n}$ up to constant factors (see Figure 6.1).

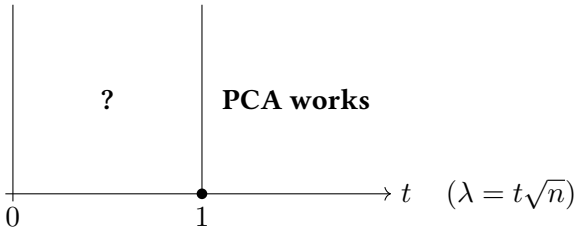

Figure 6.1: Spectral detection threshold in sparse PCA

This gives the first algorithmic threshold in the problem. The natural next question is whether one can do better if one is willing to use an exponential-time procedure that explicitly exploits the sparsity of the spike.

6.3.2 A stronger statistic

To use sparsity, consider[1]

$$\mathrm{OPT}(Y) := \max_{v\in\left\{0,\pm\frac{1}{\sqrt{k}}\right\}^n,\ \|v\|_0\le 2k} v^\top Y v.$$

This is not a polynomial-time computable statistic, in fact it is NP-Hard; however it gives information about the information-theoretic detection threshold.

Under the planted model,

$$v^\top Y v = v^\top\left(\frac{\lambda}{\sqrt{2}}\theta\theta^\top + W\right)v = \frac{\lambda}{\sqrt{2}}\langle v,\theta\rangle^2 + v^\top W v.$$

If we take $v = \theta$, then

$$\theta^\top Y\theta = \frac{\lambda}{\sqrt{2}}\langle\theta,\theta\rangle^2 + \theta^\top W\theta \approx \frac{\lambda}{\sqrt{2}} + \theta^\top W\theta.$$

Since W is Gaussian and $\|\theta\|_2^2 \approx 1$, the random variable $\theta^\top W\theta$ is approximately $\mathcal{N}(0,1)$. Hence

$$\mathrm{OPT}(Y) \ge \frac{\lambda}{\sqrt{2}} + T, \qquad T\sim\mathcal{N}(0,1),$$

under the planted model.

Under the null model, we need to control

$$\max_{v\in\left\{0,\pm\frac{1}{\sqrt{k}}\right\}^n,\ \|v\|_0\le 2k} v^\top W v.$$

For each fixed v, the quantity $v^\top W v$ is a centered Gaussian with variance of order one. The relevant tool is the following elementary bound (which we prove at the end of this section).

[1] In the proof of Theorem 6.7 we'll see we can assume $\|v\|_0 \le 2k$

Lemma 6.1. *If $Z_1, \dots, Z_M$ are possibly correlated standard Gaussian random variables, then, for $M \to \infty$,*

$$\max_{1 \le i \le M} Z_i \le 10\sqrt{\log M}$$

with probability at least 0.999.

Applying this lemma, which is proven in Section A.3, with

$$M = \left|\{v \in \{-1, 0, 1\}^n : \|v\|_0 \le 2k\}\right|,$$

we obtain, under the null model,

$$\mathrm{OPT}(Y) \le 20\sqrt{\log M}$$

with probability at least 0.999.

It remains to estimate M. A crude bound gives

$$M = \sum_{j=0}^{2k} \binom{n}{j} 2^j \le (2k+1)\binom{n}{2k} 2^{2k} \overset{k \ge 3}{\le} 2^k \binom{n}{2k} 2^{2k} = \binom{n}{2k} 2^{3k},$$

so

$$\log M \le 2k \log \frac{ne}{2k} + 2k \log 2.$$

Hence, using $\binom{n}{m} \le \left(\frac{ne}{m}\right)^m \quad \forall 1 \le m \le n$, we conclude

$$\mathrm{OPT}(Y) \le 20\sqrt{\log\left(\binom{n}{2k} 2^{3k}\right)} \le 20\sqrt{2k \log \frac{ne}{2k} + 3k \log 2} \le 40\sqrt{k}\sqrt{\log \frac{ne}{2k}}$$

with probability at least 0.999 under the null.

Combining the planted and null estimates, we conclude that $\mathrm{OPT}(Y)$ strongly detects whenever

$$\lambda \ge 50\sqrt{k}\sqrt{\log \frac{ne}{2k}}.$$

Since $k = \rho n$, this becomes

$$\lambda \ge 50\sqrt{\rho}\sqrt{\log \frac{1}{\rho}}\sqrt{n}.$$

When ρ is small, the factor $\sqrt{\rho \log(1/\rho)}$ is much smaller than 1. Thus exhaustive search detects below the PCA threshold. This can be seen as a first indication that sparse PCA exhibits a genuine computational-statistical gap (see Figure 6.2).

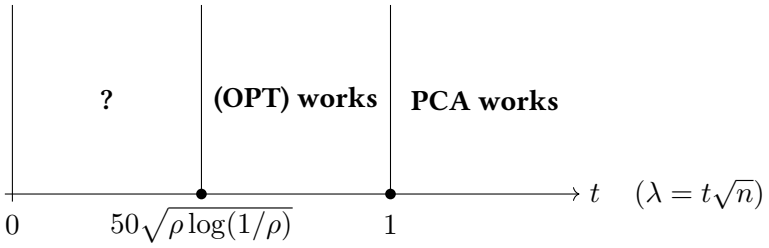

Figure 6.2: Sparse PCA: spectral threshold versus exhaustive-search threshold

6.4 The Likelihood Ratio for Gaussian Additive Models

We now return to the general Gaussian additive model

$$\mathbb{P}_1 : \quad Y = \lambda X + Z, \qquad X \sim \mu, \qquad Z \sim \mathcal{N}(0, I_N),$$

versus

$$\mathbb{P}_2 : \quad Y = Z.$$

The first step is to derive an explicit expression for the likelihood ratio.

Lemma 6.2. *For the Gaussian additive model,*

$$\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)} = \mathop{\mathbb{E}}_{X\sim\mu}\left[\exp\left(-\frac{\lambda^2}{2}\|X\|_2^2 + \lambda\langle X, Y\rangle\right)\right].$$

Proof. Note:

$$\mathbb{P}_2(Y) = \frac{1}{\sqrt{(2\pi)^{N/2}}}\exp\left(-\|Y\|_2^2/2\right)$$

Further, note:

$$\begin{aligned}
\mathbb{P}_1(Y) &= \mathop{\mathbb{E}}_{X\sim\mu}[\mathbb{P}_2(Y - \lambda X)] \\
&= \mathop{\mathbb{E}}_{X\sim\mu}\left[\frac{1}{\sqrt{(2\pi)^{N/2}}}\exp\left(-\|Y - \lambda X\|_2^2/2\right)\right] \\
&= \mathop{\mathbb{E}}_{X\sim\mu}\left[\frac{1}{\sqrt{(2\pi)^{N/2}}}\exp\left(-\|Y\|_2^2/2 - \lambda^2\|X\|_2^2/2 + \lambda\langle Y, X\rangle\right)\right]
\end{aligned}$$

Computing the ratio, we have:

$$\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)} = \mathop{\mathbb{E}}_{X\sim\mu}\left[\exp\left(-\frac{\lambda^2}{2}\|X\|_2^2 + \lambda\langle X, Y\rangle\right)\right]$$

as desired. □

This formula is the starting point for all subsequent calculations. It expresses the likelihood ratio as an average over the prior of a simple exponential function. In particular, it shows that the structure of the planted model enters only through the prior on X.

The same Gaussian computation also gives a closed formula for the full L^2 norm of the likelihood ratio.

Lemma 6.3. *For the Gaussian additive model,*

$$\left\|\frac{\mathbb{P}_1}{\mathbb{P}_2}\right\|_2^2 = \mathop{\mathbb{E}}_{X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mu}\left[\exp\left(\lambda^2\langle X_1, X_2\rangle\right)\right].$$

Proof. Note:

$$\left\|\frac{\mathbb{P}_1}{\mathbb{P}_2}\right\|_2^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[\left(\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)}\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\left[\mathop{\mathbb{E}}_{X\sim\mu}\left[\exp\left(-\frac{\lambda^2}{2}\|X\|_2^2 + \lambda\langle X, Y\rangle\right)\right]^2\right]$$

Now we make use of the "replica trick".

Replica Trick. Recall, it holds for any f that:

$$\left(\mathop{\mathbb{E}}_{X\sim\mu}[f(x)]\right)^2 = \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}(f(X_1)f(X_2))$$

Using this trick, we have:

$$\left\|\frac{\mathbb{P}_1}{\mathbb{P}_2}\right\|_2^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}\mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\exp\left(-\frac{\lambda^2}{2}(\|X_1\|_2^2+\|X_2\|_2^2)\right)\exp\left(\lambda(\langle X_1,Y\rangle+\langle X_2,Y\rangle)\right)\right]$$

By swapping the order of expectations, we obtain:

$$\left\|\frac{\mathbb{P}_1}{\mathbb{P}_2}\right\|_2^2 = \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\exp\left(-\frac{\lambda^2}{2}(\|X_1\|_2^2+\|X_2\|_2^2)\right)\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_2}[\exp(\lambda\langle Y,X_1+X_2\rangle)]\right]$$

Since the inner expectation is a Gaussian MGF, we have:

$$\begin{aligned}\left\|\frac{\mathbb{P}_1}{\mathbb{P}_2}\right\|_2^2 &= \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\exp\left(-\frac{\lambda^2}{2}(\|X_1\|_2^2+\|X_2\|_2^2)\right)\exp\left(\frac{\lambda^2}{2}\|X_1+X_2\|_2^2\right)\right]\\ &= \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\exp\left(\lambda^2\langle X_1,X_2\rangle\right)\right]\end{aligned}$$

as desired. □

This identity is extremely useful. It reduces the full χ^2-divergence to an overlap computation involving two independent samples from the prior.

In the sparse PCA model defined at the beginning of Section 6.3, this formula gives the following result.

Proposition 6.4. *Assume that $\rho \in (0, 1/4)$ is fixed. If*

$$\lambda \le 0.01\sqrt{n\rho\log\frac{1}{\rho}},$$

then strong detection between $\mathbb{P}_1$ and $\mathbb{P}_2$ is impossible.

The proof of the previous theorem is given in Section A.4.

Thus the picture of the sparse PCA phase diagram can already be sharpened: one now sees an information-theoretically impossible regime, an information-theoretically possible but conjecturally hard regime, and the spectral regime where PCA succeeds (see Figure 6.3).

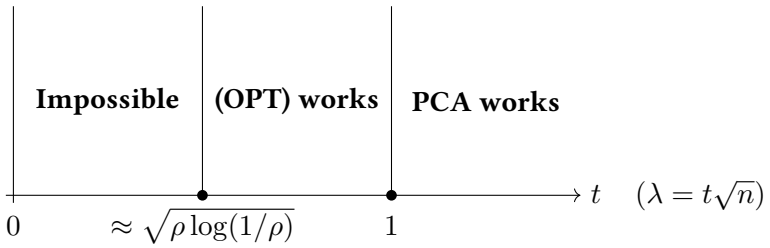

Figure 6.3: Sparse PCA: information-theoretic, exhaustive-search, and spectral regimes

At this point, however, the main computational question remains open: is there a faster algorithm in the region between the information-theoretic threshold and the PCA threshold? This is exactly the type of question for which the low-degree likelihood ratio is designed to provide evidence for.

6.5 The Low-Degree Likelihood Ratio for Gaussian Additive Models

The low-degree norm can also be computed explicitly in Gaussian additive models.

Theorem 6.5. *For the Gaussian additive model,*

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\le D} \right\|_2^2 = \mathop{\mathbb{E}}_{X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mu} \left[\sum_{d=0}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} \langle X_1, X_2 \rangle^d \right].$$

Proof. The proof will be given in Section 7.3 using Hermite polynomials. □

This theorem is the main tool of the chapter. It shows that the low-degree norm is controlled by the moments of the overlap $\langle X_1, X_2 \rangle$. In other words, the question of low-degree hardness is very conveniently reduced to the overlap distribution of the prior.

6.6 Back to Sparse PCA

We now return to sparse PCA and apply the general low-degree formula. In this model,

$$X = \theta\theta^\top,$$

where

$$\begin{cases} 0 & \text{with probability } 1 - \frac{k}{n} = 1 - \rho, \\ \sqrt{\frac{1}{k}} & \text{with probability } \frac{k}{2n}, \\ -\sqrt{\frac{1}{k}} & \text{with probability } \frac{k}{2n}, \end{cases} \qquad \rho \text{ a small constant,}$$

so for two independent draws θ_1, θ_2,

$$\langle X_1, X_2 \rangle = \langle \theta_1, \theta_2 \rangle^2.$$

Hence

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\le D} \right\|_2^2 = \mathop{\mathbb{E}}_{\theta_1, \theta_2} \left[\sum_{d=0}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} \langle \theta_1, \theta_2 \rangle^{2d} \right].$$

Thus, as in the full χ^2 calculation, everything reduces to understanding the overlap $\langle \theta_1, \theta_2 \rangle$.

The key estimate is the following local-Chernoff bound.

Lemma 6.6 (Local-Chernoff bound)**.** *For every $\eta > 0$, there exist $c_\eta > 0$ such that for all n and all $0 < t \le c_\eta \rho = c_\eta \frac{k}{n}$,*

$$\mathop{\mathbb{P}}_{\theta_1, \theta_2} \left(|\langle \theta_1, \theta_2 \rangle| \ge t \right) \le C \exp\left(-\frac{1-\eta}{2} t^2 n \right).$$

The proof can be found in Section A.5

The proof of the low-degree lower bound then proceeds by splitting the expectation into a small-overlap contribution and a large-overlap contribution. The former is controlled by the truncated exponential, and the latter by the local-Chernoff tail estimate.

Theorem 6.7. *Let $\eta > 0$. If*

$$\lambda = \sqrt{1-2\eta}\sqrt{n}$$

and $D = (\log n)^{1.01}$, then

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2^2 = 1 + \mathcal{O}(1).$$

Proof. We already know

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2^2 = \mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=0}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} \langle \theta_1, \theta_2 \rangle^{2d} \right] = 1 + \mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=1}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} \langle \theta_1, \theta_2 \rangle^{2d} \right]$$

We first condition on the event $\mathcal{E} := \{\|\theta_1\|_0, \|\theta_2\|_0 < 2k\} = \left\{\|\theta_1\|_2, \|\theta_2\|_2 < \sqrt{2}\right\}$. By union bound and Chernoff, we know $\mathbb{P}_1(\mathcal{E}^c) \leq 2e^{-\frac{1}{3}k} = 2e^{-\frac{1}{3}\rho n}$. Thus, using Cauchy-Schwarz and a crude bound, we get

$$|\langle \theta_1, \theta_2 \rangle| \leq \|\theta_1\|_2 \|\theta_2\|_2 \leq \frac{n}{k},$$

so on $\mathcal{E}^c$ we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=0}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} \langle \theta_1, \theta_2 \rangle^{2d} \mathbf{1}_{\mathcal{E}^c} \right] &\leq \mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=0}^{D} \frac{n^d}{d! 2^d} \left(\frac{n}{k}\right)^{2d} \mathbf{1}_{\mathcal{E}^c} \right] \leq \mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=0}^{D} n^{3d} \mathbf{1}_{\mathcal{E}^c} \right] \\
&= \frac{n^{3(D+1)} - 1}{n^3 - 1} \mathbb{P}_1(\mathcal{E}^c) \leq n^{3(D+1)} \mathbb{P}_1(\mathcal{E}^c) \\
&\leq 2 \exp\left(3\left((\log n)^{1.01} + 1 \right) \log n - \frac{1}{3}\rho n \right) = o(1).
\end{aligned}$$

Thus, we can work on $\mathcal{E}$ without affecting the aimed result.

Further, with the intention to take advantage of Lemma 6.6, for any $0 < \varepsilon \leq c_\eta \rho$, we have

$$\begin{aligned}
\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2^2 = 1 + o(1) &+ \overbrace{\mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=1}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} \langle \theta_1, \theta_2 \rangle^{2d} \mathbf{1}\left(|\langle \theta_1, \theta_2 \rangle| \leq \varepsilon \right) \mathbf{1}_{\mathcal{E}} \right]}^{R_1} \\
&+ \overbrace{\mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=1}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} \langle \theta_1, \theta_2 \rangle^{2d} \mathbf{1}\left(|\langle \theta_1, \theta_2 \rangle| > \varepsilon \right) \mathbf{1}_{\mathcal{E}} \right]}^{R_2}.
\end{aligned}$$

To bound R_2, considering that $\lambda = \Omega(\sqrt{n})$ in the information-theoretic regime and that we are now conditioned on $\mathcal{E}$, by Cauchy-Schwarz we know $|\langle \theta_1, \theta_2 \rangle| < 2$, so

$$\begin{aligned}
R_2 &\leq \mathop{\mathbb{E}}_{\theta_1,\theta_2} \left[\sum_{d=1}^{D} \frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!} 2^{2d} \mathbf{1}\left(|\langle \theta_1, \theta_2 \rangle| > \varepsilon \right) \right] = \mathbb{P}_1\left(|\langle \theta_1, \theta_2 \rangle| > \varepsilon \right) \sum_{d=1}^{D} \frac{\left(\sqrt{2}\lambda\right)^{2d}}{d!} \\
&\underset{\lambda = \Theta(\sqrt{n})}{\overset{\text{Lemma 6.6}}{\leq}} C \exp\left(-\frac{1-\eta}{2}\varepsilon^2 n \right) D \frac{\left(\sqrt{2}\lambda\right)^{2D}}{D!} \leq C \exp\left(-\frac{1-\eta}{2}\varepsilon^2 n + D \log\left(2\lambda^2\right) \right) = o(1).
\end{aligned}$$

For R_1:

$$
\begin{aligned}
R_1 &= \mathop{\mathbb{E}}_{\theta_1,\theta_2}\left[\mathbf{1}\left(|\langle\theta_1,\theta_2\rangle| \le \varepsilon\right)\sum_{d=1}^{D}\frac{\left(\lambda/\sqrt{2}\right)^{2d}}{d!}\langle\theta_1,\theta_2\rangle^{2d}\mathbf{1}_{\mathcal{E}}\right] \\
&\le \mathop{\mathbb{E}}_{\theta_1,\theta_2}\left[\mathbf{1}\left(\langle\theta_1,\theta_2\rangle^2 \le \varepsilon^2\right)\left(\exp\left(\frac{\lambda^2}{2}\langle\theta_1,\theta_2\rangle^2\right)-1\right)\right] \\
&= \mathop{\mathbb{E}}_{\theta_1,\theta_2}\left[\mathbf{1}\left(\langle\theta_1,\theta_2\rangle^2 \le \varepsilon^2\right)\int_0^{\langle\theta_1,\theta_2\rangle^2}\frac{\lambda^2}{2}\exp\left(\frac{\lambda^2}{2}x\right)\mathrm{d}x\right] \\
&= \mathop{\mathbb{E}}_{\theta_1,\theta_2}\left[\int_0^{\varepsilon^2}\frac{\lambda^2}{2}\exp\left(\frac{\lambda^2}{2}x\right)\mathbf{1}\left(\langle\theta_1,\theta_2\rangle^2 \le \varepsilon^2\right)\mathbf{1}\left(x \le \langle\theta_1,\theta_2\rangle^2\right)\mathrm{d}x\right] \\
&\le \mathop{\mathbb{E}}_{\theta_1,\theta_2}\left[\int_0^{\varepsilon^2}\frac{\lambda^2}{2}\exp\left(\frac{\lambda^2}{2}x\right)\mathbf{1}\left(x \le \langle\theta_1,\theta_2\rangle^2\right)\mathrm{d}x\right] \\
&= \int_0^{\varepsilon^2}\frac{\lambda^2}{2}\exp\left(\frac{\lambda^2}{2}x\right)\mathbb{P}_1\left(x \le \langle\theta_1,\theta_2\rangle^2\right)\mathrm{d}x \\
&\overset{\text{Lemma 6.6}}{\le} C\frac{\lambda^2}{2}\int_0^{\varepsilon^2}\exp\left(\frac{\lambda^2}{2}x-\frac{1-\eta}{2}nx\right)\mathrm{d}x \\
&= C\frac{\lambda^2}{2}\int_0^{+\infty}\exp\left(-\frac{\eta}{2}nx\right)\mathrm{d}x \\
&= \frac{C\lambda^2}{\eta n} = \mathcal{O}\left(\frac{\lambda^2}{n}\right) \overset{\lambda=\Theta(\sqrt{n})}{=} \mathcal{O}(1).
\end{aligned}
$$

□

This theorem is the main conclusion of the chapter. It says that the low-degree norm remains bounded below the spectral threshold. Under the low-degree conjecture, this is strong evidence that polynomial-time detection should fail there. Thus, the phase transition diagram should be the one depicted in Figure 6.4

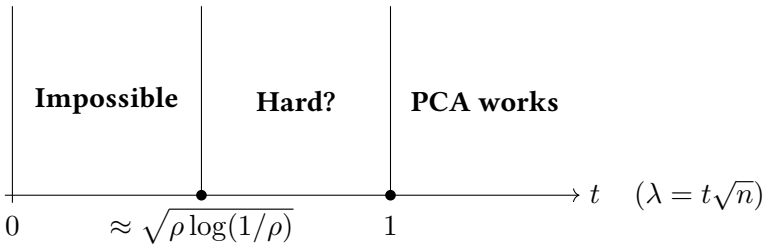

Figure 6.4: Sparse PCA: phase transition diagram

6.7 Discussion

This chapter had two main goals. The first was to show, through sparse PCA, that the computational and information-theoretic thresholds need not coincide. The second was to derive general formulas for Gaussian additive models that reduce low-degree computations to overlap calculations under the prior.

The next chapter will provide the missing technical ingredient behind these formulas. It will introduce the Hermite-polynomial machinery that completes the missing proofs above, and will then use it to study also the computational-statistical gap for Tensor PCA.

Chapter 7

Hermite Analysis and Tensor PCA

7.1 Introduction

In Chapter 6 we derived the low-degree likelihood ratio formula for Gaussian additive models:

$$\left\| \left(\frac{P_1}{P_2} \right)_{\leq D} \right\|_2^2 = \mathop{\mathbb{E}}_{X_1, X_2 \sim \mu} \left[\sum_{d=0}^{D} \frac{\lambda^{2d}}{d!} \langle X_1, X_2 \rangle^d \right].$$

This formula reduces the analysis of low-degree tests to a question about the overlap distribution of two independent samples from the prior.

The purpose of this chapter is twofold. First, we provide the Hermite-polynomial machinery that justifies the Gaussian additive formula used in the previous chapter. Second, we apply the same framework to tensor PCA. Tensor PCA provides another fundamental example of a Gaussian additive detection problem with a computational-statistical gap.

7.2 Background on Hermite polynomials

The purpose of this section is to introduce Hermite polynomials as a convenient orthonormal basis for low-degree polynomials under the standard Gaussian measure. This is exactly the structure needed to understand the low-degree likelihood ratio in Gaussian additive models.

7.2.1 Orthonormality in the univariate case

We begin with the one-dimensional case.

Definition 7.1. *The k-th Hermite polynomials are defined recursively by*

$$h_0(x) \equiv 1, \qquad h_{k+1}(x) = x h_k(x) - h_k'(x).$$

The normalized k-th Hermite polynomials are

$$\widehat{h}_k(x) = \frac{h_k(x)}{\sqrt{k!}}.$$

In particular, $\widehat{h}_k$ is a polynomial of degree k.

The basic fact is that Hermite polynomials form an orthonormal basis under the standard Gaussian inner product.

Proposition 7.2. *The set* $\left\{\widehat{h}_k(x) : k \in \mathbb{N}\right\}$ *is an orthonormal basis of* $L^2(\mathcal{N}(0,1))$ *under the inner product*

$$\langle f_1, f_2 \rangle = \mathop{\mathbb{E}}_{Y \sim \mathcal{N}(0,1)} [f_1(Y) f_2(Y)].$$

Proof. See [KWB19, Proposition B.4.]. □

Orthonormality means that for every $k, \ell \in \mathbb{N}$,

$$\left\langle \widehat{h}_k, \widehat{h}_\ell \right\rangle = \begin{cases} 1 & \text{if } k = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, any $f : \mathbb{R} \to \mathbb{R}$ with $\|f\|_2 = \sqrt{\langle f, f \rangle} < +\infty$ admits an expansion

$$f(y) = \sum_{k \geq 0} \alpha_k \widehat{h}_k(y),$$

and the coefficients are given by

$$\alpha_\ell = \left\langle f, \widehat{h}_\ell \right\rangle.$$

Indeed,

$$\left\langle \sum_{k \geq 0} \alpha_k \widehat{h}_k, \widehat{h}_\ell \right\rangle = \sum_{k \geq 0} \alpha_k \left\langle \widehat{h}_k, \widehat{h}_\ell \right\rangle = \alpha_\ell.$$

Therefore

$$f(y) = \sum_{k \geq 0} \left\langle f, \widehat{h}_k \right\rangle \widehat{h}_k(y).$$

Since each $\widehat{h}_k$ has degree k, it follows that if $p \in \mathbb{R}_{\leq D}[y]$, then

$$\left\langle p, \widehat{h}_k \right\rangle = 0 \qquad \text{for all } k > D.$$

Hence $\left\{\widehat{h}_k : k \leq D\right\}$ is an orthonormal basis of $\mathbb{R}_{\leq D}[y]$.

The next lemma identifies the orthogonal projection onto the degree-$\leq D$ polynomial space.

Lemma 7.3. *For every* $f \in L^2(\mathcal{N}(0,1))$*, the minimization problem*

$$\min_{p \in \mathbb{R}_{\leq D}[y]} \|p - f\|_2^2$$

has a unique minimizer, given by

$$p(y) = \sum_{k=0}^{D} \left\langle f, \widehat{h}_k \right\rangle \widehat{h}_k(y).$$

Proof. We use Parseval's identity, which states that for every $f \in L^2(\mathcal{N}(0,1))$,

$$\|f\|_2^2 = \sum_{k \geq 0} \left| \left\langle \widehat{h}_k, f \right\rangle \right|^2.$$

Indeed,

$$\|f\|_2^2 = \langle f, f\rangle = \left\langle \sum_k \left\langle f, \widehat{h}_k \right\rangle \widehat{h}_k, \sum_\ell \left\langle f, \widehat{h}_\ell \right\rangle \widehat{h}_\ell \right\rangle$$
$$= \sum_{k,\ell} \left\langle f, \widehat{h}_k \right\rangle \left\langle f, \widehat{h}_\ell \right\rangle \left\langle \widehat{h}_k, \widehat{h}_\ell \right\rangle = \sum_k \left\langle f, \widehat{h}_k \right\rangle^2.$$

Now let $p \in \mathbb{R}_{\le D}[y]$. Then

$$\|f - p\|_2^2 = \sum_{k\ge 0} \left\langle f - p, \widehat{h}_k \right\rangle^2.$$

Since p has degree at most D, we have $\left\langle p, \widehat{h}_k \right\rangle = 0$ for all $k > D$, so

$$\|f - p\|_2^2 = \sum_{k>D} \left\langle f, \widehat{h}_k \right\rangle^2 + \sum_{k\le D} (\left\langle f, \widehat{h}_k \right\rangle - \left\langle p, \widehat{h}_k \right\rangle)^2.$$

The first term is independent of p, so the minimum is achieved exactly when

$$\left\langle p, \widehat{h}_k \right\rangle = \left\langle f, \widehat{h}_k \right\rangle \qquad \text{for all } k \le D.$$

Thus the unique minimizer is

$$p(y) = \sum_{k=0}^{D} \left\langle f, \widehat{h}_k \right\rangle \widehat{h}_k(y).$$

□

In particular, if we denote by $f_{\le D}$ the orthogonal projection of f onto $\mathbb{R}_{\le D}[y]$, then

$$f_{\le D}(y) = \sum_{k=0}^{D} \left\langle f, \widehat{h}_k \right\rangle \widehat{h}_k(y),$$

and

$$\|f_{\le D}\|_2^2 = \sum_{k=0}^{D} \left\langle f, \widehat{h}_k \right\rangle^2.$$

7.2.2 The multivariate case

We now extend the previous discussion to the multivariate Gaussian measure.

Definition 7.4. *Let $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{N}^N$. The N-variate Hermite polynomial associated with α is*

$$H_\alpha(X) := \prod_{i=1}^{N} h_{\alpha_i}(X_i).$$

Its normalized version is

$$\widehat{H}_\alpha(X) := \frac{H_\alpha(X)}{\sqrt{\alpha_1! \cdots \alpha_N!}}.$$

Let $Y \sim \mathcal{N}(0, I_N)$, so that $Y_1, \dots, Y_N$ are i.i.d. $\mathcal{N}(0,1)$. By independence and the univariate orthonormality,

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(0,I_N)} \left[\widehat{H}_\alpha(Y)\widehat{H}_\beta(Y)\right] = \mathbf{1}(\alpha = \beta).$$

The multivariate analogue of the univariate basis theorem is then the following.

Proposition 7.5. *The family* $\left\{\widehat{H}_\alpha(X) : \alpha \in \mathbb{N}^N\right\}$ *is an orthonormal basis of* $L^2\left(\mathcal{N}\left(0, I_N\right)\right)$.

Proof. The proof is analogous to the univariate case. □

The corresponding projection theorem also carries over.

Proposition 7.6. *Fix* $f \in L^2(\mathcal{N}(0, I_N))$. *Then the orthogonal projection*

$$\min_{p \in \mathbb{R}_{\leq D}[Y]} \|p - f\|_2$$

is uniquely attained by

$$f_{\leq D} = \sum_{\alpha \in \mathbb{N}^N,\ |\alpha| \leq D} \left\langle \widehat{H}_\alpha, f \right\rangle \widehat{H}_\alpha.$$

Moreover,

$$\|f_{\leq D}\|_2^2 = \sum_{\alpha \in \mathbb{N}^N,\ |\alpha| \leq D} \left\langle \widehat{H}_\alpha, f \right\rangle^2.$$

Proof. Again, the proof is the same as in the univariate setting, using Parseval's identity and orthogonality. □

7.3 Proof of the low-degree ratio for Gaussian additive models

We now return to the Gaussian additive model from the previous chapter. Recall the setting:

$$\mathbb{P}_1 : \ Y = \lambda X + Z, \qquad X \sim \mu, \qquad Z \sim \mathcal{N}(0, I_N),$$

and

$$\mathbb{P}_2 : \ Y = Z.$$

Our goal is to compute

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2^2.$$

The Hermite basis provides the correct way to do this. By Parseval's identity in the multivariate Gaussian space,

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2^2 = \sum_{\alpha \in \mathbb{N}^N,\ |\alpha| \leq D} \left\langle \frac{\mathbb{P}_1}{\mathbb{P}_2}, \widehat{H}_\alpha \right\rangle^2.$$

Since the inner product is with respect to $\mathbb{P}_2$,

$$\left\langle \frac{\mathbb{P}_1}{\mathbb{P}_2}, \widehat{H}_\alpha \right\rangle = \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_2} \left[\frac{\mathbb{P}_1(Y)}{\mathbb{P}_2(Y)} \widehat{H}_\alpha(Y) \right] = \mathop{\mathbb{E}}_{Y \sim \mathbb{P}_1} \left[\widehat{H}_\alpha(Y) \right].$$

Therefore

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2^2 = \sum_{|\alpha| \leq D} \left(\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_1} \left[\widehat{H}_\alpha(Y) \right] \right)^2.$$

Under $\mathbb{P}_1$, we have $Y = \lambda X + Z$ with $X \sim \mu$ and $Z \sim \mathcal{N}(0, I_N)$. Hence

$$\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_1} [\widehat{H}_\alpha(Y)] = \mathop{\mathbb{E}}_{X \sim \mu} \left[\mathop{\mathbb{E}}_{Z \sim \mathcal{N}(0, I_N)} \left[\prod_{i=1}^N \widehat{h}_{\alpha_i}(\lambda X_i + Z_i) \right] \right].$$

At this point we use the following identity.

Lemma 7.7 (Translation identity, [KWB19, Proposition 2.9.]). *If* $Y \sim \mathcal{N}(\mu, 1)$, *then*

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}\left[\widehat{h}_k(Y)\right] = \frac{\mu^k}{\sqrt{k!}}.$$

Proof. It is enough to prove that

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k(Y)] = \mu^k,$$

since $\widehat{h}_k = h_k/\sqrt{k!}$.

We argue by induction on k. For $k = 0$, we have $h_0(x) \equiv 1$, so

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_0(Y)] = 1 = \mu^0.$$

Now assume that

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k(Y)] = \mu^k$$

for some $k \geq 0$. We want to prove that

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_{k+1}(Y)] = \mu^{k+1}.$$

By the defining recursion,

$$h_{k+1}(x) = xh_k(x) - h_k'(x),$$

hence

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_{k+1}(Y)] = \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[Yh_k(Y)] - \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}\left[h_k'(Y)\right].$$

Now write the expectation with respect to the density of $Y \sim \mathcal{N}(\mu, 1)$:

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[Yh_k(Y)] = \int_{\mathbb{R}} yh_k(y)\frac{1}{\sqrt{2\pi}}e^{-(y-\mu)^2/2}\,dy.$$

Since

$$y = \mu + (y - \mu),$$

this becomes

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[Yh_k(Y)] = \mu \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k(Y)] + \int_{\mathbb{R}} (y-\mu)h_k(y)\frac{1}{\sqrt{2\pi}}e^{-(y-\mu)^2/2}\,dy.$$

For the second term, observe that

$$\frac{d}{dy}\left(e^{-(y-\mu)^2/2}\right) = -(y-\mu)e^{-(y-\mu)^2/2}.$$

Hence

$$\int_{\mathbb{R}} (y-\mu)h_k(y)\frac{1}{\sqrt{2\pi}}e^{-(y-\mu)^2/2}\,dy = -\int_{\mathbb{R}} h_k(y)\frac{1}{\sqrt{2\pi}}\frac{d}{dy}\left(e^{-(y-\mu)^2/2}\right)\,dy.$$

Integrating by parts, and using that the boundary term vanishes because h_k is a polynomial while the Gaussian density decays super-exponentially, we get

$$-\int_{\mathbb{R}} h_k(y)\frac{1}{\sqrt{2\pi}}\frac{d}{dy}\left(e^{-(y-\mu)^2/2}\right)\,dy = \int_{\mathbb{R}} h_k'(y)\frac{1}{\sqrt{2\pi}}e^{-(y-\mu)^2/2}\,dy = \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k'(Y)].$$

Therefore

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[Yh_k(Y)] = \mu \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k(Y)] + \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k'(Y)].$$

Substituting this into

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_{k+1}(Y)] = \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[Yh_k(Y)] - \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k'(Y)],$$

we obtain

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_{k+1}(Y)] = \mu \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(\mu,1)}[h_k(Y)] = \mu\cdot\mu^k = \mu^{k+1}.$$

This completes the induction. □

Applying the translation identity coordinatewise gives

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,I_N)}\left[\prod_{i=1}^N \widehat{h}_{\alpha_i}(\lambda X_i + Z_i)\right] = \prod_{i=1}^N \frac{(\lambda X_i)^{\alpha_i}}{\sqrt{\alpha_i!}}.$$

Therefore

$$\begin{aligned}
\left\|\begin{pmatrix}\mathbb{P}_1\\ \mathbb{P}_2\end{pmatrix}_{\le D}\right\|_2^2 &= \sum_{|\alpha|\le D}\left(\mathop{\mathbb{E}}_{X\sim\mu}\left[\prod_{i=1}^N \frac{(\lambda X_i)^{\alpha_i}}{\sqrt{\alpha_i!}}\right]\right)^2 \\
&= \sum_{|\alpha|\le D}\left(\mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\prod_{i=1}^N \frac{(\lambda X_{1,i})^{\alpha_i}(\lambda X_{2,i})^{\alpha_i}}{\alpha_i!}\right]\right) \\
&= \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\sum_{d=0}^D\left(\sum_{\alpha\in\mathbb{N}^N,\,|\alpha|=d}\left(\prod_{i=1}^N \frac{\left(\lambda^2 X_{1,i}X_{2,i}\right)^{\alpha_i}}{\alpha_i!}\right)\right)\right] \\
&= \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\sum_{d=0}^D \frac{1}{d!}\left(\sum_{i=1}^N \lambda^2 X_{1,i}X_{2,i}\right)^d\right] \\
&= \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\sum_{d=0}^D \frac{\lambda^{2d}}{d!}\langle X_1, X_2\rangle^d\right].
\end{aligned}$$

We have therefore proved the main formula.

Theorem 7.8. *For every $D \ge 0$,*

$$\left\|\begin{pmatrix}\mathbb{P}_1\\ \mathbb{P}_2\end{pmatrix}_{\le D}\right\|_2^2 = \mathop{\mathbb{E}}_{X_1,X_2\overset{\text{i.i.d.}}{\sim}\mu}\left[\sum_{d=0}^D \frac{1}{d!}\lambda^{2d}\langle X_1, X_2\rangle^d\right].$$

This is exactly the formula that was used in the previous chapter.

7.4 Tensor PCA

We now turn to another Gaussian additive model: Tensor PCA. This example is important because it provides a computational-statistical gap in a setting different from matrices, while still fitting cleanly into the same low-degree framework.

7.4.1 The model

Definition 7.9. *A p-tensor is an element of $(\mathbb{R}^n)^{\otimes p}$, that is, an array*

$$T = (T_{i_1,\dots,i_p})_{i_1,\dots,i_p \in [n]}.$$

For $X \in \mathbb{R}^n$, the rank-one tensor $X^{\otimes p}$ is defined by

$$(X^{\otimes p})_{i_1,\dots,i_p} = X_{i_1} X_{i_2} \cdots X_{i_p}.$$

The Tensor PCA detection problem compares

$$\mathbb{P}_1 : \ X \sim \mathrm{Unif}(\{\pm 1\}^n) \text{ or } X \sim \mathrm{Unif}\left(\mathbb{S}^{n-1}\right), \quad Y = \lambda X^{\otimes p} + Z, \quad \lambda = \lambda_n > 0, \quad Z_{i_1,\dots,i_p} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

with

$$\mathbb{P}_2 : \ Y = Z.$$

7.4.2 Information-theoretic threshold

A natural way to detect is to maximize

$$(\mathrm{OPT}) = \max_{v \in \{\pm 1\}^n} \left\langle v^{\otimes p}, Y \right\rangle := \sum_{i_1,\dots,i_p} v_{i_1} \cdots v_{i_p} Y_{i_1,\dots,i_p}.$$

This serves as the analogue of the MAP estimator in the planted model. Indeed,

$$\left\langle v^{\otimes p}, Y \right\rangle = \left\langle v^{\otimes p}, \lambda X^{\otimes p} + Z \right\rangle = \lambda \left\langle v^{\otimes p}, X^{\otimes p} \right\rangle + \left\langle v^{\otimes p}, Z \right\rangle = \lambda \langle v, X \rangle^p + \left\langle v^{\otimes p}, Z \right\rangle.$$

Under $\mathbb{P}_1$, taking $v = X$ gives

$$(\mathrm{OPT}) \geq \left\langle X^{\otimes p}, Y \right\rangle = \lambda \|X\|_2^{2p} + \left\langle X^{\otimes p}, Z \right\rangle.$$

Moreover,

$$\left\langle X^{\otimes p}, Z \right\rangle \sim \mathcal{N}\left(0, \|X\|_2^{2p}\right),$$

so when $\|X\|_2^2 = n$ we obtain

$$(\mathrm{OPT}) \geq \lambda n^p - \mathcal{O}\left(n^{p/2}\right).$$

Under $\mathbb{P}_2$, for every fixed $v \in \{\pm 1\}^n$,

$$\left\langle v^{\otimes p}, Z \right\rangle \sim \mathcal{N}\left(0, n^p\right).$$

Since there are 2^n such vectors, the Gaussian maximal inequality gives

$$(\mathrm{OPT}) \leq n^{p/2} \sqrt{n \log 2}$$

up to constants, that is,

$$(\mathrm{OPT}) \leq n^{p/2 + 1/2}$$

up to constants.

Comparing the planted and null behaviors, we see that strong detection is possible when

$$\lambda n^p \gg n^{p/2+1/2},$$

that is,

$$\lambda \gg n^{1/2 - p/2}.$$

This threshold (see Figure 7.1) can in fact be shown to be tight.

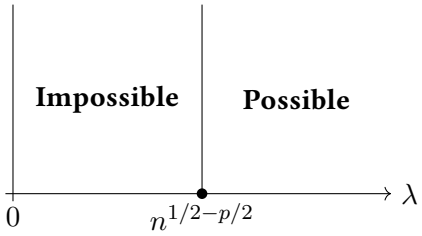

Figure 7.1: Information-theoretic threshold in tensor PCA

7.4.3 A natural polynomial-time algorithm

We now compare the information-theoretic threshold with what a natural efficient algorithm can achieve. Assume for simplicity that p is even.

Flatten the tensor into a matrix

$$M \in \mathbb{R}^{n^{p/2} \times n^{p/2}}$$

by setting

$$M_{(i_1,\dots,i_{p/2}),(i_{p/2+1},\dots,i_p)} := Y_{i_1,\dots,i_p}.$$

Then

$$M_{IJ} = \lambda \theta_I \theta_J + Z_{IJ},$$

wher

$$\theta_I := \prod_{i \in I} \theta_i.$$

Now

$$\|\theta\|_2^2 = \sum_I \theta_I^2 = n^{p/2}.$$

If we define

$$\theta' := \frac{\theta}{\|\theta\|_2},$$

then the matrix model becomes

$$M_{IJ} = \lambda n^{p/2} \theta'_I \theta'_J + Z_{IJ}.$$

This is now exactly a spiked matrix model. Therefore PCA on M works when

$$\lambda n^{p/2} \gg \sqrt{n^{p/2}},$$

that is,

$$\lambda \gg n^{-p/4}.$$

This gives the natural polynomial-time threshold in Tensor PCA.

7.4.4 A computational-statistical gap

Comparing the two thresholds, we find a clear gap: information-theoretically, strong detection is possible when $\lambda \gg n^{1/2-p/2}$; algorithmically, the natural PCA-based method works when $\lambda \gg n^{-p/4}$.

For $p \geq 3$, these are genuinely different scales. Tensor PCA is therefore another model exhibiting a computational-statistical gap (see Figure 7.2).

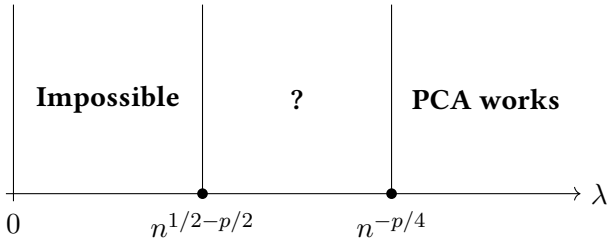

Figure 7.2: Tensor PCA: computational-statistical gap

7.4.5 A low-degree lower bound

The low-degree method predicts hardness in the intermediate regime. The proof is taken from the survey [KWB19, Section 3.1].

Proposition 7.10. *If $D \le (\log n)^2$ and $\lambda \le n^{-p/4-\varepsilon/2}$ for some $\varepsilon > 0$, then*

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D} \right\|_2^2 = 1 + o(1).$$

Proof. By the Gaussian-additive-model formula,

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D} \right\|_2^2 = \mathbb{E}\left[\sum_{d=0}^{D} \frac{\lambda^{2d}}{d!} \langle X_1, X_2\rangle^d\right].$$

In the Tensor PCA model, $X_1 = \theta_1^{\otimes p}$ and $X_2 = \theta_2^{\otimes p}$, so

$$\langle X_1, X_2\rangle = \langle \theta_1, \theta_2\rangle^p .$$

Therefore

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D} \right\|_2^2 = \mathop{\mathbb{E}}_{\theta_1,\theta_2}\left[\sum_{d=0}^{D} \frac{\lambda^{2d}}{d!} \langle \theta_1, \theta_2\rangle^{dp}\right].$$

Now $\theta_1, \theta_2 \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{\pm 1\}^n)$, so $\langle \theta_1, \theta_2\rangle$ is approximately Gaussian with variance n (see [KWB19, Appendix C]). Hence

$$\mathop{\mathbb{E}}_{\theta_1,\theta_2}[\langle \theta_1, \theta_2\rangle^{dp}] \approx (2dp-1)!!\, n^{dp/2}.$$

Using $(2dp-1)!! \le (2Dp)^{dp}$ for $d \le D$, we obtain

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D} \right\|_2^2 \le 1 + \sum_{d=1}^{D} \frac{\lambda^{2d}}{d!} (2Dp)^{dp} n^{dp/2}.$$

Since $d! \ge 1$, this is at most

$$1 + D\sum_{d=1}^{D} \left((2Dp)^p \lambda^2 n^{p/2}\right)^d .$$

Under the assumption $\lambda^2 \le n^{-p/2-\varepsilon}$, this yields

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le D} \right\|_2^2 \le 1 + \mathcal{O}\left(D(2Dp)^p n^{-\varepsilon}\right).$$

If $D = (\log n)^2$, then the right-hand side tends to 1, proving the claim. □

Thus the low-degree norm is asymptotically trivial throughout the regime

$$\lambda \le n^{-p/4-\varepsilon/2}.$$

Under the low-degree conjecture, this means that polynomial-time detection should fail there. The low-degree threshold therefore aligns with the natural algorithmic threshold rather than with the information-theoretic one (see Figure 7.3).

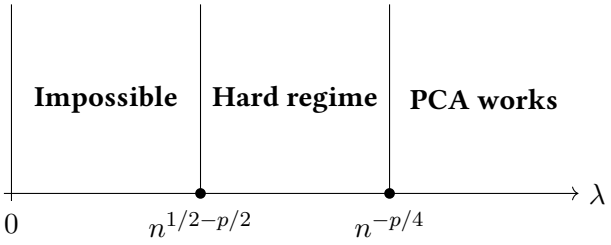

Figure 7.3: Tensor PCA: information-theoretic, hard and spectral regimes

7.5 Other examples

Before concluding the low-degree chapters, it is worth mentioning two additional models in which the same low-degree prediction has been successfully applied.

7.5.1 Gaussian sparse regression

The model is

$$\mathbb{P}_1 : \{(y_i, X_i),\ i = 1, \dots, n\}, \qquad y_i = \langle X_i, \theta \rangle + w_i,$$

where

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p), \qquad \theta \sim \text{Unif}\,\{0,1\}^p : \|\theta\|_0 = k, \qquad w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

versus

$$\mathbb{P}_2 : \{(y_i, X_i),\ i = 1, \dots, n\}, \qquad y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, k+1), \qquad X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p).$$

The current status is the following: for $k < n < k \log p$, it is not known whether a polynomial-time algorithm succeeds, while the low-degree ratio remains bounded (see [BEH+22]). On the other hand, for $n > k \log p$, methods such as LASSO and related algorithms are effective.

7.5.2 Planted clique

The model is

$$\mathbb{P}_1 : G = \mathcal{G}_{n,\frac{1}{2},k} \equiv \mathcal{G}_{n,\frac{1}{2}} \cup \text{clique}(S), \qquad S \sim \text{Unif}\left(\binom{V(G)}{k}\right),$$

versus

$$\mathbb{P}_2 : G = \mathcal{G}_{n,\frac{1}{2}}.$$

Here we know that detection is impossible for $k < 2\log_2 n$. We now prove that when $\log n \ll k \ll \sqrt{n}$, then the low-degree ratio remains bounded, but no polynomial-time algorithm is known. The following is a classical but important low-degree lower bound computation.

Proposition 7.11. *Fix* $0 < \delta < \frac{1}{2}$ *and* $0 < \varepsilon < 1$. *Consider the planted clique detection problem with*

$$\log_2 n \ll k \leq n^{1/2-\delta}.$$

Let

$$D = (\log n)^{1+\varepsilon}.$$

Then

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq D} \right\|_2^2 = 1 + o(1).$$

The proof can be found in Section A.6.

These two examples are only mentioned briefly here, but they foreshadow later chapters of the notes. In both cases, low-degree calculations point to a conjectural hard regime that matches the current algorithmic picture.

7.6 Conclusion and transition

This chapter completed the low-degree analysis of Gaussian additive models. The main technical tool was the Hermite basis, which is the natural orthonormal basis for polynomial functions under the standard Gaussian measure. Since the low-degree likelihood ratio is defined as the orthogonal projection of the likelihood ratio onto the space of low-degree polynomials, Hermite analysis gives an explicit way to compute its norm.

The next chapters develop a different formal model of restricted computation: the statistical query framework. Unlike the low-degree method, which studies polynomial functions of the data, the statistical query model restricts algorithms to accessing the data distribution only through noisy expectations of bounded query functions. This gives a rigorous oracle model in which one can prove unconditional lower bounds for broad classes of algorithms. We will see that, in several high-dimensional detection problems, the statistical query lower bounds agree closely with the predictions obtained from low-degree likelihood ratios.

Chapter 8

The Statistical Query Framework

8.1 Introduction

In the previous chapters, we developed the low-degree likelihood-ratio framework as a structured way to reason about computational hardness in detection problems. That framework is extremely useful, but it is (by far) not the only family of efficient methods that are powerful in these settings.

The goal of the present chapter is to introduce another celebrated such framework: the statistical-query (SQ) model, introduced in [Kea98]. The philosophy is simple. In many algorithmic procedures, one does not manipulate the full dataset in a completely arbitrary way. Instead, one repeatedly estimates expectations of simple functions of a single sample. This is the behavior captured by the SQ model and, more specifically, by the $\mathrm{VSTAT}(m)$ oracle.

The chapter has two main parts. First, we introduce the SQ model, discuss why it is expressive, and explain the notion of statistical dimension that controls its power. Second, we apply the general lower-bound framework to (sublinear) sparse PCA and show that SQ methods correctly predicts the hard regime there.

8.2 Detection in an i.i.d. setting

The statistical-query framework is most naturally formulated when the data come as i.i.d. samples. We therefore begin with the following special case of the detection problem discussed above.

We observe samples $X_1, \dots, X_m$ and test between the hypotheses

$$X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} p_\theta, \qquad \theta \sim \mu,$$

and

$$X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} q,$$

where q is the null distribution.

Equivalently, if we define the planted mixture

$$p := \mathop{\mathbb{E}}_{\theta \sim \mu} [p_\theta],$$

then the problem is to distinguish $p^{\otimes m}$ from $q^{\otimes m}$.

This is the natural i.i.d. restriction of the planted-vs-null detection setup studied earlier in the notes.

8.2.1 Example: sparse linear regression

Sparse regression fits this setting directly. Let

$$\theta \sim \mathrm{Unif}\{u \in \{0,1\}^p : \|u\|_0 = k\}.$$

Under the planted model, for $i = 1, \dots, m$,

$$(x_i, y_i) \overset{\text{i.i.d.}}{\sim} p_\theta,$$

where

$$x_i \sim \mathcal{N}(0, I_p), \qquad y_i = \langle x_i, \theta \rangle + w_i, \qquad w_i \sim \mathcal{N}\left(0, \sigma^2\right),$$

with w_i independent of x_i.

Under the null model,

$$(x_i, y_i) \overset{\text{i.i.d.}}{\sim} q,$$

where

$$x_i \sim \mathcal{N}(0, I_p), \qquad y_i \sim \mathcal{N}\left(0, k + \sigma^2\right),$$

independently of x_i.

Thus sparse regression is naturally an i.i.d. detection problem.

8.2.2 Example: PCA models and cloning

PCA models are less obviously i.i.d., since one usually observes a single matrix

$$Y = \lambda x x^\top + W.$$

However, as already discussed earlier in Section 4.4, the cloning point of view allows us to replace the one-sample model by an equivalent multi-sample version.

Concretely, one may instead think of observing

$$Y_i = \frac{\lambda}{\sqrt{n}} x x^\top + W_i, \qquad i = 1, \dots, m,$$

where the W_i are i.i.d. Gaussian noise matrices, versus the corresponding null model

$$Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d).$$

This does not change the statistical content of the problem, but now the model has been recast into an i.i.d. form, which makes the SQ point of view applicable.

This example is useful because it shows that the i.i.d. formulation is not as restrictive as it may initially appear.

8.3 Statistical queries

8.3.1 Motivation

Suppose we observe i.i.d. samples $X_1, \dots, X_m \sim D$, where D is either the planted law p_θ or the null law q. A very common computational primitive is to estimate quantities of the form

$$\mathop{\mathbb{E}}_{X \sim D}[h(X)]$$

for some function $h : \Omega \to [0,1]$.

The standard empirical estimator is

$$\frac{1}{m}\sum_{i=1}^{m} h(X_i).$$

By Bernstein's inequality, with probability at least 0.99 its error is bounded by

$$\left|\frac{1}{m}\sum_{i=1}^{m} h(X_i) - \mathop{\mathbb{E}}_{X\sim D}[h(X)]\right| \le \frac{1}{m}\max\left\{1, \sqrt{m}\sqrt{\mathop{\mathbb{E}}_{X\sim D}[h(X)]\left(1 - \mathop{\mathbb{E}}_{X\sim D}[h(X)]\right)}\right\}.$$

This suggests an abstract model in which the algorithm does not access samples directly, but instead queries approximate expectations of bounded functions of a single sample.

8.3.2 Definition of the $\mathrm{VSTAT}(m)$ oracle

Definition 8.1. *An $\mathrm{SQ}(m)$ algorithm for a distribution D has access to D through queries to a $\mathrm{VSTAT}(m)$ oracle. Given a function $h : \Omega \to [0,1]$, the oracle returns*

$$\mathop{\mathbb{E}}_{X\sim D}[h(X)] + \tau,$$

where the error τ is adversarial but satisfies

$$|\tau| \le \frac{1}{m}\max\left\{1, \sqrt{m}\sqrt{\mathop{\mathbb{E}}_{X\sim D}[h(X)]\left(1 - \mathop{\mathbb{E}}_{X\sim D}[h(X)]\right)}\right\}.$$

The parameter m should be thought of as a sample size. The number of queries made by the algorithm should be thought of as a proxy for computational time.

8.3.3 Why the SQ model is expressive

The SQ model is quite expressive. Many common algorithms can be interpreted as repeatedly querying expectations of simple functions.

A particularly important example is gradient descent. Suppose one wants to solve

$$\min_{w\in\mathcal{W}} \mathop{\mathbb{E}}_{X\sim D}[L(w;X)], \qquad D \in \{p_\theta, q\},$$

for some loss function L. The population gradient is

$$\mathop{\mathbb{E}}_{X\sim D}[\nabla_w L(w;X)].$$

A gradient-descent iteration takes the form

$$w_{t+1} = w_t - \eta \mathop{\mathbb{E}}_{X\sim D}[\nabla_w L(w_t;X)].$$

Thus each gradient step only requires access to expectations of functions of a single sample. In this sense, noisy gradient descent naturally fits inside the SQ model.

The same is true, at least at an abstract level, for many first-order methods, EM-type procedures, and other algorithms driven by empirical averages.

8.3.4 Limitations of the SQ model

Despite its expressive power, the SQ model is not a perfect representation of all polynomial-time algorithms.

First, queries are only allowed to depend on a single sample. Many natural algorithms compare several samples simultaneously; in principle, SQ cannot do this, and this can be problematic in principle.

Second, there is no requirement that the query function h be polynomial-time computable. This is the main conceptual limitation of the model. For instance, in a planted-clique problem one could query

$$h(G) = \mathbf{1}(G \text{ contains a clique of size } k),$$

which is not efficiently computable, and turns out to make the whole SQ framework not applicable to the important (unmodified) planted clique task.

Third, the oracle noise is adversarial rather than random. This is not realistic as a data model, but it is precisely what makes the framework strong enough to yield rigorous lower bounds.

So the point of the SQ model is not that it exactly characterizes polynomial time computability. Its point is that it is both expressive enough to capture many relevant algorithms and structured enough to be analyzed rigorously.

8.4 From successful queries to statistical dimension

We now ask: how many $\mathrm{VSTAT}(m)$ queries are needed to distinguish the planted mixture

$$p = \mathop{\mathbb{E}}_{\theta\sim\mu}[p_\theta]$$

from the null q?

Fix a query $h : \Omega \to [0, 1]$. The query is useful for distinguishing p_θ from q only if the difference

$$\left| \mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)] - \mathop{\mathbb{E}}_{X\sim q}[h(X)] \right|$$

is at least of the same order as the oracle noise. More precisely, we say that h is successful for p_θ if

$$\left| \mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)] - \mathop{\mathbb{E}}_{X\sim q}[h(X)] \right| \geq M,$$

where

$$M := \max\left\{ \frac{1}{m}, \frac{1}{\sqrt{m}} \min\left\{ \sqrt{\mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)]\left(1 - \mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)]\right)}, \sqrt{\mathop{\mathbb{E}}_{X\sim q}[h(X)]\left(1 - \mathop{\mathbb{E}}_{X\sim q}[h(X)]\right)} \right\} \right\}$$

Since

$$\mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)] - \mathop{\mathbb{E}}_{X\sim q}[h(X)] = \mathop{\mathbb{E}}_{X\sim q}\left[\left(\frac{p_\theta}{q} - 1\right) h(X)\right] = \left\langle \frac{p_\theta}{q} - 1, h \right\rangle_q,$$

the success condition can be expressed in terms of the correlation between h and the likelihood-ratio fluctuation $p_\theta/q - 1$.

If an $\mathrm{SQ}(m)$ algorithm succeeds with s queries $h_1, \ldots, h_s$, then

$$\mathop{\mathbb{P}}_{\theta\sim\mu}\left(\bigcup_{i=1}^{s} \{h_i \text{ is successful for } p_\theta\} \right) \geq 0.99.$$

Therefore, by the union bound, at least one query must satisfy

$$\mathop{\mathbb{P}}_{\theta\sim\mu}(h \text{ is successful for } p_\theta) \geq \frac{0.99}{s}.$$

This observation is the starting point of the lower-bound argument.

8.4.1 A useful simplification

The success condition can be simplified in a form that is more convenient for the theorem.

Lemma 8.2. *Let $h : \Omega \to [0,1]$ and suppose*

$$\mathop{\mathbb{E}}_{X\sim q}[h(X)] \leq \frac{1}{2}.$$

If h is successful for p_θ, then

$$\left|\mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)] - \mathop{\mathbb{E}}_{X\sim q}[h(X)]\right| \geq \sqrt{\frac{\mathop{\mathbb{E}}_{X\sim q}[h(X)]}{3m}}.$$

Proof. Set

$$a := \mathop{\mathbb{E}}_{X\sim q}[h(X)], \qquad b := \mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)], \qquad d := |a-b|.$$

We are assuming $a \leq 1/2$.

If $a \leq 3/m$, then

$$\sqrt{\frac{a}{3m}} \leq \frac{1}{m},$$

and the success condition immediately implies

$$d \geq \frac{1}{m} \geq \sqrt{\frac{a}{3m}}.$$

We may therefore assume $a > 3/m$. Suppose, toward a contradiction, that

$$d < \sqrt{\frac{a}{3m}}.$$

Since $a > 3/m$, we have

$$\sqrt{\frac{a}{3m}} \leq \frac{a}{3},$$

so $d < a/3$. Hence

$$b \in [a-d, a+d] \subseteq \left[\frac{2a}{3}, \frac{4a}{3}\right].$$

Now the function $x \mapsto x(1-x)$ is concave and symmetric around $1/2$. Since $a \leq 1/2$, its minimum on the interval $[a-d, a+d]$ is attained at $a-d$. Therefore

$$b(1-b) \geq (a-d)(1-a+d).$$

Also, since $a \leq 1/2$, we have $1-a+d \geq 1-a \geq 1/2$, and thus

$$b(1-b) \geq \frac{a-d}{2} > \frac{a}{3}.$$

Similarly,

$$a(1-a) \geq \frac{a}{2} > \frac{a}{3}.$$

Therefore

$$\min\left\{\sqrt{\frac{b(1-b)}{m}}, \sqrt{\frac{a(1-a)}{m}}\right\} > \sqrt{\frac{a}{3m}} > d.$$

Thus the variance part of the VSTAT tolerance is already larger than d. This contradicts the assumption that h is successful for p_θ.

Hence our assumption was false, and

$$d \geq \sqrt{\frac{a}{3m}}.$$

This is exactly the desired bound. □

8.5 Statistical dimension

The correct notion that measures the power of SQ algorithms is statistical dimension, discussed in [FGR+17].

Definition 8.3 ([FGR+17, Definition 2.5.]). *The statistical dimension of the testing problem $p = \mathop{\mathbb{E}}_{\theta\sim\mu}[p_\theta]$ versus q at parameter m is*

$$\mathrm{SDA}(p,q,m) := \max\left\{s \in \mathbb{N} : \forall A \text{ s.t. } \mathop{\mathbb{P}}_{\theta,\theta'\sim\mu}(A) \geq s^{-2}, \ \mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\left|\left\langle \frac{p_\theta}{q}, \frac{p_{\theta'}}{q}\right\rangle_q - 1\right| \middle| A\right] \leq \frac{1}{m}\right\}.$$

This quantity should be compared with the usual (one-sample) χ^2-divergence of the planted mixture and the null measure:

$$\chi^2(p,q) = \mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\left\langle \frac{p_\theta}{q}, \frac{p_{\theta'}}{q}\right\rangle_q - 1\right].$$

The statistical dimension is a stronger, conditional version of this quantity: instead of controlling the average over all pairs (θ,θ'), it controls the average on every event of not-too-small probability.

8.6 The general VSTAT lower bound

We can now state and prove a main theorem of the framework. The exact statement and proof of this version appears in [BBH+21, Appendix A].

Theorem 8.4 ([FGR+17, Theorem 2.7.], [BBH+21, Appendix A]). *For every $m \geq 1$, any $\mathrm{SQ}(m)$ algorithm that solves the detection problem $p = \mathop{\mathbb{E}}_{\theta\sim\mu}[p_\theta]$ versus q must use at least*

$$0.99 \cdot \mathrm{SDA}(p,q,3m)$$

queries to the $\mathrm{VSTAT}(m)$ oracle.

Proof. Assume that an $\mathrm{SQ}(m)$ algorithm succeeds using s queries. As explained earlier, at least one query $h : \Omega \to [0,1]$ must satisfy

$$\mathop{\mathbb{P}}_{\theta\sim\mu}(h \text{ is successful for } p_\theta) \geq \frac{0.99}{s}.$$

Replacing h by $1-h$ if necessary, we may assume

$$\mathop{\mathbb{E}}_{X\sim q}[h(X)] \le \frac{1}{2}.$$

Let

$$S := \{\theta : h \text{ is successful for } p_\theta\}, \qquad \alpha := \mathop{\mathbb{P}}_{\theta\sim\mu}(S).$$

Then $\alpha \ge 0.99/s$.

For $\theta \in S$, the previous lemma gives

$$\left|\left\langle \frac{p_\theta}{q} - 1, h\right\rangle_q\right| = \left|\mathop{\mathbb{E}}_{X\sim p_\theta}[h(X)] - \mathop{\mathbb{E}}_{X\sim q}[h(X)]\right| \ge \sqrt{\frac{\mathop{\mathbb{E}}_{X\sim q}[h(X)]}{3m}}.$$

Define

$$\sigma(\theta) := \operatorname{sign}\left\langle \frac{p_\theta}{q} - 1, h\right\rangle_q,$$

and

$$F_\theta := \left(\frac{p_\theta}{q} - 1\right)\sigma(\theta)\mathbf{1}_S(\theta).$$

Then

$$\alpha\sqrt{\frac{\mathop{\mathbb{E}}_{X\sim q}[h(X)]}{3m}} \le \mathop{\mathbb{E}}_{\theta\sim\mu}\left[\left\langle \frac{p_\theta}{q} - 1, h\right\rangle_q \sigma(\theta)\mathbf{1}_S(\theta)\right] = \left\langle \mathop{\mathbb{E}}_{\theta\sim\mu}[F_\theta], h\right\rangle_q.$$

By Cauchy-Schwarz,

$$\left\langle \mathop{\mathbb{E}}_{\theta\sim\mu}[F_\theta], h\right\rangle_q \le \|h\|_{L^2(q)}\left\|\mathop{\mathbb{E}}_{\theta\sim\mu}[F_\theta]\right\|_{L^2(q)}.$$

Since $0 \le h \le 1$, we have

$$\|h\|^2_{L^2(q)} = \mathop{\mathbb{E}}_{X\sim q}[h(X)^2] \le \mathop{\mathbb{E}}_{X\sim q}[h(X)].$$

Therefore

$$\alpha\sqrt{\frac{1}{3m}} \le \left\|\mathop{\mathbb{E}}_{\theta\sim\mu}[F_\theta]\right\|_{L^2(q)}.$$

Squaring both sides gives

$$\frac{\alpha^2}{3m} \le \mathop{\mathbb{E}}_{X\sim q}\left[\left(\mathop{\mathbb{E}}_{\theta\sim\mu}[F_\theta(X)]\right)^2\right].$$

By Fubini,

$$\mathop{\mathbb{E}}_{X\sim q}\left[\left(\mathop{\mathbb{E}}_{\theta\sim\mu}[F_\theta(X)]\right)^2\right] = \mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\langle F_\theta, F_{\theta'}\rangle_q\right].$$

Now

$$\langle F_\theta, F_{\theta'}\rangle_q = \mathbf{1}_S(\theta)\mathbf{1}_S(\theta')\sigma(\theta)\sigma(\theta')\left\langle \frac{p_\theta}{q} - 1, \frac{p_{\theta'}}{q} - 1\right\rangle_q.$$

Taking absolute values and using $|\sigma(\theta)\sigma(\theta')| = 1$, we obtain

$$\left|\langle F_\theta, F_{\theta'}\rangle_q\right| \le \mathbf{1}_S(\theta)\mathbf{1}_S(\theta')\left|\left\langle \frac{p_\theta}{q} - 1, \frac{p_{\theta'}}{q} - 1\right\rangle_q\right|.$$

Moreover,

$$\left\langle \frac{p_\theta}{q} - 1, \frac{p_{\theta'}}{q} - 1\right\rangle_q = \left\langle \frac{p_\theta}{q}, \frac{p_{\theta'}}{q}\right\rangle_q - 1,$$

because

$$\left\langle \frac{p_\theta}{q}, 1 \right\rangle_q = \mathop{\mathbb{E}}_{X\sim q}\left[\frac{p_\theta(X)}{q(X)}\right] = 1.$$

Therefore

$$\frac{\alpha^2}{3m} \le \mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\mathbf{1}_S(\theta)\mathbf{1}_S(\theta')\left|\left\langle \frac{p_\theta}{q}, \frac{p_{\theta'}}{q} \right\rangle_q - 1\right|\right].$$

Let

$$A := S \times S.$$

Then $\mathop{\mathbb{P}}_{\theta,\theta'\sim\mu}(A) = \alpha^2$, and the previous inequality becomes

$$\frac{\alpha^2}{3m} \le \mathop{\mathbb{P}}_{\theta,\theta'\sim\mu}(A) \mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\left|\left\langle \frac{p_\theta}{q}, \frac{p_{\theta'}}{q} \right\rangle_q - 1\right| \middle| A\right].$$

Since $\mathop{\mathbb{P}}_{\theta,\theta'\sim\mu}(A) = \alpha^2$, we conclude that

$$\mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\left|\left\langle \frac{p_\theta}{q}, \frac{p_{\theta'}}{q} \right\rangle_q - 1\right| \middle| A\right] \ge \frac{1}{3m}.$$

Now assume, for contradiction, that

$$s < 0.99 \cdot \mathrm{SDA}(p, q, 3m).$$

Then

$$\alpha \ge \frac{0.99}{s} > \frac{1}{\mathrm{SDA}(p, q, 3m)}.$$

Hence

$$\mathop{\mathbb{P}}_{\theta,\theta'\sim\mu}(A) = \alpha^2 > \frac{1}{\mathrm{SDA}(p, q, 3m)^2}.$$

But by the definition of statistical dimension, every such event A must satisfy

$$\mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\left|\left\langle \frac{p_\theta}{q}, \frac{p_{\theta'}}{q} \right\rangle_q - 1\right| \middle| A\right] \le \frac{1}{3m},$$

contradicting the lower bound just obtained.

Therefore any successful $\mathrm{SQ}(m)$ algorithm must use at least

$$0.99 \cdot \mathrm{SDA}(p, q, 3m)$$

queries. □

Chapter 9

SQ Bounds for Sparse PCA

9.1 Introduction

In the previous chapter, we introduced the statistical-query framework and the notion of statistical dimension, which gives a general method for proving lower bounds against $\mathrm{VSTAT}(m)$ algorithms. The purpose of the present chapter is to apply that framework to one of the central models of these notes: sparse PCA.

This chapter has three goals. First, we recall the sparse PCA detection problem in the multi-sample and cloned single-sample formulations, and review its information-theoretic threshold. Second, we compare this threshold with the performance of the currently best algorithms for the task, namely PCA and diagonal thresholding. This reveals the computational-statistical gap that motivates the chapter. Third, we prove an SQ lower bound showing that, in the intermediate regime, any $\mathrm{SQ}(m)$ algorithm requires super-polynomially many queries. We also discuss simple SQ upper bounds in the easy regimes, so that the full phase diagram becomes tractable from the SQ point of view. To the best of our knowledge, the material presented in this chapter is new and has not appeared in the literature.

9.2 Model setup for sparse PCA

We work in the regime $n^{\Omega(1)} = k = o(n)$. The detection problem is to test the alternative hypothesis $\mathbb{P}_1$ against the null hypothesis $\mathbb{P}_2$ from m i.i.d. samples, where sample i is given by

$$\begin{cases} \mathbb{P}_1 : \ Y_i = xx^\top + W_i, \\ \mathbb{P}_2 : \ Y_i = W_i. \end{cases}$$

Here the signal x is sampled from a prior μ with independent coordinates

$$x_j \overset{\text{i.i.d.}}{\sim} \begin{cases} \frac{1}{\sqrt{k}}, & \text{with probability } \frac{k}{2n}, \\ -\frac{1}{\sqrt{k}}, & \text{with probability } \frac{k}{2n}, \\ 0, & \text{with probability } 1 - \frac{k}{n}, \end{cases} \qquad j \in [n],$$

and the noise matrices satisfy

$$(W_i)_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

The question is: for what values of m can we achieve strong detection?

As in earlier chapters, it is useful to keep in mind that the vector x is approximately k-sparse and has Euclidean norm of order one. Indeed,

$$\|x\|_0 \sim \mathrm{Bin}\left(n, \frac{k}{n}\right), \qquad \|x\|_2^2 = \frac{\|x\|_0}{k}.$$

So, with high probability,

$$\|x\|_0 \asymp k, \qquad \|x\|_2^2 \asymp 1.$$

9.3 Information-theoretic bound

9.3.1 Cloning reduction

As in the sparse PCA discussion from earlier chapters (see Section 4.4), one can pass from the m-sample model to an equivalent single-sample model by the cloning trick. Define

$$Y := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Y_i.$$

Then under $\mathbb{P}_1$,

$$Y = \sqrt{m}\, xx^\top + W,$$

while under $\mathbb{P}_2$,

$$Y = W,$$

where again W has i.i.d. $\mathcal{N}(0,1)$ entries.

So the original multi-sample problem is statistically equivalent to the single-sample testing problem

$$\begin{cases} \mathbb{P}_1 : \; Y = \sqrt{m}\, xx^\top + W, \\ \mathbb{P}_2 : \; Y = W, \end{cases} \qquad x \sim \mu, \qquad W_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

This is the form in which the formulas for Gaussian additive models from the previous low-degree chapters apply more well-suited.

9.3.2 The χ^2 formula

By the cloning trick, the original m-sample problem is equivalent to the single-sample testing problem

$$\begin{cases} \mathbb{P}_1 : \; Y = \sqrt{m}\, xx^\top + W, \\ \mathbb{P}_2 : \; Y = W, \end{cases} \qquad x \sim \mu, \qquad W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Since this is a Gaussian additive model, the formula from the previous low-degree chapter (see Lemma 6.3) gives

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \left[\exp\left(m \langle x, x' \rangle^2\right)\right] - 1.$$

The next proposition shows that the χ^2-distance is asymptotically trivial below the information-theoretic threshold.

Proposition 9.1. *Assume*

$$k = o(n), \qquad k = n^{\Omega(1)}.$$

Let

$$\mathcal{E} := \{\|x\|_0 \le 2k\},$$

and let $\widetilde{\mu}$ be the law of x conditioned on $\mathcal{E}$ in the sparse PCA model defined in Section 9.2. If

$$m = o\left(k \log \frac{n}{k}\right),$$

then

$$\chi^2(\widetilde{P}_1, P_2) \to 0.$$

Consequently, since $\mathbb{P}(\mathcal{E}^c) \to 0$, the original Bernoulli planted model is contiguous to the null in total variation, and weak detection is impossible in the same regime.

The proof can be found in Section A.7.

Thus the single-sample sparse PCA model is information-theoretically undetectable when

$$m \ll k \log \frac{n}{k}.$$

9.3.3 Thresholding above the information-theoretic threshold

Consider

$$\mathrm{OPT} := \max_{\substack{v \in \{0, \pm 1/\sqrt{k}\}^n, \\ \|v\|_0 = k}} v^\top Y v.$$

Using the single-sample model, this can be written as

$$\mathrm{OPT} = \max_{\substack{v \in \{0, \pm 1/\sqrt{k}\}^n, \\ \|v\|_0 = k}} \left(\sqrt{m}\,\langle v, x\rangle^2 + v^\top W v\right).$$

Under $\mathbb{P}_1$, we want a candidate v correlated with x. Since $\|x\|_0 \sim \mathrm{Bin}(n, k/n)$, a standard binomial tail bound implies

$$\mathop{\mathbb{P}}_{x \sim \mu}\left(\|x\|_0 \ge \frac{k}{2}\right) \to 1.$$

On the event $\|x\|_0 \ge k/2$, we can construct an admissible vector v with nontrivial correlation with x. Let

$$r := \|x\|_0.$$

Choose a set

$$T \subseteq \mathrm{supp}(x)$$

of size $\min\{r, k\}$. Since $r \ge k/2$, we have $|T| \ge k/2$. Define

$$v_i = \frac{\mathrm{sign}(x_i)}{\sqrt{k}} \qquad \text{for } i \in T.$$

If $|T| < k$, choose the remaining $k - |T|$ coordinates of the support of v outside $\mathrm{supp}(x)$, with arbitrary signs. This gives an admissible vector

$$v \in \left\{0, \pm\frac{1}{\sqrt{k}}\right\}^n, \qquad \|v\|_0 = k.$$

Moreover, the coordinates added outside $\mathrm{supp}(x)$ do not contribute to the inner product, while for every $i \in T$,

$$v_i x_i = \frac{\mathrm{sign}(x_i)}{\sqrt{k}} x_i = \frac{1}{k}.$$

Therefore

$$\langle v, x \rangle = \sum_{i \in T} v_i x_i = \frac{|T|}{k} \geq \frac{1}{2}.$$

For this choice of v,

$$v^\top Y v = \sqrt{m} \langle v, x \rangle^2 + v^\top W v \geq \frac{\sqrt{m}}{4} + v^\top W v$$

so

$$\mathrm{OPT} \geq \frac{\sqrt{m}}{4} + v^\top W v.$$

Since $v^\top W v \sim \mathcal{N}(0,1)$ for each fixed v, we conclude that under $\mathbb{P}_1$,

$$\mathrm{OPT} \geq c\sqrt{m}$$

with high probability for some absolute constant $c > 0$.

Under $\mathbb{P}_2$, each $v^\top W v$ is again $\mathcal{N}(0,1)$. The number of admissible vectors v is at most

$$\binom{n}{k} 2^k.$$

Therefore, by the Gaussian tail,

$$\mathrm{OPT} \leq C\sqrt{\log\left(\binom{n}{k} 2^k\right)} \leq C'\sqrt{k \log \frac{n}{k}}$$

with high probability.

Comparing the two bounds shows that thresholding succeeds when

$$\sqrt{m} \gg \sqrt{k \log \frac{n}{k}},$$

that is,

$$m \gg k \log \frac{n}{k}.$$

So the information-theoretic picture is the one shown in Figure 9.1.

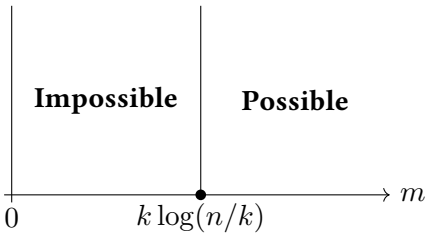

Figure 9.1: Sparse PCA phase diagram from the SQ point of view

9.4 Time-efficient algorithms

We now ask what happens if one imposes a computational constraint. The previous thresholding argument is not polynomial-time computable, because it maximizes over all k-sparse signed vectors. The question is what can be achieved by efficient algorithms.

9.4.1 PCA works for $m \gg n$

Consider again the cloned single-sample model

$$Y = \sqrt{m}\, xx^\top + W.$$

The largest eigenvalue satisfies

$$\lambda_{\max}(Y) = \max_{v \in \mathbb{S}^{n-1}} v^\top Y v = \max_{v \in \mathbb{S}^{n-1}} \left(\sqrt{m}\, \langle v, x \rangle^2 + v^\top W v \right).$$

Under $\mathbb{P}_1$, taking $v = x/\|x\|_2$ gives

$$\lambda_{\max}(Y) \geq \sqrt{m}\|x\|_2^2 + \frac{x^\top W x}{\|x\|_2^2}.$$

Since $\|x\|_2^2 \asymp 1$ with high probability and the second term is of lower order, this yields

$$\lambda_{\max}(Y) \geq c\sqrt{m}$$

with high probability.

Under $\mathbb{P}_2$, one may use the BBP/random-matrix bound recalled earlier in the notes to conclude that

$$\lambda_{\max}(Y) \sim 2\sqrt{n}$$

with high probability.

Therefore PCA succeeds when

$$\sqrt{m} \gg \sqrt{n},$$

that is,

$$m \gg n.$$

9.4.2 Diagonal thresholding works for $m \gg k^2 \log n$ when $k \ll \sqrt{n}$

The diagonal entries of Y are

$$Y_{ii} = \sqrt{m}\, x_i^2 + W_{ii}.$$

Since x_i^2 is either 0 or $1/k$, the planted model creates a small positive shift on the diagonal coordinates belonging to the support of x.

Under $\mathbb{P}_1$, for every i in the support of x,

$$Y_{ii} = \frac{\sqrt{m}}{k} + W_{ii},$$

so

$$\max_i Y_{ii} \geq \frac{\sqrt{m}}{k} - \mathcal{O}(1)$$

with high probability.

Under $\mathbb{P}_2$, the diagonal entries are i.i.d. standard Gaussians, so, as shown previously on the notes,

$$\max_i Y_{ii} \le C\sqrt{\log n}$$

with high probability.

Thus diagonal thresholding succeeds when

$$\frac{\sqrt{m}}{k} \gg \sqrt{\log n},$$

that is,

$$m \gg k^2 \log n.$$

This is particularly relevant when $k \ll \sqrt{n}$, because then $k^2 \log n \ll n$ up to logarithmic factors, so diagonal thresholding improves on PCA.

9.4.3 The resulting phase diagram

Up to logarithmic factors, Figure 9.2 shows the sparse PCA phase diagram.

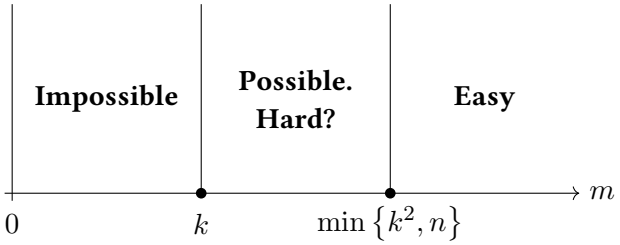

Figure 9.2: Sparse PCA phase diagram from the SQ point of view

The main purpose of the rest of the chapter is to justify the middle region from the SQ point of view.

9.5 SQ lower bound for sparse PCA

We now prepare the ingredients needed for the statistical-dimension lower bound.

9.5.1 One-sample laws and likelihood-ratio correlations

The general SQ theorem from the previous chapter is formulated in terms of one-sample laws. Therefore, for a fixed spike x, let $\mathbb{P}_x$ denote the law of a single sample

$$Y = xx^\top + W,$$

and let $\mathbb{Q}$ denote the null law

$$Y = W,$$

where the entries of W are i.i.d. $\mathcal{N}(0,1)$. The VSTAT parameter m will enter through the tolerance $1/m$ in the definition of statistical dimension; it should not be inserted into the one-sample likelihood-ratio correlation.

For fixed x, x', we compute

$$\left\langle \frac{\mathbb{P}_x}{\mathbb{Q}}, \frac{\mathbb{P}_{x'}}{\mathbb{Q}} \right\rangle_{\mathbb{Q}}.$$

Since this is a Gaussian additive model, the likelihood ratio is

$$\frac{\mathrm{d}\mathbb{P}_x}{\mathrm{d}\mathbb{Q}}(Y) = \exp\left(\left\langle Y, xx^\top \right\rangle - \frac{1}{2}\left\|xx^\top\right\|_F^2\right).$$

Thus

$$\begin{aligned}\left\langle \frac{\mathbb{P}_x}{\mathbb{Q}}, \frac{\mathbb{P}_{x'}}{\mathbb{Q}} \right\rangle_{\mathbb{Q}} &= \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}\left[\exp\left(\left\langle Y, xx^\top \right\rangle - \frac{1}{2}\left\|xx^\top\right\|_F^2\right)\exp\left(\left\langle Y, x'x'^\top \right\rangle - \frac{1}{2}\left\|x'x'^\top\right\|_F^2\right)\right] \\ &= \exp\left(\left\langle xx^\top, x'x'^\top \right\rangle\right),\end{aligned}$$

where we used the moment-generating function of a standard Gaussian vector. Finally,

$$\left\langle xx^\top, x'x'^\top \right\rangle = \left\langle x, x' \right\rangle^2.$$

Therefore

$$\left\langle \frac{\mathbb{P}_x}{\mathbb{Q}}, \frac{\mathbb{P}_{x'}}{\mathbb{Q}} \right\rangle_{\mathbb{Q}} = e^{\langle x,x'\rangle^2}.$$

Equivalently,

$$\left\langle \frac{\mathbb{P}_x}{\mathbb{Q}}, \frac{\mathbb{P}_{x'}}{\mathbb{Q}} \right\rangle_{\mathbb{Q}} - 1 = e^{\langle x,x'\rangle^2} - 1.$$

This is the quantity that must be controlled conditionally in order to lower-bound the statistical dimension. More precisely, to prove that

$$\mathrm{SDA}(\mathbb{P}, \mathbb{Q}, m) \geq s,$$

it is enough to show that for every event

$$A \subseteq \mathrm{supp}(\mu) \times \mathrm{supp}(\mu)$$

with

$$\mathop{\mathbb{P}}_{x,x'\sim\mu}(A) \geq s^{-2},$$

one has

$$\mathop{\mathbb{E}}_{x,x'\sim\mu}\left[e^{\langle x,x'\rangle^2} - 1 \;\middle|\; A\right] \leq \frac{1}{m}.$$

Thus the SQ lower bound reduces to understanding the upper tail of the overlap $\langle x, x'\rangle$.

It is useful to keep this separate from the χ^2 computation in the information-theoretic part of the chapter. There, after cloning m samples into the single matrix

$$Y = \sqrt{m}\, xx^\top + W,$$

one obtains

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) + 1 = \mathop{\mathbb{E}}_{x,x'\sim\mu}\left[\exp\left(m\left\langle x, x'\right\rangle^2\right)\right].$$

By contrast, in the SQ lower bound the one-sample correlation is $e^{\langle x,x'\rangle^2}$, and the sample size parameter m appears through the VSTAT tolerance $1/m$.

9.5.2 A rearrangement lemma

The next simple lemma allows us to reduce the supremum over all conditioning events to upper-tail events of the overlap. It is a standard rearrangement principle: among all events of a fixed probability, the conditional expectation of an increasing function is maximized by keeping the largest values of the underlying random variable.

Lemma 9.2 (Rearrangement lemma)**.** *Let Z be a nonnegative random variable, let $g : [0,\infty) \to [0,\infty)$ be increasing, and let $\alpha \in (0,1)$. Define*

$$t_\alpha := \inf\{t \geq 0 : \mathbb{P}(Z \geq t) \leq \alpha\}.$$

Then for every event A with $\mathbb{P}(A) \geq \alpha$,

$$\mathbb{E}[g(Z) \mid A] \leq \mathbb{E}[g(Z) \mid Z \geq t_\alpha].$$

Proof. It is enough to prove the statement for events A with $\mathbb{P}(A) = \alpha$, because if $\mathbb{P}(A) > \alpha$, we may choose a subset $A' \subseteq A$ with $\mathbb{P}(A') = \alpha$ and

$$\mathbb{E}[g(Z) \mid A] \leq \sup_{\mathbb{P}(B)=\alpha} \mathbb{E}[g(Z) \mid B].$$

So fix A with $\mathbb{P}(A) = \alpha$. We must show

$$\mathbb{E}[g(Z)\mathbf{1}_A] \leq \mathbb{E}[g(Z)\mathbf{1}_{\{Z \geq t_\alpha\}}].$$

Since g is increasing, the event that maximizes the integral of $g(Z)$ among all events of probability α is obtained by keeping the largest values of Z. Formally, if one decomposes A into

$$A \cap \{Z < t_\alpha\}, \qquad A \cap \{Z = t_\alpha\}, \qquad A \cap \{Z > t_\alpha\},$$

then replacing the part of A lying below t_α by an equal-probability subset of $\{Z > t_\alpha\}$ can only increase the integral of $g(Z)$, because g is increasing. Repeating this replacement until the entire mass is concentrated on the upper tail gives the desired inequality. □

We will apply this lemma with

$$Z = \langle x, x' \rangle^2, \qquad g(z) = e^z - 1.$$

Since g is increasing, the worst conditioning event for the statistical-dimension bound is an upper-tail event of $\langle x, x' \rangle^2$. Thus it suffices to control the overlap tail.

9.5.3 Proof of the SQ lower bound

We now prove the main theorem of the chapter.

Theorem 9.3 (SQ lower bound)**.** *In the sparse PCA model, for every $\varepsilon > 0$, if*

$$k \ll m \ll n^{-\varepsilon} \min\left\{k^2, n\right\}.$$

Then any $\mathrm{SQ}(m)$ algorithm requires $\exp\left(n^{\Omega(1)}\right)$-many queries.

Proof. By the general statistical-dimension theorem from the previous chapter, it is enough to show that for some $s \gg \mathrm{poly}(n)$,

$$\sup_{A:\ \mathop{\mathbb{P}}_{x,x'\sim\mu}(A)\geq s^{-2}} \mathop{\mathbb{E}}_{x,x'\sim\mu}\left[\left|\left\langle \frac{\mathbb{P}_x}{\mathbb{Q}}, \frac{\mathbb{P}_{x'}}{\mathbb{Q}}\right\rangle - 1\right| \,\middle|\, A\right] \leq \frac{1}{m}.$$

Since we are in a Gaussian additive model, the Gaussian-additive formula gives

$$\left\langle \frac{\mathbb{P}_x}{\mathbb{Q}}, \frac{\mathbb{P}_{x'}}{\mathbb{Q}}\right\rangle = e^{\langle x,x'\rangle^2}.$$

So it suffices to prove that for some $s \gg \mathrm{poly}(n)$,

$$\sup_{A:\ \mathop{\mathbb{P}}_{x,x'\sim\mu}(A)\geq s^{-2}} \mathop{\mathbb{E}}_{x,x'\sim\mu}\left[e^{\langle x,x'\rangle^2} - 1 \,\middle|\, A\right] \leq \frac{1}{m}.$$

We now estimate the overlap $\langle x, x'\rangle$. Write

$$\langle x, x'\rangle = \sum_{i=1}^{n} Z_i, \qquad Z_i := x_i x_i'.$$

The variables $Z_1, \ldots, Z_n$ are independent, centered, and satisfy

$$|Z_i| \leq \frac{1}{k}, \qquad \mathbb{E}\left[Z_i^2\right] = \frac{1}{n^2}.$$

Hence

$$\sum_{i=1}^{n} \mathbb{E}\left[Z_i^2\right] = \frac{1}{n}.$$

Applying Bernstein's inequality, for every $t > 0$ we obtain

$$\mathop{\mathbb{P}}_{x,x'\sim\mu}\left(|\langle x, x'\rangle| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2/n + \frac{2t}{3k}}\right).$$

In particular,

$$\mathop{\mathbb{P}}_{x,x'\sim\mu}\left(|\langle x, x'\rangle| \geq t\right) \leq \begin{cases} 2\exp\left(-\frac{1}{4}nt^2\right), & \text{if } t \leq \frac{3k}{n}, \\ 2\exp\left(-\frac{3}{4}tk\right), & \text{if } t > \frac{3k}{n}. \end{cases}$$

Now choose a small $\varepsilon > 0$ such that $k \gg n^{\varepsilon}$. This is possible, since we are in the polynomial-sparsity regime $k = n^{\alpha}$ with $\alpha \in (0, 1/2)$. Set

$$t := n^{\varepsilon} \max\left\{\frac{1}{\sqrt{n}}, \frac{1}{k}\right\}.$$

Then $t = o(1)$, and from the Bernstein bound above we obtain

$$\mathop{\mathbb{P}}_{x,x'\sim\mu}\left(\langle x, x'\rangle^2 \geq t^2\right) \leq 2e^{-n^{\varepsilon}}.$$

Also, replacing t by $3t$,

$$\mathop{\mathbb{P}}_{x,x'\sim\mu}\left(\langle x, x'\rangle^2 \geq 9t^2\right) \leq 2e^{-3n^{\varepsilon}}.$$

We now choose

$$s := \sqrt{2}\, e^{n^{\varepsilon}/2},$$

so that

$$s^{-2} = \frac{1}{2} e^{-n^{\varepsilon}}.$$

Let

$$A_{t_*} := \left\{ \langle x, x' \rangle^2 \geq t_*^2 \right\},$$

where t_* is the largest t such that

$$\mathop{\mathbb{P}}_{x,x' \sim \mu} (A_{t_*}) \geq s^{-2}.$$

By monotonicity of the map $z \mapsto e^{z^2} - 1$ in $|z|$, and by Lemma 9.2, the supremum over all events A with probability at least s^{-2} is attained by such an upper-tail event. Since

$$\mathop{\mathbb{P}}_{x,x' \sim \mu} \left(\langle x, x' \rangle^2 \geq t^2 \right) \leq 2e^{-n^{\varepsilon}},$$

we have

$$t_* \leq t.$$

We now estimate

$$\mathop{\mathbb{E}}_{x,x' \sim \mu} \left[e^{\langle x,x' \rangle^2} - 1 \,\middle|\, A_{t_*} \right].$$

Split according to whether $\langle x, x' \rangle^2 \leq 9t^2$ or not:

$$\mathbb{E}\left[e^{\langle x,x' \rangle^2} - 1 \,\middle|\, A_{t_*} \right] \leq \mathbb{E}\left[\left(e^{\langle x,x' \rangle^2} - 1 \right) \mathbf{1}\left(t_*^2 \leq \langle x, x' \rangle^2 \leq 9t^2 \right) \,\middle|\, A_{t_*} \right]$$
$$+ \mathbb{E}\left[\left(e^{\langle x,x' \rangle^2} - 1 \right) \mathbf{1}\left(\langle x, x' \rangle^2 > 9t^2 \right) \,\middle|\, A_{t_*} \right].$$

For the first term, since $\langle x, x' \rangle^2 \leq 9t^2$,

$$e^{\langle x,x' \rangle^2} - 1 \leq e^{9t^2} - 1.$$

For the second term, since $\langle x, x' \rangle^2 \leq 1$ always, we have

$$e^{\langle x,x' \rangle^2} - 1 \leq e - 1.$$

Therefore

$$\mathop{\mathbb{E}}_{x,x' \sim \mu} \left[e^{\langle x,x' \rangle^2} - 1 \,\middle|\, A_{t_*} \right] \leq e^{9t^2} - 1 + (e-1) \frac{\mathop{\mathbb{P}}_{x,x' \sim \mu} \left(\langle x, x' \rangle^2 > 9t^2 \right)}{\mathop{\mathbb{P}}_{x,x' \sim \mu} (A_{t_*})}.$$

Using the bounds above,

$$\mathop{\mathbb{P}}_{x,x' \sim \mu} \left(\langle x, x' \rangle^2 > 9t^2 \right) \leq 2e^{-3n^{\varepsilon}}, \qquad \mathop{\mathbb{P}}_{x,x' \sim \mu} (A_{t_*}) \geq s^{-2} = \frac{1}{2} e^{-n^{\varepsilon}}.$$

Hence

$$\frac{\mathop{\mathbb{P}}_{x,x' \sim \mu} \left(\langle x, x' \rangle^2 > 9t^2 \right)}{\mathop{\mathbb{P}}_{x,x' \sim \mu} (A_{t_*})} \leq 4e^{-2n^{\varepsilon}}.$$

Thus

$$\mathop{\mathbb{E}}_{x,x' \sim \mu} \left[e^{\langle x,x' \rangle^2} - 1 \,\middle|\, A_{t_*} \right] \leq e^{9t^2} - 1 + 4(e-1)e^{-2n^{\varepsilon}}.$$

Since $t = o(1)$,

$$e^{9t^2} - 1 = \mathcal{O}\left(t^2\right).$$

By the choice of t,

$$t^2 = n^{2\varepsilon} \max\left\{\frac{1}{n}, \frac{1}{k^2}\right\}.$$

Therefore

$$\mathop{\mathbb{E}}_{x,x' \sim \mu}\left[e^{\langle x,x'\rangle^2} - 1 \,\middle|\, A_{t_*}\right] \le C n^{2\varepsilon} \max\left\{\frac{1}{n}, \frac{1}{k^2}\right\}$$

for some absolute constant C and all large enough n.

Consequently, if

$$m \le c\, n^{-2\varepsilon} \min\left\{k^2, n\right\}$$

for a sufficiently small constant $c > 0$, then

$$\mathop{\mathbb{E}}_{x,x' \sim \mu}\left[e^{\langle x,x'\rangle^2} - 1 \,\middle|\, A_{t_*}\right] \le \frac{1}{m}.$$

This proves that

$$\mathrm{SDA}(p, q, m) \ge s = \sqrt{2}\, e^{n^{\varepsilon}/2} \gg \mathrm{poly}(n),$$

and hence any $\mathrm{SQ}(m)$ algorithm requires at least polynomially many queries. □

9.6 SQ upper bounds for sparse PCA

We now turn to the easy regimes from the SQ point of view. The goal is not to claim that every polynomial-time algorithm can be represented as an SQ algorithm. Rather, we show that the simple statistics responsible for the usual algorithmic upper bounds in sparse PCA can also be implemented using VSTAT queries. In particular, diagonal thresholding in the sparse regime and trace thresholding in the denser regime both have natural SQ implementations with the expected sample-size scaling. Thus the SQ framework is not only useful for proving lower bounds in the conjecturally hard regime; it also contains natural algorithms on the easy side of the phase diagram.

9.6.1 Diagonal-thresholding $\mathrm{VSTAT}(m)$ works for $m \gg k^2$ when $k \ll \sqrt{n}$

The construction is inspired by ordinary diagonal thresholding. Fix $\ell \in [n]$ and consider the query

$$h_\ell(Y) := \mathbf{1}\left(Y_{\ell\ell} \ge \frac{1}{2k}\right).$$

If the hypothesis is $D \in \{\mathbb{P}_x, \mathbb{Q}\}$, then the $\mathrm{VSTAT}(m)$ oracle returns

$$\mathop{\mathbb{E}}_{Y \sim D}[h_\ell(Y)] + g,$$

where

$$|g| \le \max\left\{\frac{1}{m}, \sqrt{\frac{\mathop{\mathbb{E}}_{Y \sim D}[h_\ell(Y)](1 - \mathop{\mathbb{E}}_{Y \sim D}[h_\ell(Y)])}{m}}\right\}.$$

Under $\mathbb{P}_x$, for every ℓ such that $x_\ell^2 = 1/k$,

$$\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_x}[h_\ell(Y)] = \mathop{\mathbb{P}}_{\varepsilon \sim \mathcal{N}(0,1)}\left(\frac{1}{k} + \varepsilon \ge \frac{1}{2k}\right) = \mathop{\mathbb{P}}_{\varepsilon \sim \mathcal{N}(0,1)}\left(\varepsilon \ge -\frac{1}{2k}\right).$$

Using the Taylor expansion of the Gaussian cdf at zero,

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_x}[h_\ell(Y)] = \frac{1}{2} + \Theta\left(\frac{1}{k}\right).$$

Under $\mathbb{Q}$,

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}[h_\ell(Y)] = \mathop{\mathbb{P}}_{\varepsilon\sim\mathcal{N}(0,1)}\left(\varepsilon \geq \frac{1}{2k}\right) = \frac{1}{2} - \Theta\left(\frac{1}{k}\right).$$

In both cases the variance term in the oracle error is of order $1/m$, so the oracle outputs differ by an amount of order $1/k$, up to an error of order $m^{-1/2}$. Therefore the signal dominates the oracle noise whenever

$$\frac{1}{k} \gg \frac{1}{\sqrt{m}},$$

that is,

$$m \gg k^2.$$

Thus by querying all diagonal coordinates and checking whether one of them shows the positive shift, one obtains a strong detection algorithm in the regime $m \gg k^2$.

9.6.2 Trace-thresholding $\mathrm{VSTAT}(m)$ works for $m \gg n$ when $k \gg \sqrt{n}$

When k is larger, one can use the trace instead. Consider the query

$$h(Y) := \mathbf{1}(\mathrm{Tr}(Y) \geq \xi)$$

for some fixed threshold $\xi > 0$.

Under $\mathbb{Q}$,

$$\mathrm{Tr}(Y) = \sum_{\ell=1}^{n} W_{\ell\ell} \sim \mathcal{N}(0, n),$$

so

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}[h(Y)] = \mathop{\mathbb{P}}_{z\sim\mathcal{N}(0,1)}\left(\sqrt{n}\, z \geq \xi\right) = \frac{1}{2} - \frac{\xi}{\sqrt{2\pi n}} + o\left(n^{-1/2}\right).$$

Under $\mathbb{P}_x$,

$$\mathrm{Tr}(Y) = \|x\|_2^2 + \sum_{\ell=1}^{n} W_{\ell\ell}.$$

Now

$$\mathop{\mathbb{E}}_{x\sim\mu}\left[\|x\|_2^2\right] = 1, \qquad \mathop{\mathrm{Var}}_{x\sim\mu}\left(\|x\|_2^2\right) = \frac{1}{k} - \frac{1}{n}.$$

A Bernstein bound implies

$$\mathop{\mathbb{P}}_{x\sim\mu}\left(\|x\|_2^2 \leq \frac{1}{2}\right) \leq e^{-ck}$$

for some constant $c > 0$. Hence, with high probability under $\mathbb{P}_x$,

$$\|x\|_2^2 \geq \frac{1}{2}.$$

Choosing $\xi = \frac{1}{4}$, we then get

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{P}_x}[h(Y)] \geq \mathop{\mathbb{P}}_{z\sim\mathcal{N}(0,1)}\left(\frac{1}{2} + \sqrt{n}\, z \geq \frac{1}{4}\right) = \frac{1}{2} + \frac{1}{8\sqrt{2\pi n}} + o\left(n^{-1/2}\right)$$

with probability at least $1 - e^{-ck}$ over the draw of x.

Thus the planted and null expectations differ by an amount of order $n^{-1/2}$, while the $\mathrm{VSTAT}(m)$ error is of order $m^{-1/2}$. Therefore trace thresholding works when

$$\frac{1}{\sqrt{n}} \gg \frac{1}{\sqrt{m}},$$

that is,

$$m \gg n.$$

9.6.3 A heuristic analysis of SQ gradient descent under adversarial noise: underperformance?

We finally discuss a gradient-descent-based heuristic. This part is more informal than the previous two SQ upper bounds. The point is not to give a complete algorithmic proof, but to explain what the SQ model predicts for gradient descent in this problem (see [DH21]).

Recall that, in the cloned single-sample model, we observe

$$\mathbb{P}_1 : \quad Y = xx^\top + W, \qquad \mathbb{P}_2 : \quad Y = W,$$

where x is sparse and approximately normalized. Consider the optimization problem

$$\max_{v \in \mathbb{S}^{n-1}} \frac{1}{2} v^\top Y v.$$

The gradient of the objective with respect to v is

$$\nabla_v \left(\frac{1}{2} v^\top Y v \right) = Yv.$$

Thus a population gradient step would use the quantity $\mathop{\mathbb{E}}_{Y \sim D}[Yv_t]$, where D is either the planted distribution $\mathbb{P}_x$ or the null distribution $\mathbb{P}_2$.

In the SQ model we do not observe this expectation exactly. Instead, the $\mathrm{VSTAT}(m)$ oracle returns an approximation. Therefore, under the planted model, the update takes the form

$$v_{t+1} = v_t + \eta \left(\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_x} [Yv_t] + g_t \right),$$

where $g_t \in \mathbb{R}^n$ is the adversarial error introduced by the oracle and $\eta > 0$ is the step size.

Since under $\mathbb{P}_x$ we have $Y = xx^\top + W$ and $\mathbb{E}[W] = 0$, we get

$$\mathop{\mathbb{E}}_{Y \sim \mathbb{P}_x} [Yv_t] = xx^\top v_t = \langle v_t, x \rangle \, x.$$

Hence the planted update is

$$v_{t+1} = v_t + \eta \left(\langle v_t, x \rangle \, x + g_t \right).$$

Under the null model $\mathbb{P}_2$, the signal term is absent, and the corresponding update is simply

$$v_{t+1} = v_t + \eta g_t.$$

To understand whether gradient descent can detect the planted model, we look at the evolution of the correlation with the spike, namely $\langle v_t, x \rangle$. Taking inner product with x, we obtain under $\mathbb{P}_1$

$$\langle v_{t+1}, x \rangle = \langle v_t, x \rangle + \eta \, \langle v_t, x \rangle \, \|x\|_2^2 + \eta g_t^\top x.$$

Since $\|x\|_2^2 \approx 1$, this is approximately

$$\langle v_{t+1}, x\rangle \approx (1+\eta)\langle v_t, x\rangle + \eta g_t^\top x.$$

Under $\mathbb{P}_2$, on the other hand,

$$\langle v_{t+1}, x\rangle = \langle v_t, x\rangle + \eta g_t^\top x.$$

Thus the difference between the planted and null dynamics is the multiplicative signal term

$$\eta\langle v_t, x\rangle.$$

For this term to be detectable, it must dominate the adversarial SQ error $g_t^\top x$.

Let us estimate the size of this error. If $\|v_t\|_2 = 1$, then for Gaussian noise W,

$$Wv_t \sim \mathcal{N}(0, I_n).$$

Therefore each coordinate of the vector-valued query has variance of order one. The $\mathrm{VSTAT}(m)$ oracle introduces coordinatewise error of order

$$|(g_t)_\ell| \approx m^{-1/2}.$$

Since x has about k nonzero coordinates, each of size $1/\sqrt{k}$, the worst possible adversarial alignment gives

$$\left|g_t^\top x\right| \le \sum_{\ell\in\mathrm{supp}(x)} |(g_t)_\ell|\ |x_\ell| \approx km^{-1/2}\frac{1}{\sqrt{k}} = \sqrt{\frac{k}{m}}.$$

Now consider the first step from a random initialization. If v_0 is independent of x and approximately uniform on the sphere, then

$$|\langle v_0, x\rangle| = \mathcal{O}\left(n^{-1/2}\right).$$

Indeed, one may think of the initialization as essentially random in n dimensions, so its projection onto the fixed direction x is of order $1/\sqrt{n}$.

At the first step, the signal contribution to the correlation is therefore of order

$$\eta|\langle v_0, x\rangle| \approx \eta n^{-1/2},$$

whereas the adversarial error contribution is at most

$$\eta\left|g_0^\top x\right| \approx \eta\sqrt{\frac{k}{m}}.$$

For the signal to dominate the adversarial error, we need

$$n^{-1/2} \gg \sqrt{\frac{k}{m}}.$$

Equivalently,

$$m \gg nk.$$

This explains the heuristic statement that gradient descent works only once

$$m \gg nk$$

in the regime $k \gg \sqrt{n}$. Notice that this is weaker than trace thresholding, which already works for $m \gg n$. The value of this discussion is therefore not that the original gradient descent matches (or not) the best known algorithm here, but that it illustrates how first-order dynamics based on SQ are affected by adversarial oracle noise, and can be suboptimal.

In particular, the SQ viewpoint predicts that, at a random initialization, the true gradient signal points in the planted direction only through the small initial correlation $\langle v_0, x\rangle$, while the oracle error can align adversarially with the sparse support of x. Balancing these two effects leads to the condition $m \gg nk$.

9.7 Conclusion and transition

This chapter completes the SQ picture for sparse PCA. On the one hand, the statistical-dimension argument shows that when

$$k \ll m \ll \min\left\{k^2, n\right\},$$

any $\mathrm{SQ}(m)$ algorithm requires super-polynomially many queries. On the other hand, simple SQ algorithms based on diagonal thresholding or trace thresholding succeed once

$$m \gg \min\left\{k^2, n\right\},$$

up to logarithmic factors and the regime assumptions described above.

In the next chapter, we will see how the same SQ methodology applies to another central problem of these notes: planted clique.

Chapter 10

SQ Lower Bounds for Planted Clique

10.1 Introduction

In the previous chapter, we applied the statistical-query framework to sparse PCA and saw that SQ lower bounds recover the same qualitative hard regime suggested by the best currently known efficient algorithms. We now turn to planted clique, which is perhaps the most classical model exhibiting a computational-statistical gap.

The objective of the present chapter is to explain how the SQ predictions perform with the planted-clique problem, as discussed in the influential work of [FGR+17]. The first issue is that the standard planted-clique model is not i.i.d., whereas the SQ framework is naturally formulated for i.i.d. data. The chapter therefore begins by reviewing the usual computational landscape of planted clique and then passes to the distributional bipartite planted-clique model, which is an i.i.d. version introduced in [FGR+17] that is polynomial-time equivalent to the original one. After that, we compute the basic correlation quantity

$$\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle$$

for the one-sample laws of the distributional model, and use the overlap $|x \cap x'|$ to derive the SQ lower-bound heuristic.

10.2 The planted-clique setting

We work with the usual planted-clique detection problem:

$$\mathbb{P}_1 : \ G \sim \text{uniform } k\text{-clique} \cup \mathcal{G}_{n,\frac{1}{2}}, \qquad \mathbb{P}_2 : \ G \sim \mathcal{G}_{n,\frac{1}{2}}.$$

From the statistical side, we already know from earlier chapters that the information-theoretic threshold is logarithmic, around $2\log_2 n$. The computational question is very different: what is the threshold for polynomial-time algorithms?

The first point to explain is why $\sqrt{n}$ appears naturally as the algorithmic threshold.

10.3 Why $\sqrt{n}$ is achievable by efficient methods

10.3.1 Max-degree detection

If k is sufficiently large, one can detect the planted clique simply by looking at the maximum degree. Let $x \subseteq [n]$ denote the planted clique. Under $\mathbb{P}_1$, for a vertex v outside the clique,

$$\deg(v) \sim \mathrm{Bin}\left(n-1, \frac{1}{2}\right),$$

so by the central limit theorem it is concentrated around

$$\frac{n}{2} + \mathcal{O}\left(\sqrt{n}\right).$$

For a vertex $v \in x$, we have

$$\deg(v) = (k-1) + \mathrm{Bin}\left(n-k, \frac{1}{2}\right),$$

hence

$$\deg(v) \approx \frac{n+k}{2} + \mathcal{O}\left(\sqrt{n}\right).$$

Under the null model $\mathbb{P}_2$, every vertex degree is distributed like

$$\deg(v) \sim \mathrm{Bin}\left(n-1, \frac{1}{2}\right),$$

so the maximum degree satisfies

$$\max_v \deg(v) \leq \frac{n}{2} + C\sqrt{n \log n}$$

with high probability.

By contrast, under $\mathbb{P}_1$, every planted vertex has mean degree shifted upward by about $k/2$. Therefore, if

$$k \gg \sqrt{n \log n},$$

then the maximum degree strongly distinguishes the planted and null models.

This explains why the algorithmic threshold is no larger than $\sqrt{n}$ up to logarithmic factors.

10.3.2 A quasi-polynomial-time algorithm below $\sqrt{n}$

We should also note a much slower algorithm that succeeds for far smaller cliques. Although it is not polynomial-time, it is useful because it shows that the computational picture below $\sqrt{n}$ is not completely trivial.

The algorithm is as follows:

Algorithm 10.1 Quasi-polynomial-time detection for planted clique

Require: A graph $G = ([n], E)$ and a target clique size k.

for each subset $S \subseteq [n]$ with $|S| = 10 \log_2 n$ **do**
  if S is a clique in G **then**
    Compute the common neighborhood

$$N(S) := \{v \in [n] \setminus S : \{v, u\} \in E \text{ for every } u \in S\}.$$

    if $|N(S)| \geq k$ **then**
      return $\mathbb{P}_1$.
    end if
  end if
end for
return $\mathbb{P}_2$.

If we are under $\mathbb{P}_1$, then every subset S of the planted clique of size $10 \log_2 n$ is itself a clique, and its common neighborhood contains the rest of the planted clique. Therefore the algorithm certainly declares the planted model.

If we are under $\mathbb{P}_2$, fix a subset S of size $10 \log_2 n$. Conditional on S being a clique, every outside vertex belongs to the common neighborhood of S with probability

$$2^{-|S|} = n^{-10}.$$

Hence the size of the common neighborhood is distributed as

$$\mathrm{Bin}\left(n - |S|, n^{-10}\right).$$

Therefore

$$\mathbb{P}\left(\exists S : \text{ common neighborhood of } S \text{ has size} \geq k\right) \leq \binom{n}{10 \log_2 n} \mathbb{P}\left(\mathrm{Bin}\left(n, n^{-10}\right) \geq k\right).$$

Using a standard Chernoff bound and the fact that the mean of $\mathrm{Bin}\left(n, n^{-10}\right)$ is n^{-9}, we obtain

$$\mathbb{P}\left(\mathrm{Bin}\left(n, n^{-10}\right) \geq k\right) \leq e^{-k/2}$$

for all large enough n. Also,

$$\binom{n}{10 \log_2 n} \leq n^{10 \log_2 n} = e^{C(\log n)^2}$$

for some absolute constant C. Hence

$$\mathbb{P}\left(\exists S : \text{ common neighborhood of } S \text{ has size} \geq k\right) \leq e^{C(\log n)^2 - k/2}.$$

So if

$$k \gg (\log n)^2,$$

the error probability under the null tends to zero.

The running time is

$$\binom{n}{10 \log_2 n} = n^{\mathcal{O}(\log n)},$$

which is quasi-polynomial. Thus planted clique can be detected in quasi-polynomial time far below $\sqrt{n}$, but no polynomial-time algorithm is known there.

So, we can summarize the landscape (all thresholds up to logarithmic factors) is as shown in Figure 10.1.

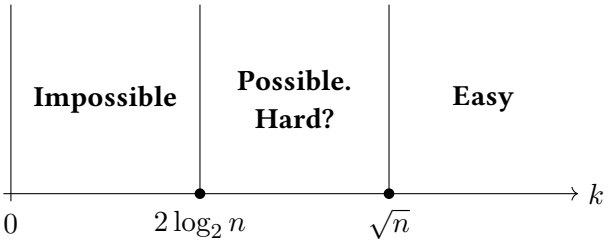

Figure 10.1: Planted clique: information-theoretic and algorithmic thresholds

10.4 Low-degree hardness below $\sqrt{n}$ and the need for an i.i.d. model

From the low-degree perspective, we already proved (see Proposition 7.11) that

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2} \right)_{\leq \log^{1+\varepsilon} n} \right\|_2 \to 1 + o(1)$$

whenever

$$0 < \varepsilon < 1, \qquad 2\log_2 n \ll k \ll n^{1/2-\delta}, \qquad 0 < \delta < 1/2.$$

Thus planted clique has a low-degree barrier below $\sqrt{n}$.

It is natural to ask whether an analogous statement can be proved in the statistical-query framework. The obstacle is that the standard planted-clique model is not i.i.d. The SQ framework, by contrast, is designed for repeated independent samples.

An i.i.d. analogue is *distributional bipartite planted clique*. We now introduce that model.

10.5 Distributional bipartite planted clique

Let $x \in \{0,1\}^n$ be the indicator vector of the planted right-side clique, with

$$\|x\|_0 = k.$$

We think of observing the rows of a bipartite adjacency matrix, one row at a time. Each row is an element of $\{0,1\}^n$.

Definition 10.2 (Distributional bipartite clique [FGR+17, Problem 1.1.])**.** *The one-sample null law is*

$$q = \mathrm{Unif}(\{0,1\}^n).$$

Given a planted support vector $x \in \{0,1\}^n$ with $\|x\|_0 = k$, the one-sample planted law p_x is defined as follows:

- *with probability $1 - \frac{k}{n}$, sample $Y \sim \mathrm{Unif}(\{0,1\}^n)$;*
- *with probability $\frac{k}{n}$, sample Y uniformly among all binary vectors that are equal to 1 on $\mathrm{supp}(x)$.*

Equivalently, if $Y \sim p_x$, then

$$Y \sim p_x = \left(1 - \frac{k}{n}\right) \mathrm{Unif}(\{0,1\}^n) + \frac{k}{n}\, \mathbf{1}(\mathrm{supp}(x)) \cup \mathrm{Unif}(\{0,1\}^{[n]\setminus \mathrm{supp}(x)}).$$

Thus the i.i.d. detection problem becomes:

$$\text{either } A_1, \dots, A_n \overset{\text{i.i.d.}}{\sim} p_x \text{ for some } x \sim \mu, \qquad \text{or } A_1, \dots, A_n \overset{\text{i.i.d.}}{\sim} q,$$

where μ is the uniform distribution over all k-subsets of $[n]$.

The reason this model is relevant is given by the following theorem.

Theorem 10.3. *Distributional planted clique is polynomial-time equivalent to bipartite planted clique.*

Proof. See [FGR+17, Theorems A.2., A.3.]. □

The point is that SQ lower bounds for the distributional model should be interpreted as evidence for hardness in the original planted-clique problem as well.

10.6 Computing the basic correlation quantity

To prove an SQ lower bound, we need to control

$$\mathop{\mathbb{E}}_{x,x' \overset{\text{i.i.d.}}{\sim} \mu} \left[\left| \left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle - 1 \right| \middle| A \right]$$

for events A of non-negligible probability.

So we first compute the inner product explicitly.

Let $Y \in \{0,1\}^n$. Since q is uniform,

$$q(Y) = 2^{-n}.$$

Also, by definition of p_x,

$$p_x(Y) = \left(1 - \frac{k}{n}\right) 2^{-n} + \frac{k}{n} 2^{-(n-k)} \mathbf{1}(Y|_x = 1),$$

where $\mathbf{1}(Y|_x = 1)$ denotes the indicator that $Y_j = 1$ for every $j \in \operatorname{supp}(x)$.

Hence

$$\frac{p_x}{q}(Y) = 2^n p_x(Y) = 1 - \frac{k}{n} + \frac{k}{n} 2^k \mathbf{1}(Y|_x = 1).$$

We now compute the inner product. Since $Y \sim q$ is uniform on $\{0,1\}^n$,

$$\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle = \mathop{\mathbb{E}}_{Y \sim q} \left[\frac{p_x}{q}(Y) \frac{p_{x'}}{q}(Y) \right].$$

Expanding,

$$\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle = \mathop{\mathbb{E}}_{Y \sim q} \left[\left(1 - \frac{k}{n} + \frac{k}{n} 2^k \mathbf{1}(Y|_x = 1)\right) \left(1 - \frac{k}{n} + \frac{k}{n} 2^k \mathbf{1}(Y|_{x'} = 1)\right) \right].$$

Now

$$\mathop{\mathbb{P}}_{Y \sim q}(Y|_x = 1) = 2^{-k}, \qquad \mathop{\mathbb{P}}_{Y \sim q}(Y|_{x'} = 1) = 2^{-k},$$

and

$$\mathop{\mathbb{P}}_{Y \sim q}(Y|_x = 1,\ Y|_{x'} = 1) = 2^{-2k + |x \cap x'|}.$$

Therefore

$$\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle = \left(1 - \frac{k}{n}\right)^2 + 2\left(1 - \frac{k}{n}\right)\frac{k}{n} + \frac{k^2}{n^2} 2^{|x \cap x'|}.$$

The first two terms simplify:

$$\left(1 - \frac{k}{n}\right)^2 + 2\left(1 - \frac{k}{n}\right)\frac{k}{n} = 1 - \frac{k^2}{n^2}.$$

So we obtain the exact formula

$$\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle = 1 + \frac{k^2}{n^2}\left(2^{|x \cap x'|} - 1\right).$$

It follows that

$$\mathop{\mathbb{E}}_{x,x' \overset{\text{i.i.d.}}{\sim} \mu} \left[\left| \left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle - 1 \right| \,\middle|\, A \right] = \mathop{\mathbb{E}}_{x,x' \overset{\text{i.i.d.}}{\sim} \mu} \left[\frac{k^2}{n^2}\left(2^{|x \cap x'|} - 1\right) \,\middle|\, A \right].$$

10.7 The SQ lower bound below $\sqrt{n}$

We now connect the correlation computation to the statistical dimension definition. Let

$$R := \left|x \cap x'\right|, \qquad x, x' \overset{\text{i.i.d.}}{\sim} \mu = \mathrm{Unif}\left(\binom{[n]}{k}\right).$$

From the previous section,

$$\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle - 1 = \frac{k^2}{n^2}\left(2^R - 1\right).$$

Thus the SQ lower bound reduces to controlling conditional moments of 2^R under events of non-negligible probability.

Recall the Definition 8.3 of statistical dimension. To prove that

$$\mathrm{SDA}(p, q, m) \geq S,$$

it is enough to show that, for every event

$$A \subseteq \mathrm{supp}(\mu) \times \mathrm{supp}(\mu), \qquad \mathop{\mathbb{P}}_{x,x' \overset{\text{i.i.d.}}{\sim} \mu}(A) \geq S^{-2},$$

we have

$$\mathop{\mathbb{E}}_{x,x' \overset{\text{i.i.d.}}{\sim} \mu} \left[\left| \left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle - 1 \right| \,\middle|\, A \right] \leq \frac{1}{m}.$$

In our setting the absolute value is unnecessary, since the correlation is always at least 1. Therefore it is enough to prove

$$\frac{k^2}{n^2} \mathop{\mathbb{E}}_{x,x' \overset{\text{i.i.d.}}{\sim} \mu} \left[2^R - 1 \mid A\right] \leq \frac{1}{m}.$$

Since the i.i.d. bipartite model has n samples, the natural VSTAT parameter is $m = \Theta(n)$. Thus the condition becomes

$$\mathop{\mathbb{E}}_{x,x' \overset{\text{i.i.d.}}{\sim} \mu} \left[2^R - 1 \mid A\right] \lesssim \frac{n}{k^2}.$$

We now prove that this holds for all events A of probability at least S^{-2}, where

$$S = \exp\left(c\left(\log\frac{n}{k^2}\right)^2\right)$$

for a sufficiently small constant $c > 0$, provided $k \le n^{1/2-\varepsilon}$.

Since $k \le n^{1/2-\varepsilon}$, we have $\frac{n}{k^2} \to \infty$ polynomially fast. The overlap

$$R = \left|x \cap x'\right|$$

has a $\mathrm{Hyp}(n, k, k)$ distribution. Thus, for a fixed x,

$$\mathbb{P}(R = s) = \frac{\binom{k}{s}\binom{n-k}{k-s}}{\binom{n}{k}} \le \binom{k}{s}\left(\frac{k}{n-k}\right)^s \le \left(\frac{Ck^2}{ns}\right)^s$$

for a universal constant $C > 0$.

We now prove the conditional moment bound. Choose

$$r_0 = \left\lfloor \log_2\left(\frac{n}{k^2}\right) - B \right\rfloor$$

with $B > 0$ any constant. Then

$$2^{r_0} < \frac{n}{k^2}.$$

For any event A with $\mathbb{P}(A) \ge S^{-2}$,

$$\begin{aligned}
\mathbb{E}\left[2^R \mid A\right] &= \frac{1}{\mathbb{P}(A)}\left(\mathbb{E}\left[2^R \mathbf{1}_{A\cap\{R\le r_0\}}\right] + \mathbb{E}\left[2^R \mathbf{1}_{A\cap\{R>r_0\}}\right]\right) \\
&\le \frac{1}{\mathbb{P}(A)}\left(2^{r_0}\mathbb{P}(A\cap\{R\le r_0\}) + \mathbb{E}\left[2^R \mathbf{1}_{\{R>r_0\}}\right]\right) \\
&\le 2^{r_0} + \frac{1}{\mathbb{P}(A)}\mathbb{E}\left[2^R \mathbf{1}_{\{R>r_0\}}\right].
\end{aligned}$$

Using the hypergeometric probability inequality and the fact that $\frac{k^2}{n} = o(1)$ we get

$$\begin{aligned}
\mathbb{E}\left[2^R \mathbf{1}_{\{R>r_0\}}\right] &= \sum_{r=r_0+1}^{k} 2^r \mathbb{P}(R=r) \le \sum_{r=r_0+1}^{k}\left(\frac{2Ck^2}{nr}\right)^r \le \sum_{r=r_0+1}^{k}\left(\frac{2Ck^2}{nr_0}\right)^r \\
&= \left(\frac{2Ck^2}{nr_0}\right)^{r_0+1} \frac{1-\left(\frac{2Ck^2}{nr_0}\right)^{k-r_0}}{1-\frac{2Ck^2}{nr_0}} \le 2\left(\frac{2Ck^2}{nr_0}\right)^{r_0+1} \\
&= 2\exp\left(-(r_0+1)\log\frac{nr_0}{2Ck^2}\right)
\end{aligned}$$

Since r_0 is of order $\log\frac{n}{k^2}$, this last sum is at most

$$\exp\left(-c_0\left(\log\frac{n}{k^2}\right)^2\right)$$

for some constant $c_0 > 0$. Taking

$$S = \exp\left(c\left(\log\frac{n}{k^2}\right)^2\right)$$

with $\frac{c_0}{2} > c > 0$, we get

$$\frac{1}{\mathbb{P}(A)}\mathbb{E}\left[2^R \mathbf{1}_{\{R>r_0\}}\right] \le S^2 \exp\left(-c_0\left(\log\frac{n}{k^2}\right)^2\right) \le 1$$

for all large n. Therefore

$$\mathbb{E}\left[2^R - 1 \mid A\right] \le 2^{r_0} + \frac{1}{\mathbb{P}(A)}\left(\mathbb{E}\left[2^R \mathbf{1}_{\{R>r_0\}}\right] - 1\right) \le 2^{r_0} < \frac{n}{k^2}.$$

Consequently,

$$\mathbb{E}\left[\left|\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle - 1\right| \,\middle|\, A\right] < \frac{k^2}{n^2}\frac{n}{k^2} = \frac{1}{n}.$$

This, by Definition 8.3, shows that

$$\mathrm{SDA}(p, q, \Theta(n)) \ge \exp\left(c\left(\log \frac{n}{k^2}\right)^2\right).$$

By the SQ lower bound Theorem 8.4, any SQ algorithm using a VSTAT oracle with sample size $\Theta(n)$ must make at least

$$\exp\left(\Omega\left(\left(\log \frac{n}{k^2}\right)^2\right)\right)$$

queries to solve the distributional bipartite planted-clique detection problem.

Remark 10.4 (The square-root scale)**.** *The SQ lower bound obtained above has the expected transition at the square-root scale. As k approaches the square-root scale, the quantity n/k^2 becomes bounded, and the exponent*

$$\left(\log \frac{n}{k^2}\right)^2$$

degenerates. Thus the SQ lower bound disappears precisely around $k \approx \sqrt{n}$. This matches the algorithmic picture discussed earlier in the chapter: degree-based methods become effective at the square-root scale, while below this scale no polynomial-time algorithm is known. Therefore, in the distributional bipartite planted-clique model, the SQ framework recovers the same qualitative computational threshold predicted by the usual planted-clique conjecture.

10.8 Conclusion

In this chapter we studied planted clique from the statistical-query point of view. The standard planted-clique model is not an i.i.d. model, so we first passed to the distributional bipartite planted-clique model, where each row of the bipartite adjacency matrix is an independent sample. This allowed us to apply the SQ framework developed earlier.

The explicit correlation formula

$$\left\langle \frac{p_x}{q}, \frac{p_{x'}}{q} \right\rangle = 1 + \frac{k^2}{n^2}\left(2^{|x\cap x'|} - 1\right)$$

was the key computation that showed that the SQ lower bound is controlled by the overlap $|x \cap x'|$ between two independent planted supports. Since this overlap is typically very small when $k \ll \sqrt{n}$, the correlations between different planted distributions remain small on all sufficiently large conditioning events. This yields a large statistical dimension and therefore a large SQ query lower bound.

For another quite influential and pedagogical SQ lower bound we direct the interested reader to [DKS17] for the study of the so-called Non-Gaussian component analysis setting.

Chapter 11

Almost Equivalence Between Low-Degree and Statistical Queries

11.1 Introduction

We now discuss an important bridge between two computational lower-bound frameworks that have appeared throughout these notes: low-degree polynomials and statistical queries. The result is from [BBH+21]. It shows that, under suitable assumptions, low-degree likelihood-ratio lower bounds and VSTAT lower bounds are almost equivalent.

Some assumptions, though, seem necessary. Consider the original one-sample planted clique model. If $k \gg \log n$, then a single $\mathrm{VSTAT}(4)$ query

$$h(G) = \mathbf{1}\{G \text{ contains a clique of size } k\}$$

can distinguish the planted and null models (left as exercise for the reader). This is not reflected by low-degree polynomials (and rightfully so): as we discussed, the low-degree likelihood ratio for planted clique remains bounded below the $\sqrt{n}$ scale. Therefore a completely general equivalence between SQ and low-degree cannot hold. We will see here that a condition that works, is that there should be no *very* powerful one-sample high-degree test. This may sound circular, but it is natural in noise-robust models, where adding a small amount of noise in one-sample *washes out* high-degree information. The goal of this chapter is to make this precise.

11.2 The i.i.d. Detection Setting

We focus on the i.i.d. planted-vs-null testing problem. There is a parameter $\theta \sim \mu$, and conditional on θ the samples are drawn from p_θ. The null sample distribution is q. Thus the m-sample testing problem is

$$\mathbb{P}_1 : Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} p_\theta \quad \text{for } \theta \sim \mu,$$

against

$$\mathbb{P}_2 : Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} q.$$

Let

$$D_\theta := \frac{p_\theta}{q}$$

be the one-sample likelihood ratio, and let

$$D := \frac{p}{q} = \mathop{\mathbb{E}}_{\theta\sim\mu}[D_\theta]$$

be the averaged one-sample likelihood ratio.

For the m-sample model, the likelihood ratio conditional on θ is

$$D_\theta^{\otimes m}(Y_1, \dots, Y_m) = \prod_{i=1}^{m} D_\theta(Y_i).$$

The averaged m-sample likelihood ratio is

$$\mathop{\mathbb{E}}_{\theta\sim\mu}\left[D_\theta^{\otimes m}\right].$$

All L^2 norms and inner products in this chapter are taken with respect to the appropriate null distribution. For one sample this is q, and for m samples this is $q^{\otimes m}$.

11.3 Samplewise Degree

The usual low-degree method uses the total degree of a polynomial in all observed variables. In the i.i.d. setting it is useful to refine this notion and keep track of the degree used in each sample.

Definition 11.1 (Samplewise degree)**.** *Let $f : (\mathbb{R}^n)^m \to \mathbb{R}$ be a polynomial in m samples*

$$x_1, \dots, x_m \in \mathbb{R}^n.$$

We say that f has samplewise degree (d, k) if it can be written as a linear combination of monomials that have degree at most d in each individual sample x_i, and have nonzero degree in at most k of the samples.

Thus d controls the degree per sample, while k controls how many samples the monomial actually uses.

This notion is related to the usual total degree as follows.

Remark 11.2. *If a polynomial has samplewise degree (d, k), then its usual total degree is at most dk. Conversely, if a polynomial has usual total degree at most d, then it has samplewise degree at most (d, d).*

The first implication is immediate: at most k samples appear, each with degree at most d. For the second, a monomial of total degree at most d can involve at most d samples nontrivially, and its degree in each individual sample is also at most d.

11.4 Samplewise Low-Degree Likelihood Ratios

Let $D_\theta^{\le d}$ denote the projection of D_θ onto one-sample polynomials of degree at most d, and define

$$D_\theta^{>d} := D_\theta - D_\theta^{\le d}.$$

Since

$$\mathop{\mathbb{E}}_{Y\sim q}[D_\theta(Y)] = 1,$$

the constant component of $D_\theta^{\le d}$ is 1. Therefore it is useful to write

$$A_\theta := D_\theta^{\le d} - 1.$$

Then A_θ is the nonconstant degree-$\le d$ part of the one-sample likelihood ratio.

For m samples, define the (d,k)-low-degree likelihood ratio by projecting the m-sample likelihood ratio onto the subspace of samplewise degree (d,k):

$$\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} := \left(\mathop{\mathbb{E}}_{\theta\sim\mu}\left[D_\theta^{\otimes m}\right]\right)_{\le d,k}.$$

Equivalently,

$$\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} = \mathop{\mathbb{E}}_{\theta\sim\mu}\left[\left(D_\theta^{\otimes m}\right)_{\le d,k}\right].$$

Let us compute this projection explicitly. Since

$$D_\theta^{\le d} = 1 + A_\theta,$$

the samplewise degree (d,k) projection of $D_\theta^{\otimes m}$ is

$$\left(D_\theta^{\otimes m}\right)_{\le d,k} = \sum_{\substack{S\subseteq[m]\\ |S|\le k}} A_\theta^{\otimes S},$$

where

$$A_\theta^{\otimes S} := \prod_{i\in S} A_\theta(Y_i),$$

and the empty product is 1.

Hence

$$\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1 = \sum_{\substack{S\subseteq[m]\\ 1\le|S|\le k}} \mathop{\mathbb{E}}_{\theta\sim\mu}\left[A_\theta^{\otimes S}\right].$$

The following lemma is the key decomposition of the samplewise low-degree norm.

Lemma 11.3 (Samplewise low-degree norm decomposition)**.** *For every d,k,m,*

$$\left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1\right\|^2_{L^2(q^{\otimes m})} = \sum_{t=1}^{k}\binom{m}{t}\mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\left(\left\langle D_\theta^{\le d}, D_{\theta'}^{\le d}\right\rangle_{L^2(q)} - 1\right)^t\right].$$

Moreover, for every $t\ge 0$,

$$\mathop{\mathbb{E}}_{\theta,\theta'\sim\mu}\left[\left(\left\langle D_\theta^{\le d}, D_{\theta'}^{\le d}\right\rangle_{L^2(q)} - 1\right)^t\right] \ge 0.$$

Proof. Recall that

$$A_\theta = D_\theta^{\le d} - 1.$$

Since A_θ is orthogonal to constants in $L^2(q)$,

$$\langle A_\theta, 1\rangle_{L^2(q)} = 0.$$

Also,

$$\langle A_\theta, A_{\theta'}\rangle_{L^2(q)} = \left\langle D_\theta^{\le d} - 1, D_{\theta'}^{\le d} - 1\right\rangle = \left\langle D_\theta^{\le d}, D_{\theta'}^{\le d}\right\rangle - 1.$$

We have

$$\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\leq d,k} - 1 = \sum_{\substack{S\subseteq[m]\\ 1\leq|S|\leq k}} \mathop{\mathbb{E}}_{\theta}\left[A_\theta^{\otimes S}\right].$$

Therefore,

$$\left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\leq d,k} - 1\right\|^2 = \sum_{\substack{S,S'\subseteq[m]\\ 1\leq|S|,|S'|\leq k}} \left\langle \mathop{\mathbb{E}}_{\theta}\left[A_\theta^{\otimes S}\right], \mathop{\mathbb{E}}_{\theta'}\left[A_{\theta'}^{\otimes S'}\right]\right\rangle.$$

If $S \neq S'$, then there exists an index i belonging to exactly one of the two sets. In the i-th sample coordinate, one factor is A_θ and the other factor is the constant 1. Since A_θ is orthogonal to constants, the inner product is zero. Hence only the terms $S = S'$ remain.

For a fixed S with $|S| = t$,

$$\left\|\mathop{\mathbb{E}}_{\theta}\left[A_\theta^{\otimes S}\right]\right\|^2 = \mathop{\mathbb{E}}_{\theta,\theta'}\left[\langle A_\theta, A_{\theta'}\rangle^t\right] = \mathop{\mathbb{E}}_{\theta,\theta'}\left[\left(\left\langle D_\theta^{\leq d}, D_{\theta'}^{\leq d}\right\rangle - 1\right)^t\right].$$

There are $\binom{m}{t}$ subsets S of size t, so summing over $t = 1, \ldots, k$ gives the desired identity.

Finally,

$$\mathop{\mathbb{E}}_{\theta,\theta'}\left[\langle A_\theta, A_{\theta'}\rangle^t\right] = \left\|\mathop{\mathbb{E}}_{\theta}\left[A_\theta^{\otimes t}\right]\right\|^2 \geq 0.$$

This proves the nonnegativity claim. □

This lemma is the precise form of the *replica trick* in the samplewise low-degree setting.

11.5 From Low-Degree Lower Bounds to SQ Lower Bounds

We now prove the first direction: a sufficiently strong samplewise low-degree lower bound implies an SQ lower bound.

We use the following characterization of statistical dimension. In this chapter,

$$\mathrm{SDA}(m') \geq r$$

means that for every event B in the pair space (θ, θ') satisfying

$$\mathbb{P}(B) \geq \frac{1}{r^2},$$

one has

$$\mathop{\mathbb{E}}_{\theta,\theta'}\left[\left|\langle D_\theta, D_{\theta'}\rangle_{L^2(q)} - 1\right| \,\middle|\, B\right] \leq \frac{1}{m'}.$$

Thus m' plays the role of the effective sample size of the VSTAT oracle, while r is the statistical dimension lower bound.

Theorem 11.4 (Low-degree lower bound implies SQ lower bound)**.** *Let $d, k \in \mathbb{N}$, with k even. Suppose that for some $\delta, \varepsilon > 0$,*

$$\left\|\mathop{\mathbb{E}}_{\theta\sim\mu}\left[(D_\theta^{>d})^{\otimes k}\right]\right\| \leq \delta$$

and

$$\left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\leq d,k} - 1\right\| \leq \varepsilon.$$

Then, for every $r > 0$,

$$\mathrm{SDA}\left(\frac{m}{2r^{2/k}\left(k\varepsilon^{2/k} + \delta^{2/k}m\right)}\right) \geq r.$$

Proof. Let

$$X(\theta, \theta') := \langle D_\theta, D_{\theta'}\rangle - 1.$$

We want to prove that for every event B with $\mathbb{P}(B) \geq 1/r^2$,

$$\mathbb{E}[|X| \mid B] \leq \frac{1}{m'},$$

where

$$m' := \frac{m}{2r^{2/k}\left(k\varepsilon^{2/k} + \delta^{2/k}m\right)}.$$

By Hölder's inequality,

$$\mathbb{E}[|X| \mid B] = \frac{\mathbb{E}[|X|\mathbf{1}_B]}{\mathbb{P}(B)} \leq \frac{\mathbb{E}[|X|^k]^{1/k}\mathbb{P}(B)^{1-1/k}}{\mathbb{P}(B)} = \left(\frac{\mathbb{E}\left[|X|^k\right]}{\mathbb{P}(B)}\right)^{1/k}.$$

Since $\mathbb{P}(B) \geq 1/r^2$, it is enough to prove

$$\mathbb{E}\left[|X|^k\right]^{1/k} \leq \frac{1}{m'r^{2/k}}.$$

We now bound this k-th moment.

Write

$$A_\theta := D_\theta^{\leq d} - 1, \qquad B_\theta := D_\theta^{>d}.$$

Then

$$D_\theta - 1 = A_\theta + B_\theta.$$

Because A_θ is degree at most d and $B_{\theta'}$ is orthogonal to all degree-$\leq d$ functions,

$$\langle A_\theta, B_{\theta'}\rangle = 0.$$

Also both A_θ and B_θ are orthogonal to constants. Hence

$$X(\theta, \theta') = \langle D_\theta - 1, D_{\theta'} - 1\rangle = \langle A_\theta, A_{\theta'}\rangle + \langle B_\theta, B_{\theta'}\rangle.$$

Define

$$X_{\leq d} := \langle A_\theta, A_{\theta'}\rangle = \left\langle D_\theta^{\leq d}, D_{\theta'}^{\leq d}\right\rangle - 1$$

and

$$X_{>d} := \langle B_\theta, B_{\theta'}\rangle.$$

Then $X = X_{\leq d} + X_{>d}$.

By the triangle inequality in L^k,

$$\mathbb{E}\left[|X|^k\right]^{1/k} \leq \mathbb{E}\left[|X_{\leq d}|^k\right]^{1/k} + \mathbb{E}\left[|X_{>d}|^k\right]^{1/k}.$$

Since k is even,

$$\mathbb{E}\left[|X_{\leq d}|^k\right] = \mathbb{E}\left[X_{\leq d}^k\right] = \left\|\mathop{\mathbb{E}}_{\theta}\left[A_\theta^{\otimes k}\right]\right\|^2.$$

By the samplewise low-degree norm decomposition,

$$\binom{m}{k}\mathbb{E}\left[X_{\le d}^k\right] \le \left\|\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1\right\|^2 \le \varepsilon^2.$$

Therefore

$$\mathbb{E}\left[|X_{\le d}|^k\right]^{1/k} \le \frac{\varepsilon^{2/k}}{\binom{m}{k}^{1/k}}.$$

Using

$$\binom{m}{k} \ge \left(\frac{m}{k}\right)^k$$

for $m \ge k$, we get

$$\mathbb{E}\left[|X_{\le d}|^k\right]^{1/k} \le \frac{k\varepsilon^{2/k}}{m}.$$

Similarly,

$$\mathbb{E}\left[|X_{>d}|^k\right] = \mathbb{E}\left[\langle B_\theta, B_{\theta'}\rangle^k\right] = \left\|\mathbb{E}_\theta\left[B_\theta^{\otimes k}\right]\right\|^2 \le \delta^2,$$

so

$$\mathbb{E}\left[|X_{>d}|^k\right]^{1/k} \le \delta^{2/k}.$$

Combining the two estimates,

$$\mathbb{E}\left[|X|^k\right]^{1/k} \le \frac{k\varepsilon^{2/k}}{m} + \delta^{2/k}.$$

In particular,

$$\mathbb{E}\left[|X|^k\right]^{1/k} \le 2\left(\frac{k\varepsilon^{2/k}}{m} + \delta^{2/k}\right) = \frac{1}{m'r^{2/k}},$$

by the definition of m'. Therefore

$$\mathbb{E}[|X| \mid B] \le \frac{1}{m'}.$$

Since this holds for every B with $\mathbb{P}(B) \ge 1/r^2$, we conclude

$$\mathrm{SDA}(m') \ge r.$$

□

11.5.1 Interpreting the theorem

Let us spell out what Theorem 11.4 gives in the regime that typically appears in applications. We usually choose the samplewise degree parameters slightly above logarithmic scale, for instance

$$d \approx k \approx (\log m)^{1.1}.$$

Suppose that the high-degree one-sample term is very small, say

$$\delta \le m^{-k/2},$$

and that the samplewise low-degree norm is bounded by a constant independent of m:

$$\varepsilon = O(1).$$

More generally, the same discussion applies whenever

$$\log \varepsilon = o(k),$$

since then

$$\varepsilon^{2/k} = 1 + o(1).$$

Under these assumptions,

$$\delta^{2/k} m \le (m^{-k/2})^{2/k} m = 1.$$

Moreover, if $\varepsilon = O(1)$, then

$$\varepsilon^{2/k} = 1 + o(1).$$

Therefore the denominator in Theorem 11.4 satisfies

$$2r^{2/k}\left(k\varepsilon^{2/k} + \delta^{2/k} m\right) = 2r^{2/k}\left(k(1+o(1)) + 1\right).$$

Since $k \to \infty$, this is

$$2r^{2/k} k(1+o(1)).$$

Thus the theorem gives, for every $r > 0$,

$$\mathrm{SDA}\left(\frac{m}{2r^{2/k}k(1+o(1))}\right) \ge r.$$

Now choose

$$r = m^{\varepsilon' k/2}$$

for a small constant $\varepsilon' > 0$. Then

$$r^{2/k} = m^{\varepsilon'},$$

and hence

$$\mathrm{SDA}\left(\frac{m^{1-\varepsilon'}}{2k(1+o(1))}\right) \ge m^{\varepsilon' k/2}.$$

Since in our applications k is polylogarithmic in m, the factor $2k(1+o(1))$ is negligible at the level of polynomial exponents. Thus we may summarize the conclusion as

$$\mathrm{SDA}\left(m^{1-\varepsilon'}\right) \ge m^{\Theta(k)}.$$

In words, with roughly $m^{1-\varepsilon'}$ samples, any SQ algorithm requires $m^{\Theta(k)}$ queries.

More generally, setting

$$r = \left(\frac{m}{m'}\right)^{k/2}$$

gives

$$r^{2/k} = \frac{m}{m'}.$$

The theorem then yields

$$\mathrm{SDA}\left(\frac{m'}{2k(1+o(1))}\right) \ge \left(\frac{m}{m'}\right)^{k/2}.$$

Again ignoring the polylogarithmic factor $2k(1+o(1))$, this is the useful family of lower bounds

$$\mathrm{SDA}(m') \ge \left(\frac{m}{m'}\right)^{k/2}, \qquad m' \le m.$$

This is the form that is most directly comparable with the converse direction.

11.6 From SQ Lower Bounds to Low-Degree Lower Bounds

We now discuss the reverse direction: sufficiently strong SQ lower bounds imply low-degree likelihood-ratio lower bounds.

We state the theorem in the latter form.

Theorem 11.5 (SQ lower bounds imply samplewise low-degree lower bounds)**.** *Let $M_n \to \infty$. Suppose that, for every $m' \le M_n m$,*

$$\mathrm{SDA}(m') \ge \left(\frac{M_n m}{m'}\right)^k.$$

Then, for every d,

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1 \right\|_{L^2(\mathbb{P}_2)} = o(1).$$

Before proceeding with the convenient proof, we first need an easy result from probability theory, stated as a fact:

Fact 11.6. *If $p > q > 0$, then*

$$\mathbb{E}\left[|X^q|\right] \le 2^{\frac{q}{p}} \sup_A \left\{\mathbb{P}(A)^p \mathbb{E}[|X| \mid A]^q\right\} \frac{p}{p-q}.$$

Proof. See [HL19, Fact A.2.] □

Proof of Theorem 11.5. Let

$$X(\theta, \theta') := \langle D_\theta, D_{\theta'} \rangle_{L^2(q)} - 1.$$

The assumption

$$\mathrm{SDA}\left(m'\right) \ge \left(\frac{M_n m}{m'}\right)^k$$

means the following: for every event A in the pair space (θ, θ') such that

$$\mathbb{P}(A) \ge \left(\frac{m'}{M_n m}\right)^{2k},$$

we have

$$\mathbb{E}[|X| \mid A] \le \frac{1}{m'}.$$

We first derive moment bounds for X. Fix any event A with $\mathbb{P}(A) > 0$ and choose

$$m' := M_n m\, \mathbb{P}(A)^{1/(2k)}.$$

Then $m' \le M_n m$, and by construction

$$\mathbb{P}(A) = \left(\frac{m'}{M_n m}\right)^{2k}.$$

Hence the SDA condition gives

$$\mathbb{E}[|X| \mid A] \le \frac{1}{M_n m\, \mathbb{P}(A)^{1/(2k)}}.$$

Equivalently,

$$\mathbb{P}(A)^{1/(2k)}\mathbb{E}[|X| \mid A] \le \frac{1}{M_n m}.$$

Therefore, for every $1 \le t \le k$,

$$\mathbb{P}(A)^{t/(2k)}\mathbb{E}[|X| \mid A]^t \le \left(\frac{1}{M_n m}\right)^t.$$

We now apply Fact 11.6 with

$$p = 2k, \qquad q = t.$$

Since $t \le k$, we have $p > q$. The fact gives

$$\mathbb{E}[|X|^t] \le 2^{t/(2k)} \frac{2k}{2k-t} \sup_A \left\{ \mathbb{P}(A)^{t/(2k)}\mathbb{E}[|X| \mid A]^t \right\}.$$

Using the previous bound and $t \le k$,

$$2^{t/(2k)} \frac{2k}{2k-t} \le 4.$$

Thus

$$\mathbb{E}[|X|^t] \le 4\left(\frac{1}{M_n m}\right)^t \qquad \text{for every } 1 \le t \le k.$$

We now relate these moment bounds to the samplewise low-degree norm. Let

$$A_\theta := D_\theta^{\le d} - 1.$$

Then

$$\langle A_\theta, A_{\theta'} \rangle = \left\langle D_\theta^{\le d}, D_{\theta'}^{\le d} \right\rangle - 1.$$

For every $t \ge 1$, the replica trick gives

$$\mathop{\mathbb{E}}_{\theta,\theta'} \left[\langle A_\theta, A_{\theta'} \rangle^t \right] = \left\| \mathop{\mathbb{E}}_{\theta} \left[A_\theta^{\otimes t} \right] \right\|^2.$$

Since A_θ is the degree-$\le d$ projection of $D_\theta - 1$, the tensor

$$A_\theta^{\otimes t}$$

is the orthogonal projection of

$$(D_\theta - 1)^{\otimes t}$$

onto the corresponding tensor-product low-degree subspace. Orthogonal projection is a contraction in L^2, so

$$\left\| \mathop{\mathbb{E}}_{\theta} \left[A_\theta^{\otimes t} \right] \right\| \le \left\| \mathop{\mathbb{E}}_{\theta} \left[(D_\theta - 1)^{\otimes t} \right] \right\|.$$

Applying the replica trick again,

$$\left\| \mathop{\mathbb{E}}_{\theta} \left[(D_\theta - 1)^{\otimes t} \right] \right\|^2 = \mathop{\mathbb{E}}_{\theta,\theta'} \left[(\langle D_\theta, D_{\theta'} \rangle - 1)^t \right] \le \mathbb{E}\left[|X|^t \right].$$

Therefore

$$\mathop{\mathbb{E}}_{\theta,\theta'} \left[\left(\left\langle D_\theta^{\le d}, D_{\theta'}^{\le d} \right\rangle - 1 \right)^t \right] \le 4\left(\frac{1}{M_n m}\right)^t.$$

Using the samplewise low-degree norm decomposition,

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1 \right\|^2 = \sum_{t=1}^{k} \binom{m}{t} \mathop{\mathbb{E}}_{\theta,\theta'} \left[\left(\left\langle D_\theta^{\le d}, D_{\theta'}^{\le d} \right\rangle - 1 \right)^t \right].$$

Hence

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1 \right\|^2 \le 4 \sum_{t=1}^{k} \binom{m}{t} \left(\frac{1}{M_n m}\right)^t.$$

Using

$$\binom{m}{t} \le \left(\frac{em}{t}\right)^t,$$

we obtain

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1 \right\|^2 \le 4 \sum_{t=1}^{k} \left(\frac{e}{M_n t}\right)^t.$$

Since $M_n \to \infty$, the right-hand side is $o(1)$. Therefore

$$\left\| \left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right)_{\le d,k} - 1 \right\| = o(1),$$

as claimed. □

11.7 Discussion of the High-Degree Assumption

The direction from low-degree lower bounds to SQ lower bounds required the assumption

$$\left\| \mathop{\mathbb{E}}_{\theta\sim\mu} \left[\left(D_\theta^{>d} \right)^{\otimes k} \right] \right\| \le \delta.$$

This is the main additional assumption in the equivalence. It says that the averaged k-fold tensor of the high-degree one-sample likelihood ratio is small. Intuitively, it rules out the possibility that the problem has a very effective high-degree one-sample test.

As discussed at the beginning of the chapter, planted clique in the original one-sample graph model violates this assumption. The high-degree statistic

$$h(G) = \mathbf{1}\{G \text{ contains a } k\text{-clique}\}$$

distinguishes the planted and null distributions when $k \gg \log n$, even though low-degree polynomials do not see the planted clique below the $\sqrt{n}$ scale.

However, the assumption is natural for noise-robust problems, which the planted clique problem is not. The reason is that adding a small amount of noise strongly damps high-degree components.

11.7.1 Noise operators

Let Q be the null distribution and consider a linear operator

$$T : L^2(Q) \to L^2(Q).$$

We say that T is a (d, ε)-operator if it contracts the degree-$> d$ subspace by a factor at most ε:

$$\|Tf\|_{L^2(Q)} \le \varepsilon \|f\|_{L^2(Q)} \qquad \text{for every } f \in L^2(Q)^{>d}.$$

Equivalently, if T is diagonalizable in an orthogonal polynomial basis, all eigenfunctions of degree greater than d have eigenvalues of absolute value at most ε.

Many natural noise operators have this property. For example, the Ornstein–Uhlenbeck operator on Gaussian space contracts degree-r Hermite polynomials by a factor ρ^r. Hence it is a (d, ρ^{d+1})-operator.

Similarly, the discrete noise operator on the hypercube contracts degree-r Fourier characters by a factor ρ^r, and is therefore also a (d, ρ^{d+1})-operator.

11.7.2 Noise robustness in spiked models

Consider a rank-one Gaussian spiked model

$$Y = \lambda\theta\theta^\top + W, \qquad W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Let $D_{\theta,\lambda}$ denote the likelihood ratio at signal strength λ.

If we apply a small Ornstein–Uhlenbeck noise step to the observation, then

$$Y \mapsto \sqrt{1-\rho^2}\, Y + \rho W', \qquad W'_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

independent of Y. Under the planted model, this becomes

$$\sqrt{1-\rho^2}(\lambda\theta\theta^\top + W) + \rho W' = \lambda\sqrt{1-\rho^2}\,\theta\theta^\top + \widetilde{W},$$

where

$$\widetilde{W}_{ij} := \sqrt{1-\rho^2}\, W_{ij} + \rho W'_{ij} \sim \mathcal{N}(0,1).$$

Thus the model is invariant under noise injection, except that the signal-to-noise ratio is rescaled.

At the likelihood-ratio level, this means that, for the corresponding noise operator T_ρ,

$$D_{\theta,\lambda} = T_\rho D_{\theta,\frac{\lambda}{\sqrt{1-\rho^2}}}.$$

Therefore, taking the projection onto degrees strictly larger than d gives

$$D^{>d}_{\theta,\lambda} = T_\rho \left(D^{>d}_{\theta,\frac{\lambda}{\sqrt{1-\rho^2}}} \right).$$

Indeed, the degree decomposition is taken with respect to the Hermite basis of $L^2(Q)$, where Q is the standard Gaussian null distribution. The Ornstein-Uhlenbeck operator T_ρ preserves each Hermite degree and acts on the degree-r component by multiplication by ρ^r. Hence T_ρ commutes with the projections onto degree $\le d$ and degree $> d$.

Since T_ρ contracts degree-$> d$ functions by ρ^{d+1},

$$\|T_\rho f\| \le \rho^{d+1}\|f\| \qquad \text{for } f \in L^2(Q)^{>d}.$$

Hence

$$\left\| \mathop{\mathbb{E}}_{\theta} \left[\left(D^{>d}_{\theta,\lambda} \right)^{\otimes k} \right] \right\|^2 = \left\| \mathop{\mathbb{E}}_{\theta} \left[\left(T_\rho D^{>d}_{\theta,\frac{\lambda}{\sqrt{1-\rho^2}}} \right)^{\otimes k} \right] \right\|^2 \le \rho^{2(d+1)k} \left\| \mathop{\mathbb{E}}_{\theta} \left[\left(D^{>d}_{\theta,\frac{\lambda}{\sqrt{1-\rho^2}}} \right)^{\otimes k} \right] \right\|^2.$$

Thus, if the final norm at the slightly larger signal strength $\frac{\lambda}{\sqrt{1-\rho^2}}$ is merely bounded by a constant, then choosing d and k moderately large makes the high-degree contribution exponentially small.

For example, if $\rho = 0.1$ and $k = \Omega(\log n)$, then

$$\rho^{2(d+1)k}$$

is already polynomially or superpolynomially small. This explains why the high-degree assumption is mild in many Gaussian noise-robust problems.

11.8 Conclusion and transition

This completes the SQ part of the notes. The SQ framework gave a way to translate correlation bounds between planted distributions into query lower bounds. In sparse PCA, it recovered the known hard regime for SQ algorithms. In planted clique, after passing to an i.i.d. bipartite version, it recovered the square-root barrier. Together with the low-degree framework, this provides strong unconditional evidence that natural broad classes of algorithms fail below the conjectured computational thresholds.

The next part of the notes takes a different approach. Instead of proving unconditional lower bounds against a restricted algorithmic model, we study reductions. Reductions allow us to transfer hardness from one average-case problem to another. In particular, the planted clique conjecture can be used as a starting point for conditional hardness results in other high-dimensional inference problems, such as sparse PCA. This gives a complementary perspective on computational-statistical gaps: low-degree and SQ methods give unconditional evidence within restricted models, while reductions explain how hardness in one canonical problem propagates to many others.

Chapter 12

Reductions: From Planted Clique to Sparse PCA

12.1 Introduction

In the previous chapters, we developed several kinds of *unconditional* lower bounds against restricted classes of algorithms or test statistics, such as low-degree polynomials and statistical queries. We now turn to a different kind of evidence for computational hardness: *conditional* lower bounds obtained by reductions.

The guiding idea is the following. Suppose there is a detection problem $\mathcal{P}'$ which we believe is computationally hard, and suppose we can efficiently transform instances of $\mathcal{P}'$ into instances of another problem $\mathcal{P}$ in such a way that planted instances map close to planted instances and null instances map close to null instances. Then any efficient algorithm for $\mathcal{P}$ would give an efficient algorithm for $\mathcal{P}'$. Thus, assuming $\mathcal{P}'$ is hard, the problem $\mathcal{P}$ must also be hard.

The canonical source of such reductions in high-dimensional statistics is planted clique. In this chapter we explain how planted clique can be reduced to sparse PCA, as addressed in [BBH18] (see also [BR13; MW15] for two prior similar and very influential reductions to the topics discussed in this chapter). The reduction has two main ingredients: a rejection kernel, which maps Bernoulli graph entries into approximately Gaussian entries, and Gaussian cloning, which changes the sparsity and signal strength parameters.

12.2 Reductions between detection problems

First, we formally define the reduction between detection problems as follows.

Let $\mathcal{P}$ and $\mathcal{P}'$ be two detection problems. The problem $\mathcal{P}$ is to distinguish $\mathbb{P}_1$ from $\mathbb{P}_2$ on a common observation space Ω, while $\mathcal{P}'$ is to distinguish $\mathbb{P}'_1$ from $\mathbb{P}'_2$ on a common observation space Ω'.

Definition 12.1. *We say that $\mathcal{P}'$ reduces to $\mathcal{P}$ in polynomial time, and write*

$$\mathcal{P}' \leq_p \mathcal{P},$$

if there exists a polynomial-time computable map

$$\varphi : \Omega' \to \Omega$$

such that

$$\mathrm{TV}\left(\underset{x\sim\mathbb{P}_2'}{\mathrm{Law}}(\varphi(x)),\mathbb{P}_2\right)\le 0.0001 \qquad \left(\textit{i.e. } \varphi\left(\mathbb{P}_2'\right)\overset{\mathrm{TV}}{\approx}\mathbb{P}_2\right),$$

and

$$\mathrm{TV}\left(\underset{x\sim\mathbb{P}_1'}{\mathrm{Law}}(\varphi(x)),\mathbb{P}_1\right)\le 0.0001 \qquad \left(\textit{i.e. } \varphi\left(\mathbb{P}_1'\right)\overset{\mathrm{TV}}{\approx}\mathbb{P}_1\right)$$

in the asymptotic regime of the problem.

Remark 12.2. *The constant* 0.0001 *is, of course, an* arbitrarily small constant.

Remark 12.3. *Suppose* $\dim(\Omega)=N$, $\dim(\Omega')=\mathrm{poly}(N)=N'$, *then here* polynomial-time *means* $\mathrm{poly}(N)$*-time, in which case we ask the two* TV *distances in the definition to go to* 0 *as* $N\to\infty$.

We will use the following lemma to show conditional lower bounds via reduction.

Lemma 12.4. *If* $P'\le_p\mathcal{P}$ *and strong detection is possible for* $\mathcal{P}$ *in polynomial-time, then strong detection is possible for* $\mathcal{P}'$ *in polynomial-time.*

Proof. Let

$$\mathcal{A}:\Omega\to\{\mathbb{P}_1,\mathbb{P}_2\}$$

be a polynomial-time test for $\mathcal{P}$ such that

$$\underset{x\sim\mathbb{P}_2}{\mathbb{P}}\left(\mathcal{A}(x)=\mathbb{P}_1\right)+\underset{x\sim\mathbb{P}_1}{\mathbb{P}}\left(\mathcal{A}(x)=\mathbb{P}_2\right)\to 0.$$

Let $\mathcal{A}':\Omega'\to\{\mathbb{P}_1',\mathbb{P}_2'\}$ be a test for $\mathcal{P}'$ such that

$$\mathcal{A}'(x)=\begin{cases}\mathbb{P}_1' & \text{if } \mathcal{A}(\varphi(x))=\mathbb{P}_1,\\ \mathbb{P}_2' & \text{if } \mathcal{A}(\varphi(x))=\mathbb{P}_2.\end{cases}$$

Since both φ and $\mathcal{A}$ are polynomial-time computable, so is $\mathcal{A}'$. Now let's bound the Type I error. Under the null distribution $\mathbb{P}_2'$, by the definition of $\mathcal{A}'$,

$$\underset{x\sim\mathbb{P}_2'}{\mathbb{P}}\left(\mathcal{A}'(x)=\mathbb{P}_1'\right)=\underset{x\sim\mathbb{P}_2'}{\mathbb{P}}\left(\mathcal{A}(\varphi(x))=\mathbb{P}_1\right),$$

so

$$\begin{aligned}\left|\underset{x\sim\mathbb{P}_2'}{\mathbb{P}}\left(\mathcal{A}(\varphi(x))=\mathbb{P}_1\right)-\underset{x\sim\mathbb{P}_2}{\mathbb{P}}\left(\mathcal{A}(x)=\mathbb{P}_1\right)\right| &\le \mathrm{TV}\left(\underset{x\sim\mathbb{P}_2'}{\mathrm{Law}}\left(\mathcal{A}\left(\varphi\left(\mathbb{P}_2'\right)\right)\right),\mathcal{A}(\mathbb{P}_2)\right)\\ &\le \mathrm{TV}\left(\underset{x\sim\mathbb{P}_2'}{\mathrm{Law}}(\varphi(x)),\mathbb{P}_2\right)\to 0,\end{aligned}$$

where the first inequality is straightforward from the definition of TV and the second inequality is the data-processing inequality.

As the strong detection property of $\mathcal{A}$ ensures $\underset{x\sim\mathbb{P}_2}{\mathbb{P}}\left(\mathcal{A}(x)=\mathbb{P}_1\right)\to 0$, then the Type I error of $\mathcal{A}'$ also tends to zero:

$$\underset{x\sim\mathbb{P}_2'}{\mathbb{P}}\left(\mathcal{A}'(x)=\mathbb{P}_1'\right)=\underset{x\sim\mathbb{P}_2'}{\mathbb{P}}\left(\mathcal{A}(\varphi(x))=\mathbb{P}_1\right)\to 0.$$

Similarly,

$$\underset{x\sim\mathbb{P}_1'}{\mathbb{P}}\left(\mathcal{A}'(x)=\mathbb{P}_2'\right)\to 0.$$

Therefore $\mathcal{A}'$ achieves strong detection for P'. □

The following simple lemma shows that reduction (specifically the TV tending to zero variant) satisfies transitivity. It follows from simple triangle inequality.

Lemma 12.5. *If* $\mathcal{P}_1\le_p\mathcal{P}_2$ *and* $\mathcal{P}_2\le_p\mathcal{P}_3$, *then* $P_1\le_p P_3$.

12.3 Reduction from Planted Clique to Sparse PCA

Now, we know that it suffices to map instances of an assumed *hard* problem to instances of our problem of interest, as long as they are close in TV distance. The most *classic* choice of hard problem is the planted clique problem. It turns out that the planted clique problem (and its variants) can be reduced to a lot of problems, including planted independent set, sparse PCA, tensor PCA, (robust) sparse linear regression, etc. In this section, we will give a reduction from planted clique to sparse PCA.

12.3.1 Review: Planted Clique and Sparse PCA

Planted Clique

Recall that in the detection version of the planted clique problem we are asked to distinguish between the following two distributions:

$$\mathbb{P}_1 : \mathcal{G}_{n,1/2} \cup \text{Clique}(S), \qquad S \sim \text{Uniform}\left(\binom{[n]}{k}\right);$$
$$\mathbb{P}_2 : \mathcal{G}_{n,1/2}.$$

Thus, the adjacency matrix of the planted model contains a hidden $k \times k$ submatrix with only 1 entries except for the diagonal.

We know the phase diagram (ignoring logarithmic terms) is as shown in Figure 12.1.

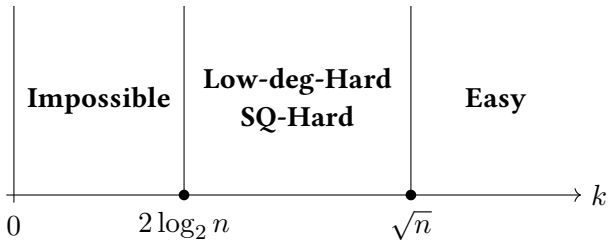

Figure 12.1: Planted clique phase diagram for detection

Now we'll see that when $k \gg \sqrt{n \log n} = \widetilde{\mathcal{O}}(\sqrt{n})$, the following polynomial algorithm solves the problem:

Algorithm 12.6 Top-degree detection for planted clique

1: For every vertex $v \in [n]$, compute its degree $d(v) := \deg_G(v)$.
2: Let $\widehat{S}$ be the set of the k vertices with largest degree in G.
3: **if** $\widehat{S}$ is a clique in G **then**
4: **return** $\mathbb{P}_1$.
5: **else**
6: **return** $\mathbb{P}_2$.
7: **end if**

To prove the algorithm above achieves strong detection we need the following result:

Lemma 12.7. *Let* $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bin}\left(n, \frac{1}{2}\right)$. *Then*

$$\max_{1 \le i \le n} \left| X_i - \frac{n}{2} \right| = \mathcal{O}\left(\sqrt{n \log n}\right)$$

with high probability.

Proof. By Hoeffding's inequality, for every $t > 0$,

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

Taking a union bound over $i = 1, \ldots, n$, we get

$$\mathbb{P}\left(\max_{1\leq i\leq n} |X_i - \mathbb{E}[X_i]| \geq t\right) \leq 2n\exp\left(-\frac{2t^2}{n}\right).$$

Choose

$$t = C\sqrt{n\log n}.$$

Then

$$2n\exp\left(-\frac{2t^2}{n}\right) = 2n\exp(-2C^2\log n) = 2n^{1-2C^2}.$$

For $C > \frac{1}{\sqrt{2}}$, this tends to zero, and the result follows. □

By the lemma, under the null model all degrees satisfy

$$\deg(v) = \frac{n}{2} + \mathcal{O}\left(\sqrt{n\log n}\right)$$

uniformly over all vertices, with high probability.

Under the planted model, if $v \notin S$, then again

$$\deg(v) = \frac{n}{2} + \mathcal{O}\left(\sqrt{n\log n}\right)$$

with high probability, whereas if $v \in S$, then

$$\deg(v) = k - 1 + \mathrm{Bin}\left(n-k, \frac{1}{2}\right) = \frac{n}{2} + \frac{k}{2} + \mathcal{O}\left(\sqrt{n\log n}\right)$$

with high probability, uniformly over planted vertices.

Thus the degree gap between planted and non-planted vertices is of order k, while the maximal random fluctuation is of order $\sqrt{n\log n}$. Therefore, if

$$k \gg \sqrt{n\log n},$$

then with high probability every planted vertex has larger degree than every non-planted vertex. Consequently, selecting the k vertices of largest degree recovers the planted clique, and checking whether those vertices form a clique gives a polynomial-time strong detector.

Sparse PCA

Now recall that in the detection version of the normalized sparse PCA problem we are asked to distinguish between the following two distributions:

$$\mathbb{P}_1 : Y = \lambda xx^\top + W, \qquad W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),\ x \in \{0,1\}^n,\ \|x\|_0 = k = n^\alpha,\ \alpha \in (0,1),\ \lambda = \lambda_n > 0,$$
$$\mathbb{P}_2 : Y = W.$$

Thus the planted model contains a hidden $k \times k$ submatrix with entries shifted by λ, which reminds the adjacency matrix of the planted clique problem under the planted model.

The phase diagram (again ignoring logarithmic terms) is as shown in Figure 12.2.

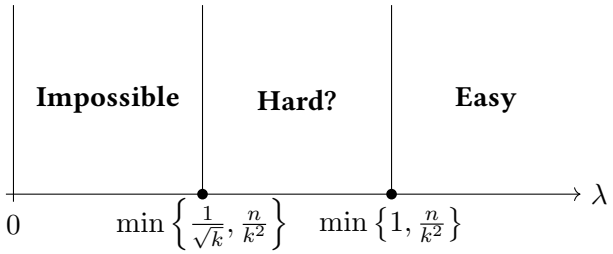

Figure 12.2: Sparse PCA phase diagram for detection

There are two elementary polynomial-time algorithms at the upper threshold.

First, if $\lambda \gg \sqrt{\log n} = \widetilde{\mathcal{O}}(1)$, one can consider the following algorithm:

Algorithm 12.8 Diagonal-thresholding detection for sparse PCA

1: Compute the largest diagonal entry $D(Y) := \max_{i \in [n]} Y_{ii}$.
2: **if** $D(Y) \geq \lambda/2$ **then**
3: **return** $\mathbb{P}_1$.
4: **else**
5: **return** $\mathbb{P}_2$.
6: **end if**

Under $\mathbb{P}_1$, for $i \in \mathrm{Supp}(x) := \{j : x_j = 1\}$, we have

$$Y_{ii} = \lambda + W_{ii}.$$

Using Lemma 6.1, we get

$$\max_i Y_{ii} \geq \lambda + \mathcal{O}\left(\sqrt{\log n}\right)$$

with high probability. Under $\mathbb{P}_2$, using Lemma 6.1 again, we have

$$\max_i W_{ii} = \mathcal{O}\left(\sqrt{\log n}\right)$$

with high probability. Therefore, it follows that the diagonal test succeeds when $\lambda \gg \sqrt{\log n} = \widetilde{\mathcal{O}}(1)$.

Second, if $\lambda \gg \dfrac{n}{k^2}$, one can consider the statistic below:

Algorithm 12.9 Global-sum detection for sparse PCA

1: Compute the global sum statistic $T(Y) := \sum_{i=1}^n \sum_{j=1}^n Y_{ij}$.
2: **if** $T(Y) \geq \lambda k^2/2$ **then**
3: **return** $\mathbb{P}_1$.
4: **else**
5: **return** $\mathbb{P}_2$.
6: **end if**

Under $\mathbb{P}_1$, considering $\sum_{i,j} W_{ij} \sim \mathcal{N}\left(0, n^2\right)$ and so $\sum_{i,j} W_{ij} = \mathcal{O}(n)$,

$$\sum_{i,j} Y_{ij} = \lambda k^2 + \sum_{i,j} W_{ij} = \lambda k^2 + \mathcal{O}(n),$$

whereas under $\mathbb{P}_2$, using the same reasoning,

$$\sum_{i,j} Y_{ij} = \mathcal{O}(n).$$

Thus the sum test succeeds when $\lambda k^2 \gg n$, that is, $\lambda \gg \frac{n}{k^2}$.

Combining the two tests gives polynomial-time detection whenever $\lambda \gg \min\left\{1, \frac{n}{k^2}\right\}$, up to logarithmic factors. Interestingly, no better polynomial-time algorithm is known in the regimes considered here.

One can prove a low-degree lower bound for sparse PCA. Yet, the point of this section is that one can also prove a reduction: if one finds a polynomial-time algorithm for sparse PCA for any $k = n^\alpha$, $\alpha < 2/3$ and $\lambda \ll \min\left\{1, \frac{n}{k^2}\right\}$, then one can also solve the planted clique problem for some clique size $k \ll \sqrt{n}$. We will need two techniques to do the reduction: *rejection kernel* and *Gaussian cloning*.

12.3.2 Rejection Kernel: the case $k < \sqrt{n}$

We first consider the regime where the sparse PCA sparsity is below the square-root scale. In this regime,

$$k < \sqrt{n} \qquad \Longrightarrow \qquad \min\left\{1, \frac{n}{k^2}\right\} = 1.$$

Thus the conjectured algorithmic threshold is $\lambda \asymp 1$, up to logarithmic factors. We will reduce planted clique to sparse PCA with a small constant, or slightly subconstant, value of λ.

Consider the adjacency matrix A (symmetric) of the input graph in the planted clique problem. Under $\mathbb{P}_1$, A has a $k \times k$ block of 1's (except for the diagonal), and the rest off-diagonal entries are i.i.d. $\mathrm{Bern}(1/2)$, and under $\mathbb{P}_2$, all entries are i.i.d. $\mathrm{Bern}(1/2)$.

Meanwhile, in the sparse PCA problem, the input is a real-valued matrix Y. Under $\mathbb{P}_1$, Y has a $k \times k$ block of i.i.d. $\mathcal{N}(\lambda, 1)$ entries and the rest are i.i.d. $\mathcal{N}(0, 1)$ entries, and under $\mathbb{P}_2$, all entries are i.i.d. $\mathcal{N}(0, 1)$. Thus, one idea is to construct a (randomized) polynomial-time computable map

$$\varphi : \{0, 1\} \to \mathbb{R} \quad \text{such that} \quad \varphi(1) \approx \mathcal{N}(\lambda, 1), \qquad \varphi(\mathrm{Bern}(1/2)) \approx \mathcal{N}(0, 1).$$

With this φ, we could map entry-wise the adjacency matrix from the planted clique problem to the input matrix of sparse PCA.

Suppose first that

$$\varphi(1) \sim \mathcal{N}(\lambda, 1)$$

and

$$\varphi(0) \sim f$$

for some density f. If $B \sim \mathrm{Bern}(1/2)$, then requiring $\varphi(B) \sim \mathcal{N}(0, 1)$ gives, at the level of densities,

$$\frac{1}{2}\phi_\lambda + \frac{1}{2} f = \phi_0,$$

where

$$\phi_a(x) := \frac{1}{\sqrt{2\pi}} e^{-(x-a)^2/2}.$$

Hence, formally, it must hold that

$$f = 2\phi_0 - \phi_\lambda.$$

The problem is that this function is not nonnegative everywhere. Yet, notice that the set on which it is nonnegative is

$$S := \{x : 2\phi_0(x) - \phi_\lambda(x) \geq 0\}.$$

Solving the inequality,

$$2\phi_0(x) \geq \phi_\lambda(x) \iff \log 2 - \frac{x^2}{2} \geq -\frac{(x-\lambda)^2}{2} \iff x \leq \frac{\lambda}{2} + \frac{\log 2}{\lambda}.$$

Thus

$$S = \left\{x : x \leq \frac{\lambda}{2} + \frac{\log 2}{\lambda}\right\}.$$

The rejection kernel keeps the positive part of $2\phi_0 - \phi_\lambda$ and renormalizes it. Define

$$Z := \int_S (2\phi_0(x) - \phi_\lambda(x))\,\mathrm{d}x,$$

and

$$\widetilde{f}(x) := \frac{(2\phi_0(x) - \phi_\lambda(x))\mathbf{1}_S(x)}{Z}.$$

Sampling from $\widetilde{f}$ can be done efficiently by rejection sampling, since $2\phi_0$ is an envelope for the unnormalized density on S. This is the key idea that makes this construction possible.

To formally define and prove this, we first describe it in a slightly more general way for any $p \in (0,1)$ instead $p = 1/2$. Define

$$S_p := \{x \in \mathbb{R} : \phi_0(x) \geq p\phi_\lambda(x)\}.$$

Since

$$\frac{\phi_\lambda(x)}{\phi_0(x)} = \exp\left(\lambda x - \frac{\lambda^2}{2}\right),$$

we have the explicit description

$$S_p = \left(-\infty,\ \frac{\lambda}{2} + \frac{\log(1/p)}{\lambda}\right].$$

Let

$$Z_p := \int_{S_p} \frac{\phi_0(x) - p\phi_\lambda(x)}{1-p}\,\mathrm{d}x$$

and define

$$\widetilde{f}_p(x) := \frac{1}{Z_p}\frac{\phi_0(x) - p\phi_\lambda(x)}{1-p}\mathbf{1}\{x \in S_p\}.$$

Lemma 12.10 (Rejection kernel for a Bernoulli background)**.** *Let $X \sim \mathrm{Bern}(p)$ and define the randomized kernel RK_p by*

$$\mathrm{RK}_p(1) \sim \mathcal{N}(\lambda, 1), \qquad \mathrm{RK}_p(0) \sim \widetilde{f}_p.$$

Set

$$a_p := \mathbb{P}_{G\sim\mathcal{N}(0,1)}(G \notin S_p), \qquad b_p := \mathbb{P}_{G\sim\mathcal{N}(\lambda,1)}(G \notin S_p).$$

Then

$$\mathrm{TV}\left(\mathrm{Law}\big(\mathrm{RK}_p(X)\big), \mathcal{N}(0,1)\right) = pb_p - a_p \leq pb_p.$$

In particular,

$$\mathrm{TV}\left(\mathrm{Law}\big(\mathrm{RK}_p(X)\big), \mathcal{N}(0,1)\right) \leq p\,\mathbb{P}_{G\sim\mathcal{N}(\lambda,1)}(G \notin S_p).$$

Proof. By the definition of S_p, the numerator

$$\phi_0(x) - p\phi_\lambda(x)$$

is nonnegative on S_p. Thus it remains only to verify that $Z_p > 0$ and to estimate the resulting mixture. By definition,

$$a_p = \int_{S_p^c} \phi_0(x)\,\mathrm{d}x, \qquad b_p = \int_{S_p^c} \phi_\lambda(x)\,\mathrm{d}x.$$

On S_p^c, we have

$$\phi_0(x) < p\phi_\lambda(x).$$

Integrating this inequality over S_p^c gives

$$a_p \le pb_p.$$

Furthermore,

$$\begin{aligned} Z_p &= \frac{1}{1-p}\int_{S_p} \big(\phi_0(x) - p\phi_\lambda(x)\big)\,\mathrm{d}x \\ &= \frac{(1-a_p) - p(1-b_p)}{1-p} \\ &= 1 + \frac{pb_p - a_p}{1-p}. \end{aligned}$$

Consequently,

$$Z_p \ge 1,$$

so $\widetilde{f}_p$ is a well-defined probability density.

Let m_p denote the density of $\mathrm{RK}_p(X)$. Since $X \sim \mathrm{Bern}(p)$,

$$m_p(x) = p\phi_\lambda(x) + (1-p)\widetilde{f}_p(x).$$

We compute its total variation distance from ϕ_0 separately on S_p and S_p^c.

For $x \in S_p$,

$$m_p(x) = p\phi_\lambda(x) + \frac{\phi_0(x) - p\phi_\lambda(x)}{Z_p},$$

and hence

$$\begin{aligned} m_p(x) - \phi_0(x) &= p\phi_\lambda(x) + \frac{\phi_0(x) - p\phi_\lambda(x)}{Z_p} - \phi_0(x) \\ &= \left(1 - \frac{1}{Z_p}\right)\big(p\phi_\lambda(x) - \phi_0(x)\big). \end{aligned}$$

Because $Z_p \ge 1$ and $p\phi_\lambda(x) \le \phi_0(x)$ on S_p, this difference is nonpositive. Therefore

$$\begin{aligned} \int_{S_p} |m_p(x) - \phi_0(x)|\,\mathrm{d}x &= \left(1 - \frac{1}{Z_p}\right)\int_{S_p} \big(\phi_0(x) - p\phi_\lambda(x)\big)\,\mathrm{d}x \\ &= \left(1 - \frac{1}{Z_p}\right)(1-p)Z_p \\ &= (1-p)(Z_p - 1) \\ &= pb_p - a_p. \end{aligned}$$

For $x \in S_p^c$, the density $\widetilde{f}_p$ vanishes, so

$$m_p(x) = p\phi_\lambda(x).$$

Moreover, $p\phi_\lambda(x) > \phi_0(x)$ on S_p^c. Thus

$$\begin{aligned}\int_{S_p^c} |m_p(x) - \phi_0(x)| \, \mathrm{d}x &= \int_{S_p^c} \big(p\phi_\lambda(x) - \phi_0(x)\big) \, \mathrm{d}x \\ &= pb_p - a_p.\end{aligned}$$

Combining the two regions,

$$\begin{aligned}\mathrm{TV}\left(\mathrm{Law}\big(\mathrm{RK}_p(X)\big), \mathcal{N}(0,1)\right) &= \frac{1}{2}\int_{\mathbb{R}} |m_p(x) - \phi_0(x)| \, \mathrm{d}x \\ &= \frac{1}{2}\left((pb_p - a_p) + (pb_p - a_p)\right) \\ &= pb_p - a_p \\ &\le pb_p.\end{aligned}$$

□

Remark 12.11 (Gaussian-tail estimate)**.** *Since*

$$S_p^c = \left(\frac{\lambda}{2} + \frac{\log(1/p)}{\lambda}, \infty\right),$$

we have, for $G \sim \mathcal{N}(0,1)$*,*

$$\begin{aligned}b_p &= \mathbb{P}\left(\lambda + G > \frac{\lambda}{2} + \frac{\log(1/p)}{\lambda}\right) \\ &= \mathbb{P}\left(G > \frac{\log(1/p)}{\lambda} - \frac{\lambda}{2}\right).\end{aligned}$$

If

$$\lambda^2 \le \log(1/p),$$

then

$$\frac{\log(1/p)}{\lambda} - \frac{\lambda}{2} \ge \frac{\log(1/p)}{2\lambda}.$$

Using the standard bound $\mathbb{P}(G \ge t) \le e^{-t^2/2}$*, we obtain*

$$b_p \le \exp\left(-\frac{\log^2(1/p)}{8\lambda^2}\right).$$

Consequently,

$$\mathrm{TV}\left(\mathrm{Law}\big(\mathrm{RK}_p(X)\big), \mathcal{N}(0,1)\right) \le p\exp\left(-\frac{\log^2(1/p)}{8\lambda^2}\right).$$

Thus, for every fixed $p \in (0,1)$*, the error is* $\exp(-\Omega_p(\lambda^{-2}))$ *as* $\lambda \to 0$*.*

Remark 12.12 (Sampling from $\widetilde{f}_p$)**.** *The density* $\widetilde{f}_p$ *can be sampled by rejection sampling. Repeatedly draw*

$$Y \sim \mathcal{N}(0,1), \qquad U \sim \mathrm{Unif}[0,1],$$

independently, and accept Y if

$$Y \in S_p \quad \text{and} \quad U \le 1 - p\exp\left(\lambda Y - \frac{\lambda^2}{2}\right).$$

Indeed, the unnormalized density of an accepted proposal is

$$\phi_0(y)\mathbf{1}\{y \in S_p\}\left(1 - p\exp\left(\lambda y - \frac{\lambda^2}{2}\right)\right) = \big(\phi_0(y) - p\phi_\lambda(y)\big)\mathbf{1}\{y \in S_p\},$$

which is proportional to $\widetilde{f}_p$. The acceptance probability is

$$\int_{S_p}\big(\phi_0(y) - p\phi_\lambda(y)\big)\,\mathrm{d}y = (1-p)Z_p \ge 1 - p.$$

Hence, when p is a fixed constant strictly smaller than one, the expected number of proposals is bounded by $(1-p)^{-1}$.

Remark 12.13. *(Almost there!) Naively using the rejection kernel per entry allow us to do the reduction from planted clique with $k = o(\sqrt{n})$ to a slightly **modified** sparse PCA model (with zero-out diagonal and symmetric entries) when*

$$\lambda \le \frac{0.01}{\sqrt{\log n}}$$

and $k = o(\sqrt{n})$. Indeed,

$$\mathop{\mathbb{P}}_{z\sim\mathcal{N}(\lambda,1)}(z \notin S) \le \exp\left(-\frac{\log^2 2}{8\lambda^2}\right) = \mathcal{O}\left(n^{-3}\right).$$

Then, by applying RK *entrywise symmetrically on A (i.e. $Y_{ij} = Y_{ji} = \mathrm{RK}(A_{ij})$), and by a coupling argument, we have based on two remarks above*

$$\mathrm{TV}(\mathrm{Law}(\mathrm{RK}(A)), \mathrm{Law}(Y)) \le n^2 \cdot \mathcal{O}\left(n^{-3}\right) \to 0,$$

where A, Y are from the planted/null distributions of the two problems. Therefore, the rejection kernel gives a polynomial-time reduction from k-planted clique to this modified (k,λ)*-sparse PCA, when $\lambda \le \frac{0.01}{\sqrt{\log n}} = \widetilde{\mathcal{O}}(1)$. Since, for $k = o(\sqrt{n})$, the conjectured polynomial-time sparse PCA threshold is $\lambda \asymp 1$ up to logarithmic factors, this proves the desired reduction but to modified sparse PCA in the sub-square-root sparsity regime.*

Completing the reduction for $k = o(\sqrt{n})$

We now need to address the issue that in sparse PCA the observation matrix has non-zero diagonal and is not symmetric.

We now create two independent directed copies of the graph and then embed them as the two triangular halves of a random principal minor. This follows by a combination of two nice tricks: graph-cloning [BBH19, Lemma 6.2] and diagonal-planting construction of [BBH19, Lemma 6.4].

Set

$$N := 2n, \qquad Q := 2^{-1/2}.$$

Let $\mathsf{BerNull}_{N,Q}$ denote the law of an $N \times N$ matrix with i.i.d. $\mathrm{Bern}(Q)$ entries. Let $\mathsf{BerPlant}_{N,k,Q}$ denote the following planted law: choose $S^\star \subseteq [N]$ uniformly with $|S^\star| = k$, set

$$M_{ab} = 1 \qquad \text{for every } a, b \in S^\star,$$

including when $a = b$, and sample every remaining entry independently from $\mathrm{Bern}(Q)$.

The following key proposition resolves both the symmetry and the diagonal mismatch between the two models.

Proposition 12.14 (Cloning and planting the diagonal)**.** *There is a randomized polynomial-time map*

$$\Psi : \mathcal{G}_n \longrightarrow \{0,1\}^{N\times N}$$

and constants $c, C > 0$ such that, if $G \sim G(n, 1/2)$, then

$$d_{\mathrm{TV}}\left(\mathcal{L}\big(\Psi(G)\big), \mathsf{BerNull}_{N,Q}\right) \le e^{-cN}.$$

If G is drawn from the planted-clique model with a uniformly random planted clique of size k, then

$$d_{\mathrm{TV}}\left(\mathcal{L}\big(\Psi(G)\big), \mathsf{BerPlant}_{N,k,Q}\right) \le C\frac{k}{\sqrt{N}} + e^{-cN}.$$

Proof. Let A_{ij} be the adjacency indicator of the unordered pair $\{i, j\}$, for $1 \le i < j \le n$. Independently for every such pair, generate two bits

$$\big(B^+_{ij}, B^-_{ij}\big) \in \{0,1\}^2$$

as follows. If $A_{ij} = 1$, set

$$\big(B^+_{ij}, B^-_{ij}\big) = (1,1).$$

If $A_{ij} = 0$, set

$$\begin{aligned}
\mathbb{P}\big((B^+_{ij}, B^-_{ij}) = (1,0) \mid A_{ij} = 0\big) &= \sqrt{2} - 1,\\
\mathbb{P}\big((B^+_{ij}, B^-_{ij}) = (0,1) \mid A_{ij} = 0\big) &= \sqrt{2} - 1,\\
\mathbb{P}\big((B^+_{ij}, B^-_{ij}) = (0,0) \mid A_{ij} = 0\big) &= 3 - 2\sqrt{2}.
\end{aligned}$$

These probabilities are nonnegative and sum to one.

If $A_{ij} \sim \mathrm{Bern}(1/2)$, then

$$\big(B^+_{ij}, B^-_{ij}\big) \sim \mathrm{Bern}(Q)^{\otimes 2}$$

exactly, because

$$Q^2 = \frac{1}{2}, \qquad Q(1-Q) = \frac{\sqrt{2}-1}{2}, \qquad (1-Q)^2 = \frac{3-2\sqrt{2}}{2}.$$

On the other hand, every planted edge is mapped to $(1, 1)$. Thus the two directed copies have independent $\mathrm{Bern}(Q)$ background entries and the same planted clique. This is the $p = 1$, $q = 1/2$, $t = 2$ specialization of the graph-cloning lemma in [BBH19, Lemma 6.2].

Choose a set $V \subseteq [N]$ uniformly with $|V| = n$, and independently choose a uniformly random bijection

$$\pi : V \longrightarrow [n].$$

For distinct $a, b \in [N]$, define

$$M_{ab} := \begin{cases} B^+_{\min\{\pi(a),\pi(b)\},\, \max\{\pi(a),\pi(b)\}}, & a, b \in V,\ a < b,\\ B^-_{\min\{\pi(a),\pi(b)\},\, \max\{\pi(a),\pi(b)\}}, & a, b \in V,\ a > b,\\ \text{an independent } \mathrm{Bern}(Q) \text{ variable}, & \{a, b\} \not\subseteq V. \end{cases}$$

Consequently, the upper- and lower-triangular entries are independent.

It remains to generate the diagonal. Independently sample

$$L \sim \mathrm{Bin}(N, Q).$$

If $L \geq n$, choose $T \subseteq [N] \setminus V$ uniformly with $|T| = L - n$. If $L < n$, set $T = \varnothing$. Finally, set

$$M_{aa} := \mathbf{1}\{a \in V \cup T\}, \qquad a \in [N].$$

Under the null, conditional on $L = \ell \geq n$, the set $V \cup T$ is a uniformly random ℓ-subset of $[N]$. This is exactly the conditional distribution of the support of N i.i.d. $\mathrm{Bern}(Q)$ variables given that their sum is ℓ. Moreover, conditional on V, all off-diagonal entries are i.i.d. $\mathrm{Bern}(Q)$, and their law does not depend on V. Since $N = 2n$ and $Q > 1/2$, a Chernoff bound gives

$$\mathbb{P}(L < n) \leq e^{-cN}.$$

This proves the null bound.

Now suppose that G contains a planted clique $S \subseteq [n]$, and define its image

$$S^\star := \pi^{-1}(S) \subseteq V.$$

By symmetry, $S^\star$ is a uniformly random k-subset of $[N]$. Conditional on $S^\star$, all off-diagonal entries in $S^\star \times S^\star$ equal one, and every other off-diagonal entry is an independent $\mathrm{Bern}(Q)$ variable.

Conditional on $S^\star$ and $L = \ell \geq n$, the diagonal-one set $V \cup T$ is uniformly distributed among all ℓ-subsets of $[N]$ containing $S^\star$. In the target planted model, the diagonal-one set has the same conditional distribution given its size, but its size is

$$L_\star := k + \mathrm{Bin}(N - k, Q)$$

rather than $L \sim \mathrm{Bin}(N, Q)$. Therefore

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathcal{L}(\mathrm{diag} M \mid S^\star), \mathcal{L}(\mathrm{diag} M_\star \mid S^\star)\right) \\ \leq \mathbb{P}(L < n) + d_{\mathrm{TV}}\big(\mathcal{L}(L), \mathcal{L}(L_\star)\big). \end{aligned}$$

To bound the second term, write

$$X \sim \mathrm{Bin}(N - k, Q), \qquad Z \sim \mathrm{Bin}(k, Q),$$

independently. Then

$$L \overset{d}{=} X + Z, \qquad L_\star \overset{d}{=} X + k.$$

By convexity of total variation,

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathcal{L}(L), \mathcal{L}(L_\star)\big) &\leq \mathbb{E}_Z\left[d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(X + k - Z)\big)\right] \\ &\leq \mathbb{E}[k - Z] \max_j \mathbb{P}(X = j) \\ &\leq C_Q \frac{k}{\sqrt{N}}. \end{aligned}$$

Here we used the elementary bound

$$d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(X + r)\big) \leq r \max_j \mathbb{P}(X = j)$$

for a unimodal integer-valued random variable. Together with $\mathbb{P}(L < n) \leq e^{-cN}$, this proves the planted bound. □

We now can easily complete the reduction by simply "Gaussianizing" the entries independently via the Rejection Kernel.

Corollary 12.15 (Reduction to the i.i.d. Gaussian sparse-PCA model)**.** *There are constants $c_0, c_1, C > 0$ such that, for every*

$$0 < \lambda \le \frac{c_0}{\sqrt{\log N}},$$

there is a randomized polynomial-time map

$$\Phi : \mathcal{G}_n \longrightarrow \mathbb{R}^{N \times N}$$

such that, under the null,

$$d_{\mathrm{TV}}\left(\mathcal{L}\big(\Phi(G)\big), \mathcal{N}(0,1)^{\otimes N^2}\right) \le CN^{-1} + e^{-c_1 N}.$$

Under the planted model,

$$d_{\mathrm{TV}}\left(\mathcal{L}\big(\Phi(G)\big), \mathcal{L}\left(\lambda \mathbf{1}_{S^\star}\mathbf{1}_{S^\star}^\top + W\right)\right) \le C\left(\frac{k}{\sqrt{N}} + N^{-1}\right) + e^{-c_1 N},$$

where $S^\star$ is uniform over the k-subsets of $[N]$ and, independently, the entries of W are i.i.d. $\mathcal{N}(0,1)$.

Proof. For fixed $Q = 2^{-1/2}$, the Gaussian rejection-kernel lemma above, Lemma 12.10 above, gives a randomized polynomial-time map

$$\mathrm{RK}_\lambda : \{0,1\} \longrightarrow \mathbb{R}$$

such that

$$d_{\mathrm{TV}}\left(\mathcal{L}(\mathrm{RK}_\lambda(1)), \mathcal{N}(\lambda, 1)\right) \le CN^{-3}$$

and

$$d_{\mathrm{TV}}\left(\mathcal{L}(\mathrm{RK}_\lambda(B)), \mathcal{N}(0,1)\right) \le CN^{-3}, \qquad B \sim \mathrm{Bern}(Q),$$

provided $\lambda \le c_0/\sqrt{\log N}$.

Apply this kernel independently to all N^2 entries of the matrix produced by Proposition 12.14. By data processing, the error in Proposition 12.14 does not increase. Conditional on the planted support, tensorization of total variation contributes at most

$$CN^2 \cdot N^{-3} = CN^{-1}.$$

The two stated bounds follow. □

12.3.3 Gaussian Cloning: the case $k > \sqrt{n}$

Remember that the hard regime we want to show for sparse PCA, via a reduction from planted clique, is $\lambda \le \min\left\{1, \frac{n}{k^2}\right\}$, so given the above it remains to do the proof for $k > \sqrt{n}$. We do this interestingly by reducing sparse PCA for $k > \sqrt{n}$ to instances of sparse PCA with $k < \sqrt{n}$ but smaller lambda. By transitivity and the previous reduction, this suffices.

Specifically, we will give the reduction

$$\left(n^{\alpha_1}, \widetilde{\mathcal{O}}(1)\right) - \text{sparse PCA} \le_p \left(n^{\alpha_2}, \lambda\right) - \text{sparse PCA},$$

for

$$\alpha_1 < 1/2 < \alpha_2, \quad \lambda \asymp n^{-\frac{\alpha_2 - \alpha_1}{1-\alpha_1}}.$$

Note that then

$$\alpha_1 = \frac{1}{2} - o(1) \quad \Longrightarrow \quad \lambda \asymp n^{1-2\alpha_2+o(1)} \approx \frac{n}{k^2}.$$

To prove this, we use Gaussian cloning. The idea is to transform each Gaussian entry into several independent Gaussian entries with a smaller mean. This increases both the ambient dimension and the sparsity, while decreasing the signal strength in a controlled way.

Lemma 12.16 (Gaussian cloning)**.** *For $d \geq 2$, there exists a polynomial-time computable map*

$$\varphi : \mathbb{R} \to \mathbb{R}^d$$

such that, for all $\lambda \in \mathbb{R}$,

$$X \sim \mathcal{N}(\lambda, 1) \qquad \Longrightarrow \qquad \varphi(X) \sim \mathcal{N}\left(\frac{\lambda}{\sqrt{d}}\mathbf{1}_d, I_d\right),$$

where $\mathbf{1}_d$ denotes the all one vector in $\mathbb{R}^d$. In particular, if $X \sim \mathcal{N}(0,1)$, then $\varphi(X) \sim \mathcal{N}(0, I_d)$.

Proof. Let U be an orthogonal $d \times d$ matrix whose first column is $\frac{1}{\sqrt{d}}\mathbf{1}_d$. Given $X \in \mathbb{R}$, sample independent random variables

$$Z_2, \ldots, Z_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

and define

$$Z := (X, Z_2, \ldots, Z_d) \in \mathbb{R}^d.$$

Set

$$\varphi(X) := UZ.$$

If $X \sim \mathcal{N}(\lambda, 1)$, then

$$Z \sim \mathcal{N}(\lambda e_1, I_d).$$

Since U is orthogonal,

$$UZ \sim \mathcal{N}(\lambda U e_1, I_d).$$

By construction,

$$Ue_1 = \frac{1}{\sqrt{d}}\mathbf{1}_d,$$

so

$$\varphi(X) \sim \mathcal{N}\left(\frac{\lambda}{\sqrt{d}}\mathbf{1}_d, I_d\right).$$

□

Let Y be the input matrix of (k, λ)-sparse PCA. If we apply φ for $d = 4$ entrywise, and arrange each output in a 2×2 block, then we get a $2n \times 2n$ matrix (see Figure 12.3). If Y is from the planted distribution, then $\varphi(Y)$ will contain a $2k \times 2k$ block of i.i.d. $\mathcal{N}(\lambda/2, 1)$ entries and the rest are i.i.d. $\mathcal{N}(0,1)$. And if Y is from the null distribution, then all the entries of Y are i.i.d. $\mathcal{N}(0,1)$. Now we have a sparse PCA instance with $n' = 2n$, $k' = 2k$, $\lambda' = \lambda/2$.

Iterating r times gives

$$n_r = 2^r n, \qquad k_r = 2^r k, \qquad \lambda_r = 2^{-r}\lambda.$$

Now suppose we start with a sparse PCA instance produced by the rejection-kernel reduction, with sparsity

$$k = n^{\alpha_1}, \qquad \alpha_1 < \frac{1}{2},$$

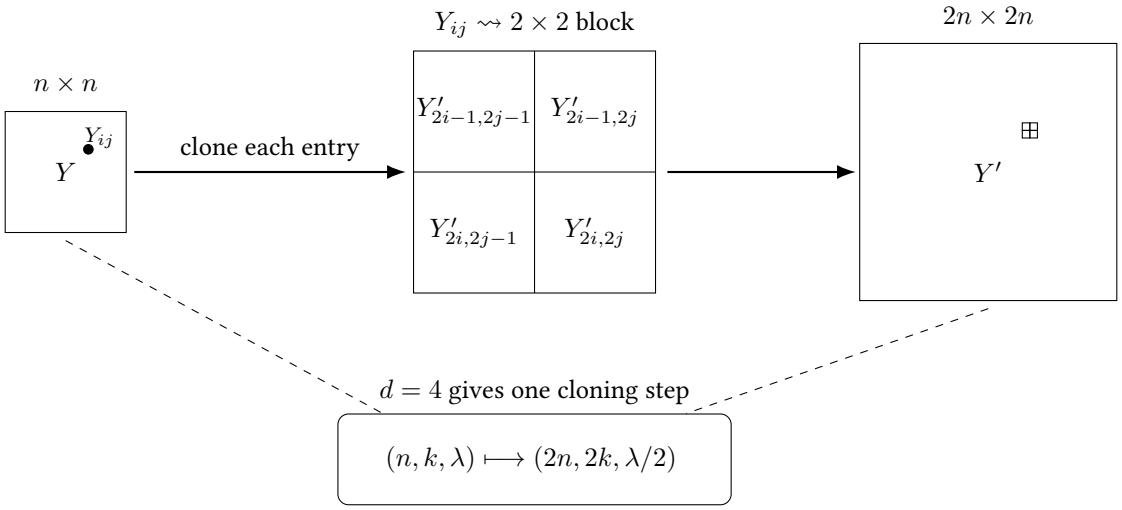

Figure 12.3: One step of Gaussian cloning. A scalar Gaussian variable is transformed into d independent Gaussian variables with mean divided by $\sqrt{d}$. In the reduction, we take $d = 4$, so each matrix entry is replaced by a 2×2 block. Therefore an $n \times n$ matrix becomes a $2n \times 2n$ matrix, the planted support size doubles, and the signal strength is divided by 2.

and signal strength $\lambda = \widetilde{\mathcal{O}}(1)$. We want to obtain an instance with sparsity exponent

$$\alpha_2 > \frac{1}{2}.$$

Choose

$$r = \varepsilon \log_2 n \quad \Longrightarrow \quad 2^r = n^\varepsilon,$$

and after r cloning steps,

$$n' = n^{1+\varepsilon}, \qquad k' = n^{\alpha_1+\varepsilon}, \qquad \lambda' = n^{-\varepsilon}\lambda.$$

We want

$$k' = \left(n'\right)^{\alpha_2} \quad \Longrightarrow \quad k' = \left(n^{1+\varepsilon}\right)^{\alpha_2} = n^{\alpha_2(1+\varepsilon)},$$

so we require

$$\alpha_1 + \varepsilon = \alpha_2(1+\varepsilon) \quad \Longrightarrow \quad \varepsilon = \frac{\alpha_2 - \alpha_1}{1-\alpha_2} \quad \Longrightarrow \quad k' = \left(n'\right)^{\alpha_2}, \quad \lambda' = n^{-\varepsilon}\lambda = \left(n'\right)^{-\varepsilon/(1+\varepsilon)}\lambda.$$

Since

$$\frac{\varepsilon}{1+\varepsilon} = \frac{\alpha_2 - \alpha_1}{1-\alpha_1} \quad \Longrightarrow \quad \lambda' = \left(n'\right)^{-(\alpha_2-\alpha_1)/(1-\alpha_1)}\widetilde{\mathcal{O}}(1).$$

If we take

$$\alpha_1 = \frac{1}{2} - o(1) \quad \Longrightarrow \quad \frac{\alpha_2 - \alpha_1}{1-\alpha_1} = 2\alpha_2 - 1 + o(1) \quad \Longrightarrow \quad \lambda' = \left(n'\right)^{1-2\alpha_2+o(1)}.$$

Since $k' = (n')^{\alpha_2}$,

$$\frac{n'}{(k')^2} = \left(n'\right)^{1-2\alpha_2}.$$

Thus cloning transforms the sub-square-root sparse PCA instance into a super-square-root sparse PCA instance at signal strength

$$\lambda' \approx \frac{n'}{(k')^2},$$

up to subpolynomial factors. This completes the reduction to the conjectural hard threshold

$$\lambda \approx \min\left\{1, \frac{n}{k^2}\right\}.$$

Remark 12.17. *The reason we take*

$$\alpha_1 = \frac{1}{2} - o(1)$$

is that the sharp sparse PCA threshold in the super-square-root regime is recovered only if the starting sparsity exponent is arbitrarily close to the square-root scale.

Indeed, after Gaussian cloning, the signal strength becomes

$$\lambda' = \left(n'\right)^{-(\alpha_2-\alpha_1)/(1-\alpha_1)} \widetilde{\mathcal{O}}(1).$$

On the other hand, the conjectural sparse PCA algorithmic threshold in the target regime is

$$\frac{n'}{\left(k'\right)^2} = \left(n'\right)^{1-2\alpha_2} = \left(n'\right)^{-(2\alpha_2-1)}.$$

Thus, to match the exponent of the conjectural threshold, we want

$$\frac{\alpha_2 - \alpha_1}{1-\alpha_1} \approx 2\alpha_2 - 1.$$

This becomes exact in the limit $\alpha_1 \to \frac{1}{2}$. Therefore we choose $\alpha_1 = \frac{1}{2} - o(1)$, so that the discrepancy between the two exponents is only $o(1)$ and can be absorbed into the $\widetilde{\mathcal{O}}$ notation.

Equivalently, the reduction starts from planted clique hardness at sparsity just below the square-root scale and then uses Gaussian cloning to transfer this hardness to the regime $k' > \sqrt{n'}$. Starting significantly below the square-root scale would still give a valid reduction, but it would only yield hardness at a polynomially smaller signal strength and would no longer match the conjectural threshold $\lambda \asymp n'/\left(k'\right)^2$.

12.4 Conclusion

The reduction from planted clique to sparse PCA has two conceptual components. The rejection kernel turns Bernoulli graph entries into approximately Gaussian entries, producing sparse PCA instances at signal strength $\lambda = \widetilde{\mathcal{O}}(1)$ below the square-root sparsity scale. Gaussian cloning then transfers hardness to larger sparsity exponents by increasing the dimension and support size while decreasing the signal strength in a controlled way.

Together, these arguments show that an efficient sparse PCA detector below the threshold

$$\lambda \asymp \min\left\{1, \frac{n}{k^2}\right\}$$

would imply an efficient planted clique detector below the $\sqrt{n}$ threshold. Thus the planted clique conjecture provides conditional evidence that the known sparse PCA algorithms are essentially optimal, up to logarithmic and subpolynomial factors, in the parameter regimes covered by the reduction.

Chapter 13

Reductions *inside* a problem: Planted Clique

13.1 Introduction

In the previous chapter, we discussed a reduction between different detection problems. There is, however, another useful type of reduction: a reduction *inside* a single problem from weak to strong detection. That is, instead of reducing one model to another, one may ask whether different notions of success for the same model are actually equivalent.

We focus on planted clique. The standard belief is that, in the regime $k \ll \sqrt{n}$, even weak detection is computationally hard. A natural question is whether weak (at any reasonable level) and strong detection could nevertheless be computationally different.

Question 13.1. *More precisely, could there exist a polynomial-time algorithm that achieves weak detection for planted clique, while strong detection remains polynomial-time hard?*

The result discussed in this chapter, proven in [HS24], shows that, in a precise sense, this cannot happen: any polynomial-time algorithm that improves sufficiently over the trivial edge-counting advantage can be amplified to a polynomial-time strong detector.

13.2 A baseline weak detector: edge counting

Let Z denote the number of edges of the observed graph. It is easy to see the distributional behavior of the statistics in the two models:

$$\mathbb{P}_1 : \; Z = \text{Bin}\left(\tbinom{n}{2} - \tbinom{k}{2}, \tfrac{1}{2}\right) + \tbinom{k}{2} \approx \tfrac{n^2+k^2}{4} + \sqrt{n^2-k^2}\mathcal{N}(0,1);$$
$$\mathbb{P}_2 : \; Z = \text{Bin}\left(\tbinom{n}{2}, \tfrac{1}{2}\right) \approx \tfrac{n^2}{4} + n\mathcal{N}(0,1).$$

Hence, let us define the edge-counting test

$$\mathcal{T}_{\text{edge}}(G) = \begin{cases} \mathbb{P}_1 & \text{if } Z \geq \dfrac{n^2}{4} + \dfrac{k^2}{8}, \\ \mathbb{P}_2 & \text{otherwise} \end{cases}.$$

Then let us compute the Type I+II errors:

$$\mathbb{P}_1 : \mathbb{P}_1\left(Z \geq \frac{n^2}{4} + \frac{k^2}{8}\right) \approx \mathbb{P}\left(\mathcal{N}(0,1) \geq -\frac{k^2}{n}\right) = \frac{1}{2} + \Theta\left(\frac{k^2}{n}\right);$$
$$\mathbb{P}_2 : \mathbb{P}_2\left(Z \leq \frac{n^2}{4} + \frac{k^2}{8}\right) \approx \mathbb{P}\left(\mathcal{N}(0,1) \leq \frac{k^2}{n}\right) = \frac{1}{2} + \Theta\left(\frac{k^2}{n}\right).$$

Consequently, the sum of Type I and Type II errors of this test is $1 - \Theta\left(\frac{k^2}{n}\right)$, so edge counting gives a weak advantage of order k^2/n.

Question 13.2. *Can a polynomial-time algorithm do substantially better than this, assuming no strong detection is possible?*

13.3 Weak detection implies strong detection

The full theorem of [HS24] proves a broad equivalence between many formulations of the planted clique conjecture, including search and decision versions, weak and strong success probabilities, binomial-size planted cliques, adversarial-size planted cliques, and fixed-size planted cliques. In particular, their results do address the fixed-size planted clique model.

In these notes, however, we do not reproduce the full proof of that theorem. Instead, we isolate one of its main mechanisms: the shrinking reduction. This reduction is especially transparent and already captures the central idea relevant for us, namely that a weak decision advantage substantially larger than the edge-counting advantage can be amplified to strong detection.

The shrinking reduction is most naturally formulated in the binomial planted clique model.

Definition 13.3. *The* binomial planted clique model*, denoted by $\widetilde{\mathcal{G}}_{n,\frac{1}{2},k}$, is the model in which each vertex is included in the planted set independently with probability k/n, and then a clique is planted on the resulting random set. Thus the planted clique size is random, with distribution* $\mathrm{Bin}\left(n, \frac{k}{n}\right)$*. Intuitively,*

$$\widetilde{\mathcal{G}}_{n,\frac{1}{2},k} \equiv \mathcal{G}_{n,\frac{1}{2},\mathrm{Bin}\left(n,\frac{k}{n}\right)}.$$

This model appears naturally when one passes to random induced subgraphs. Indeed, if G is a larger planted graph and G_n is a uniformly random induced subgraph on n vertices, then the number of planted vertices inherited by G_n is random. In the binomial model, this inherited planted set is again binomial with the appropriate mean. This is precisely why the shrinking reduction fits the binomial formulation so cleanly.

By contrast, in the fixed-size model

$$\mathcal{G}_{n,\frac{1}{2},k},$$

the planted clique has exactly size k. A random induced subgraph of a larger fixed-size planted instance does not inherit exactly k planted vertices; it inherits a hypergeometric number of planted vertices. Therefore the induced subgraph is not exactly distributed as $\mathcal{G}_{n,\frac{1}{2},k}$, but rather as a mixture of fixed-size planted clique models with random clique size. For this reason, the shrinking argument alone does not directly prove the fixed-size version.

Theorem 13.4. *Suppose that for some constant $\varepsilon > 0$ and some clique size $k \ll \sqrt{n}$, there is a polynomial-time algorithm $\mathcal{A}_{\mathrm{weak}}$ that distinguishes*

$$\widetilde{\mathcal{G}}_{n,\frac{1}{2},k} = \mathbb{P}_1 \qquad \text{vs} \qquad \mathcal{G}_{n,\frac{1}{2}} = \mathbb{P}_2,$$

and whose sum of Type I and Type II errors is at most

$$1 - n^{\varepsilon}\frac{k^2}{n}.$$

Then there is also a polynomial-time algorithm $\mathcal{A}_{strong}$ that distinguishes

$$\mathcal{G}_{N,\frac{1}{2},K} = \mathbb{P}'_1 \qquad vs \qquad \mathcal{G}_{N,\frac{1}{2}} = \mathbb{P}'_2,$$

and achieves strong detection for planted clique for clique size $K = N^\beta$ for some constant $\beta < 1/2$.

To prove this, [HS24] use a nice concentration inequality from [GLS+15].

Theorem 13.5. *Let $n, N \in \mathbb{N}$. For a graph G on N vertices consider G_n to be a uniformly at random induced subgraph on n vertices. For any event $\mathcal{E}$ on n-vertex graphs, let $\forall G$ N-vertex graph, $f(G) = \mathbb{P}(G_n \in \mathcal{E})$. Then if G is sampled from the planted clique measure $\mathcal{G}_{N,\frac{1}{2},k}$ it holds, w.h.p. that*

$$\lim_{n\to\infty} \left| f(G) - \mathop{\mathbb{E}}_{G'\sim\mathcal{G}_{N,\frac{1}{2},k}} \left[f\left(G'\right)\right] \right| = \mathcal{O}\left(\frac{n}{N}\right),$$

which is equivalent to

$$\left| \mathbb{P}\left(G_n \in \mathcal{E}\right) - \mathop{\mathbb{P}}_{\substack{G_n\sim \mathrm{Unif}\left(\binom{V(G')}{n}\right) \\ G'\sim\mathcal{G}_{N,\frac{1}{2},k}}} \left(G_n \in \mathcal{E}\right) \right| = \mathcal{O}\left(\frac{n}{N}\right).$$

Proof. It is a consequence of [GLS+15, Theorem 1.1]. □

Now, let's see how the reduction works.

Proof of Theorem 13.4. In the detection problem

$$\widetilde{\mathcal{G}}_{n,\frac{1}{2},k} = \mathbb{P}_1 \qquad \text{vs} \qquad \mathcal{G}_{n,\frac{1}{2}} = \mathbb{P}_2,$$

assume for some $0 < \alpha < \frac{1}{2}$ it holds $k = n^\alpha$, and that $\exists \varepsilon > 0$ and a poly-time algorithm $\mathcal{A}_{\text{weak}}$ such that $\mathrm{err}\left(\mathcal{A}_{\text{weak}}\right) \leq 1 - n^\varepsilon \frac{k^2}{n} = 1 - n^{\varepsilon+2\alpha-1}$.

Now, for any N we will prove that strong detection is possible in the detection problem

$$\mathcal{G}_{N,\frac{1}{2},K} = \mathbb{P}'_1 \qquad \text{vs} \qquad \mathcal{G}_{N,\frac{1}{2}} = \mathbb{P}'_2,$$

where

$$K := N^\beta, \qquad \beta := \frac{1-\alpha-\frac{\varepsilon}{2}}{2-2\alpha-\frac{\varepsilon}{2}} < \frac{1}{2}.$$

To do so, let

$$n := N^{1/(2-2\alpha-\varepsilon/2)}, \qquad \mathcal{E} \ : \ G_n \in \mathcal{E} \iff \mathcal{A}_{\text{weak}}\left(G_n\right) = \mathbb{P}_1,$$

i.e. $\mathcal{E}$ is the event of all graphs for which $\mathcal{A}_{\text{weak}}$ outputs $\mathbb{P}_1$. With these choices, it holds that

$$\frac{Kn}{N} = n^\alpha = k.$$

To define $\mathcal{A}_{\text{strong}}$, we previously need some other definitions. Given an input graph G on N vertices, define

$$f(G) := \mathop{\mathbb{P}}_{G_n\sim\mathrm{Unif}\left(\binom{V(G)}{n}\right)} \left(\mathcal{A}_{\text{weak}}(G_n) = \mathbb{P}_1\right).$$

Thus $f(G)$ is the fraction of n-vertex induced subgraphs on which the weak detector $\mathcal{A}$ declares the planted model.

We also define the null acceptance probability of $\mathcal{A}_{\text{weak}}$ by

$$\mu_0 := \mathop{\mathbb{P}}_{G' \sim \mathcal{G}_{n,\frac{1}{2}}} \left(\mathcal{A}_{\text{weak}}\left(G'\right) = \mathbb{P}_1 \right).$$

This quantity need not be available in closed form, but it can be estimated in polynomial time. Indeed, sample independent graphs

$$G^{(1)}, \dots, G^{(R_0)} \overset{\text{i.i.d.}}{\sim} \mathcal{G}_{n,\frac{1}{2}},$$

run $\mathcal{A}_{\text{weak}}$ on each of them, and set

$$\widehat{\mu}_0 := \frac{1}{R_0} \sum_{r=1}^{R_0} \mathbf{1}\left\{ \mathcal{A}_{\text{weak}}\left(G^{(r)}\right) = \mathbb{P}_1 \right\}.$$

The random variables $\mathbf{1}\left\{ \mathcal{A}_{\text{weak}}\left(G^{(r)}\right) = \mathbb{P}_1 \right\}$ are i.i.d. Bernoulli random variables with mean μ_0. Therefore, by Hoeffding's inequality,

$$\mathbb{P}\left(|\widehat{\mu}_0 - \mu_0| > \eta_N \right) \le 2 \exp\left(-2 R_0 \eta_N^2 \right).$$

Thus, it is enough to choose

$$\eta_N := \frac{c}{2} \frac{n \log N}{N}, \qquad R_0 = C \eta_N^{-2} \log N$$

for c sufficiently small constant and C sufficiently large constant. Since $n = N^{1/(2-2\alpha-\varepsilon/2)}$, then for these choices of parameters, $\widehat{\mu}_0$ is computable to the required accuracy (for what follows) whp in $\text{poly}(n)$ time.

Identically, we can compute an estimation $\widehat{f}(G)$ of $f(G)$ to the required accuracy whp in $\text{poly}(N)$ time by sampling independent uniformly random induced subgraphs

$$G_n^{(1)}, \dots, G_n^{(R_1)} \sim \text{Unif}\left(\binom{V(G)}{n} \right).$$

Given all this, the algorithm $\mathcal{A}_{\text{strong}}$ to be considered is defined by

$$\mathcal{A}_{\text{strong}}(G) = \mathbb{P}_2' \quad \Longleftrightarrow \quad \left| \widehat{f}(G) - \widehat{\mu}_0 \right| \le \frac{n}{N} \log N.$$

We now prove that this gives strong detection.

First suppose that

$$G \sim \mathcal{G}_{N,\frac{1}{2}}.$$

Then a uniformly random induced subgraph G_n is distributed exactly as $\mathcal{G}_{n,\frac{1}{2}}$. As $\mathcal{G}_{N,\frac{1}{2}} \equiv \mathcal{G}_{N,\frac{1}{2},0}$, by Theorem 13.5 we have whp

$$|f(G) - \mu_0| = \mathcal{O}\left(\frac{n}{N} \right).$$

Combining this with the estimation bounds gives

$$\left| \widehat{f}(G) - \widehat{\mu}_0 \right| \overset{\triangle}{\le} |f(G) - \mu_0| + \left| \widehat{f}(G) - f(G) \right| + |\widehat{\mu}_0 - \mu_0| \le \mathcal{O}\left(\frac{n}{N} \right) + 2\eta_N < \frac{n \log N}{N}$$

whp for all sufficiently large N. Therefore $\mathcal{A}_{\text{strong}}$ outputs $\mathbb{P}_2'$ whp under the null.

Now suppose that

$$G \sim \mathcal{G}_{N,\frac{1}{2},K}.$$

Here is precisely where it is convenient to formulate the weak detector in the binomial planted clique model. Indeed, by construction, a uniformly random induced subgraph on n vertices inherits a planted set distributed as

$$\mathrm{Bin}\left(n, \frac{K}{N}\right),$$

whose mean is

$$\frac{Kn}{N} = n^{\alpha} = k.$$

At the level of the reduction argument, this is the reason for the choice of K, N, and n: the induced subgraph is meant to be an instance of the smaller binomial-planted-clique problem on which $\mathcal{A}_{\text{weak}}$ has advantage. By the assumption $\mathrm{err}\left(\mathcal{A}_{\text{weak}}\right) \leq n^{\varepsilon}\frac{k^2}{n}$, we have

$$\left| \mathop{\mathbb{P}}_{G' \sim \widetilde{\mathcal{G}}_{n,\frac{1}{2},k}} \left(\mathcal{A}_{\text{weak}}\left(G'\right) = \mathbb{P}_1\right) - \mu_0 \right| \geq n^{\varepsilon}\frac{k^2}{n} = n^{\varepsilon+2\alpha-1}.$$

Moreover, by the definition of f and by Fubini,

$$\mathop{\mathbb{E}}_{G' \sim \mathcal{G}_{N,\frac{1}{2},K}} \left[f\left(G'\right)\right] = \mathop{\mathbb{P}}_{G' \sim \widetilde{\mathcal{G}}_{n,\frac{1}{2},k}} \left(\mathcal{A}_{\text{weak}}\left(G'\right) = \mathbb{P}_1\right) = \mu_1.$$

Thus the assumed advantage of $\mathcal{A}_{\text{weak}}$ gives the separation

$$\left| \mathop{\mathbb{E}}_{G' \sim \mathcal{G}_{N,\frac{1}{2},K}} \left[f\left(G'\right)\right] - \mu_0 \right| \geq n^{\varepsilon+2\alpha-1} \overset{N=n^{2-2\alpha-\varepsilon/2}}{=} n^{\varepsilon/2}\frac{n}{N} \gg \frac{n}{N}\log N.$$

Then Theorem 13.5 gives, whp,

$$\left| f(G) - \mathop{\mathbb{E}}_{G' \sim \mathcal{G}_{N,\frac{1}{2},K}} \left[f\left(G'\right)\right] \right| = \mathcal{O}\left(\frac{n}{N}\right).$$

Therefore, whp,

$$|f(G) - \mu_0| \overset{\triangle}{\geq} n^{\varepsilon+2\alpha-1} - \mathcal{O}\left(\frac{n}{N}\right) \gg \frac{n}{N}\log N.$$

After accounting for the two estimation errors,

$$\left|\widehat{f}(G) - \widehat{\mu}_0\right| \overset{\triangle}{\geq} |f(G) - \mu_0| - \left|\widehat{f}(G) - f(G)\right| - |\widehat{\mu}_0 - \mu_0| > \frac{n}{N}\log N$$

whp for all sufficiently large N. Therefore $\mathcal{A}_{\text{strong}}$ outputs $\mathbb{P}_1'$ whp under the planted model. □

Part III

Approximate Recovery

Chapter 14

Low-Degree Estimation Lower Bounds

14.1 Introduction

Up to this point, the notes have focused on computational lower bounds for detection. In a detection problem, one observes data Y and tries to decide whether it was generated from a structured distribution $\mathbb{P}_1$ or from a null distribution $\mathbb{P}_2$. This framework is powerful because computational lower bounds can often be formulated through restricted testing classes, such as low-degree polynomials or statistical queries.

As we introduced in Definition Definition 1.7, approximate recovery is different. Here there is no null model. A hidden parameter is sampled from a prior, the data are generated conditionally on this parameter, and the goal is to estimate the parameter itself. More precisely, let

$$\theta \sim \mu, \qquad Y \sim \mathbb{P}_\theta.$$

An estimator is a measurable function $\mathcal{A}(Y)$, and the natural objective is to minimize the mean-squared error

$$\min_{\mathcal{A}} \mathop{\mathbb{E}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta} \left[\|\mathcal{A}(Y) - \theta\|_2^2\right].$$

The statistically optimal estimator is the posterior mean

$$\mathcal{A}_{\mathrm{Bayes}}(Y) = \mathop{\mathbb{E}}_{\theta\sim\mu}[\theta \mid Y].$$

However, the formula

$$\mathop{\mathbb{E}}_{\theta\sim\mu}[\theta \mid Y] = \sum_{\theta'} \theta' \mathbb{P}\left(\theta' \mid Y\right)$$

may involve a sum over an exponentially large parameter space. Thus the Bayes estimator is statistically optimal but may be computationally infeasible.

One could try to define the optimal polynomial-time estimator by writing

$$\min_{\mathcal{A}\,:\,\mathrm{time}(\mathcal{A})\le n^{100}} \mathop{\mathbb{E}}_{\theta,Y} \left[\|\mathcal{A}(Y) - \theta\|_2^2\right].$$

As discussed in Section 5.1, this is not a tractable object to analyze directly. The main idea of this chapter is to replace the class of polynomial-time estimators by low-degree polynomial estimators. This gives a restricted but analyzable model of computation, analogous to the low-degree likelihood-ratio method for detection.

The chapter follows three steps. First, we explain why one has to be careful when comparing detection and estimation: there can be genuine detection-estimation gaps. Then we introduce the low-degree mean-squared error from [SW22] and reduce its analysis to a correlation maximization problem. Finally, in the Gaussian additive model, we use Hermite analysis to obtain an explicit upper bound on low-degree correlation, and we apply it to sparse PCA. All non-trivial technical work of this chapter is based on the important work [SW22].

14.2 A Caution: Detection-Estimation Gaps

Before developing lower bounds for estimation, it is important to understand why detection and recovery should not be conflated. Sometimes detection is much easier than estimation, and a test that succeeds at detecting structure may give essentially no information about where the structure is.

Consider the nonnegative sparse PCA model

$$\mathbb{P}_1 : \qquad Y = \lambda xx^\top + W, \qquad x \sim \mathrm{Unif}\left(\left\{v \in \left\{0, \frac{1}{\sqrt{k}}\right\}^n : \|v\|_0 = k\right\}\right), \qquad W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

and compare it with the null model

$$\mathbb{P}_2 : \qquad Y = W.$$

If the entries of x were signed, that is, if the nonzero entries were $\pm 1/\sqrt{k}$ (model already discussed in Chapters 6 and 9), then detection is impossible below the information-theoretic threshold, easy above the known polynomial-time threshold, and conjecturally hard in between. The usual qualitative picture is shown in Figure 14.1.

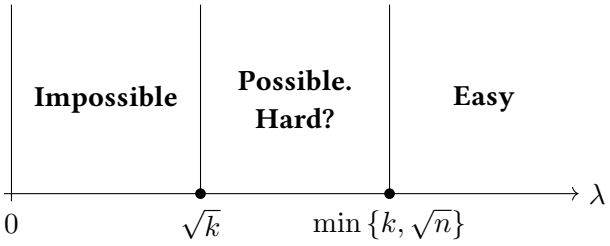

Figure 14.1: Signed-Sparse PCA phase diagram for detection

For estimation in the nonnegative model, the same qualitative diagram remains the relevant one: recovering the support of x is still hard in the intermediate regime. But detection becomes easier because nonnegativity introduces a global mean shift. Indeed, summing all entries of Y gives

$$\sum_{i,j} Y_{ij} = \lambda \sum_{i,j} x_i x_j + \sum_{i,j} W_{ij}.$$

Under $\mathbb{P}_1$,

$$\sum_{i,j} x_i x_j = \left(\sum_i x_i\right)^2 = \left(k \cdot \frac{1}{\sqrt{k}}\right)^2 = k,$$

so

$$\sum_{i,j} Y_{ij} = \lambda k + \mathcal{O}(n).$$

Under $\mathbb{P}_2$,

$$\sum_{i,j} Y_{ij} = \mathcal{O}(n).$$

Therefore the sum-test detects the planted model whenever

$$\lambda k \gg n,$$

or equivalently

$$\lambda \gg \frac{n}{k}.$$

This test has nothing to do with finding the support of x. It only detects the global positive bias produced by the nonnegative spike. Thus, if $\sqrt{n} \ll k \ll n^{2/3}$, detection can become easy at $\lambda \gg n/k$ (see Figure 14.2), while estimation may still have a hard phase. If $k \gg n^{2/3}$, the detection hard phase can disappear altogether, even though recovery remains meaningful (see Figure 14.3).

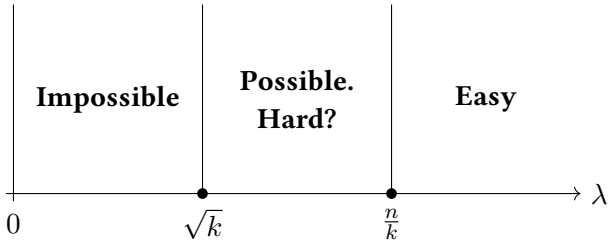

Figure 14.2: Nonnegative Sparse PCA phase diagram for detection when $\sqrt{n} \ll k \ll n^{2/3}$

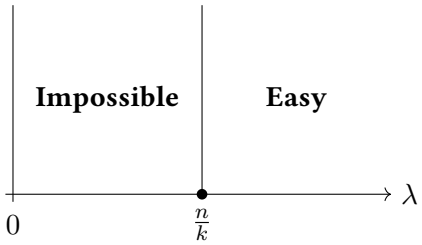

Figure 14.3: Nonnegative Sparse PCA phase diagram for detection when $k \gg n^{2/3}$

This example is the main warning: lower bounds for detection do not automatically imply lower bounds for estimation, and upper bounds for detection do not automatically give recovery algorithms. The rest of this chapter develops a framework that targets estimation directly.

14.3 Low-Degree Lower Bounds for Estimation

14.3.1 The scalar reduction

A natural low-degree analogue of the MMSE is

$$\mathrm{MMSE}_{\le D} := \min_{\mathcal{A}(Y)\in\mathbb{R}_{\le D}[Y]} \mathop{\mathbb{E}}_{\theta,Y} \left[\|\mathcal{A}(Y)-\theta\|_2^2\right],$$

where $\mathbb{R}_{\le D}[Y]$ denotes the space of polynomials of degree at most D in the entries of Y.

In many models, the prior is symmetric across coordinates. In that case, it is enough to study one coordinate. Indeed,

$$\frac{1}{n}\mathop{\mathbb{E}}_{\theta,Y}\left[\left\|\mathbb{E}[\theta \mid Y]-\theta\right\|_2^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{\theta,Y}\left[\Big(\mathbb{E}[\theta_i \mid Y]-\theta_i\Big)^2\right].$$

By symmetry, each term in the sum is the same, so the normalized MMSE equals

$$\mathop{\mathbb{E}}_{\theta,Y}\left[\Big(\mathbb{E}[\theta_1 \mid Y] - \theta_1\Big)^2\right].$$

This motivates the scalar low-degree MMSE

$$\mathrm{MMSE}_{\le D} := \min_{A \in \mathbb{R}_{\le D}[Y]} \mathop{\mathbb{E}}_{\theta,Y}\left[(A(Y) - \theta_1)^2\right].$$

The following correlation formulation is more convenient. Define

$$\mathrm{corr}_{\le D} := \max_{A \in \mathbb{R}_{\le D}[Y],\, A \neq 0} \frac{\mathop{\mathbb{E}}_{\theta,Y}[A(Y)\theta_1]}{\sqrt{\mathop{\mathbb{E}}_{\theta,Y}[A(Y)^2]}}.$$

Equivalently, one can maximize over all degree-D polynomials normalized by

$$\mathop{\mathbb{E}}_{\theta,Y}\left[A(Y)^2\right] = 1.$$

Lemma 14.1. *We have*

$$\mathrm{MMSE}_{\le D} = \mathop{\mathbb{E}}_{\theta}\left[\theta_1^2\right] - \mathrm{corr}^2_{\le D}.$$

Proof. For any polynomial $A \in \mathbb{R}_{\le D}[Y]$,

$$\mathop{\mathbb{E}}_{\theta,Y}\left[(\theta_1 - A(Y))^2\right] = \mathop{\mathbb{E}}_{\theta}\left[\theta_1^2\right] - 2\mathop{\mathbb{E}}_{\theta,Y}[\theta_1 A(Y)] + \mathop{\mathbb{E}}_{\theta,Y}\left[A(Y)^2\right].$$

The first term is independent of A. Now fix a nonzero polynomial A and optimize over scalar multiples tA. The risk becomes

$$\mathop{\mathbb{E}}_{\theta}\left[\theta_1^2\right] - 2t\mathop{\mathbb{E}}_{\theta,Y}[\theta_1 A(Y)] + t^2\mathop{\mathbb{E}}_{\theta,Y}\left[A(Y)^2\right].$$

The minimizing value of t is

$$t_\star = \frac{\mathop{\mathbb{E}}_{\theta,Y}[\theta_1 A(Y)]}{\mathop{\mathbb{E}}_{\theta,Y}[A(Y)^2]}.$$

Substituting this value gives

$$\min_{t \in \mathbb{R}} \mathop{\mathbb{E}}_{\theta,Y}\left[(\theta_1 - tA(Y))^2\right] = \mathop{\mathbb{E}}_{\theta}\left[\theta_1^2\right] - \frac{\mathop{\mathbb{E}}_{\theta,Y}[\theta_1 A(Y)]^2}{\mathop{\mathbb{E}}_{\theta,Y}[A(Y)^2]}.$$

Finally, maximizing the last quotient over nonzero $A \in \mathbb{R}_{\le D}[Y]$ gives

$$\mathrm{MMSE}_{\le D} = \mathop{\mathbb{E}}_{\theta}\left[\theta_1^2\right] - \mathrm{corr}^2_{\le D}.$$

□

Thus a lower bound on the low-degree MMSE is equivalent to an upper bound on the best low-degree correlation with the target coordinate.

14.3.2 A linear-algebra formulation

Let $\{h_\alpha\}_{\alpha\in\mathcal{I}_D}$ be a basis of $\mathbb{R}_{\le D}[Y]$, where $\mathcal{I}_D$ indexes monomials, or any other convenient polynomial basis, of degree at most D. We do not assume that this basis is orthonormal. Write

$$A(Y) = \sum_{\alpha\in\mathcal{I}_D} \widehat{A}_\alpha h_\alpha(Y).$$

Then

$$\mathop{\mathbb{E}}_{\theta,Y}[A(Y)\theta_1] = \sum_{\alpha\in\mathcal{I}_D} \widehat{A}_\alpha \mathop{\mathbb{E}}_{\theta,Y}[h_\alpha(Y)\theta_1].$$

Define

$$c_\alpha := \mathop{\mathbb{E}}_{\theta,Y}[h_\alpha(Y)\theta_1].$$

Then the numerator is

$$\left\langle \widehat{A}, c \right\rangle.$$

Similarly,

$$\mathop{\mathbb{E}}_{\theta,Y}\left[A(Y)^2\right] = \sum_{\alpha,\beta} \widehat{A}_\alpha \widehat{A}_\beta \mathop{\mathbb{E}}_{\theta,Y}\left[h_\alpha(Y)h_\beta(Y)\right].$$

Define the Gram matrix

$$P_{\alpha,\beta} := \mathop{\mathbb{E}}_{\theta,Y}\left[h_\alpha(Y)h_\beta(Y)\right].$$

Then

$$\mathop{\mathbb{E}}_{\theta,Y}\left[A(Y)^2\right] = \widehat{A}^\top P \widehat{A}.$$

Assuming the basis is linearly independent in $L^2(\mathbb{P})$, the matrix P is positive definite. Indeed, for any coefficient vector a,

$$a^\top P a = \mathop{\mathbb{E}}_{\theta,Y}\left[\left(\sum_\alpha a_\alpha h_\alpha(Y)\right)^2\right],$$

which is zero only if the corresponding polynomial is zero in $L^2(\mathbb{P})$.

Therefore

$$\mathrm{corr}^2_{\le D} = \max_{\widehat{A}\neq 0} \frac{\left\langle c, \widehat{A}\right\rangle^2}{\widehat{A}^\top P \widehat{A}}.$$

Setting $g = P^{1/2}\widehat{A}$, this becomes

$$\mathrm{corr}^2_{\le D} = \max_{g\neq 0} \frac{\left\langle P^{-1/2}c, g\right\rangle^2}{\|g\|_2^2} = c^\top P^{-1} c.$$

This formula is exact, but it is usually not directly usable: the Gram matrix P is computed under the planted distribution, which is generally not a product measure, and P^{-1} can be very complicated. The next section explains how, in Gaussian additive models, one can avoid computing P^{-1} exactly.

14.4 Gaussian Additive Models

14.4.1 Jensen's trick

Consider the Gaussian additive model

$$Y = X + Z, \qquad X \sim \mu, \qquad Z \sim \mathcal{N}(0, I_N),$$

where X is the signal and Z is independent Gaussian noise. Let x_1 denote the scalar quantity we want to estimate from Y. In applications, x_1 will be one coordinate, or one support indicator, of the underlying hidden structure.

For any polynomial A,

$$\mathop{\mathbb{E}}_{X,Z}\left[A(Y)^2\right] = \mathop{\mathbb{E}}_{X,Z}\left[A(X+Z)^2\right].$$

By Jensen's inequality, conditioning on Z gives

$$\mathop{\mathbb{E}}_{X,Z}\left[A(X+Z)^2\right] \geq \mathop{\mathbb{E}}_{Z}\left[\left(\mathop{\mathbb{E}}_{X}[A(X+Z)]\right)^2\right].$$

Thus

$$\mathrm{corr}^2_{\leq D} \leq \max_{A \in \mathbb{R}_{\leq D}[Y],\, A \neq 0} \frac{\mathop{\mathbb{E}}_{X,Z}[A(X+Z)x_1]^2}{\mathop{\mathbb{E}}_{Z}\left[\left(\mathop{\mathbb{E}}_{X}[A(X+Z)]\right)^2\right]}.$$

The advantage is that the denominator is now an L^2 norm under the standard Gaussian measure in Z. This allows us to use the Hermite basis.

14.4.2 Translation identity

Let $\widehat{h}_k$ denote the normalized probabilists' Hermite polynomial of degree k. Thus the family of polynomials $\left\{\widehat{h}_k : k \geq 0\right\}$ is orthonormal in $L^2(\mathcal{N}(0,1))$.

Lemma 14.2 (Translation identity)**.** *For every $\mu, z \in \mathbb{R}$ and every $k \in \mathbb{N}$,*

$$\widehat{h}_k(\mu + z) = \sum_{\ell=0}^{k} \sqrt{\frac{\ell!}{k!}} \binom{k}{\ell} \mu^{k-\ell} \widehat{h}_\ell(z).$$

In particular, if $Z \sim \mathcal{N}(0,1)$, then

$$\mathop{\mathbb{E}}_{Z}\left[\widehat{h}_k(\mu + Z)\right] = \frac{\mu^k}{\sqrt{k!}}.$$

The proof is given in Section A.8.

For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathbb{N}^N$, write

$$|\alpha| := \sum_{i=1}^{N} \alpha_i, \qquad \alpha! := \prod_{i=1}^{N} \alpha_i!, \qquad X^\alpha := \prod_{i=1}^{N} X_i^{\alpha_i},$$

and

$$\binom{\alpha}{\beta} := \prod_{i=1}^{N} \binom{\alpha_i}{\beta_i}.$$

Let

$$\widehat{H}_\alpha(z) := \prod_{i=1}^{N} \widehat{h}_{\alpha_i}(z_i).$$

The multivariate translation identity follows coordinatewise.

14.4.3 A general low-degree correlation bound

We now state the main Gaussian-additive bound.

Theorem 14.3 ([SW22, Theorem 2.2.]). *Consider the Gaussian additive model*

$$Y = X + Z, \qquad X \sim \mu, \qquad Z \sim \mathcal{N}(0, I_N),$$

and let x_1 be the scalar quantity to be estimated. Define coefficients κ_α recursively by

$$\kappa_0 = \mathop{\mathbb{E}}_{X}[x_1],$$

and, for $\alpha \neq 0$,

$$\kappa_\alpha = \mathop{\mathbb{E}}_{X}\left[x_1 X^\alpha\right] - \sum_{0 \leq \beta < \alpha} \kappa_\beta \binom{\alpha}{\beta} \mathop{\mathbb{E}}_{X}\left[X^{\alpha-\beta}\right],$$

where $\beta < \alpha$ means $\beta \leq \alpha$ coordinatewise and $\beta \neq \alpha$. Then

$$\mathrm{corr}^2_{\leq D} \leq \sum_{\alpha \in \mathbb{N}^N,\, |\alpha| \leq D} \frac{\kappa_\alpha^2}{\alpha!}.$$

Proof. Let $A \in \mathbb{R}_{\leq D}[Y]$. Expand A in the normalized multivariate Hermite basis:

$$A(Y) = \sum_{|\alpha| \leq D} \widehat{A}_\alpha \widehat{H}_\alpha(Y).$$

The numerator is

$$\mathop{\mathbb{E}}_{X,Z}[A(X+Z)x_1] = \sum_{|\alpha| \leq D} \widehat{A}_\alpha \mathop{\mathbb{E}}_{X,Z}\left[\widehat{H}_\alpha(X+Z)x_1\right].$$

Using the translation identity and then averaging over Z,

$$\mathop{\mathbb{E}}_{Z}\left[\widehat{H}_\alpha(X+Z)\right] = \frac{X^\alpha}{\sqrt{\alpha!}}.$$

Thus

$$\mathop{\mathbb{E}}_{X,Z}[A(X+Z)x_1] = \sum_{|\alpha| \leq D} \widehat{A}_\alpha \frac{\mathop{\mathbb{E}}_{X}\left[x_1 X^\alpha\right]}{\sqrt{\alpha!}}.$$

Define

$$c_\alpha := \frac{\mathop{\mathbb{E}}_{X}\left[x_1 X^\alpha\right]}{\sqrt{\alpha!}}.$$

Then the numerator is $\left\langle \widehat{A}, c \right\rangle$.

We now lower-bound the denominator using Jensen's trick. Define

$$g(Z) := \mathop{\mathbb{E}}_{X}[A(X+Z)].$$

Then

$$\mathop{\mathbb{E}}_{X,Z}\left[A(X+Z)^2\right] \geq \mathop{\mathbb{E}}_{Z}\left[g(Z)^2\right].$$

Expand g in the Hermite basis:

$$g(Z) = \sum_{|\gamma|\leq D} \widehat{g}_\gamma \widehat{H}_\gamma(Z).$$

Since the Hermite basis is orthonormal under $Z \sim \mathcal{N}(0, I_N)$,

$$\mathop{\mathbb{E}}_{Z}\left[g(Z)^2\right] = \sum_{|\gamma|\leq D} \widehat{g}_\gamma^2.$$

We compute $\widehat{g}_\gamma$. By definition,

$$\widehat{g}_\gamma = \mathop{\mathbb{E}}_{Z}\left[g(Z)\widehat{H}_\gamma(Z)\right] = \mathop{\mathbb{E}}_{X,Z}\left[A(X+Z)\widehat{H}_\gamma(Z)\right].$$

Substituting the Hermite expansion of A,

$$\widehat{g}_\gamma = \sum_{|\alpha|\leq D} \widehat{A}_\alpha \mathop{\mathbb{E}}_{X,Z}\left[\widehat{H}_\alpha(X+Z)\widehat{H}_\gamma(Z)\right].$$

Using the multivariate translation identity, only terms with $\gamma \leq \alpha$ survive by orthonormality. Hence

$$\mathop{\mathbb{E}}_{Z}\left[\widehat{H}_\alpha(X+Z)\widehat{H}_\gamma(Z)\right] = \mathbf{1}_{\gamma\leq\alpha}\sqrt{\frac{\gamma!}{\alpha!}}\binom{\alpha}{\gamma}X^{\alpha-\gamma}.$$

Therefore

$$\widehat{g}_\gamma = \sum_{\alpha\geq\gamma} \widehat{A}_\alpha \sqrt{\frac{\gamma!}{\alpha!}}\binom{\alpha}{\gamma}\mathop{\mathbb{E}}_{X}\left[X^{\alpha-\gamma}\right].$$

Define the upper-triangular matrix M by

$$M_{\gamma,\alpha} := \mathbf{1}_{\gamma\leq\alpha}\sqrt{\frac{\gamma!}{\alpha!}}\binom{\alpha}{\gamma}\mathop{\mathbb{E}}_{X}\left[X^{\alpha-\gamma}\right].$$

Then

$$\widehat{g} = M\widehat{A},$$

and hence

$$\mathop{\mathbb{E}}_{X,Z}\left[A(X+Z)^2\right] \geq \left\|M\widehat{A}\right\|_2^2.$$

The matrix M is triangular with diagonal entries equal to 1, because $M_{\alpha,\alpha} = 1$. Hence M is invertible.

We have shown that

$$\mathrm{corr}_{\leq D} \leq \max_{\widehat{A}\neq 0} \frac{\left\langle c, \widehat{A}\right\rangle}{\left\|M\widehat{A}\right\|_2}.$$

Let $u = M\widehat{A}$. Then $\widehat{A} = M^{-1}u$, and so

$$\left\langle c, \widehat{A}\right\rangle = \left\langle c, M^{-1}u\right\rangle = \left\langle M^{-\top}c, u\right\rangle.$$

Therefore

$$\mathrm{corr}_{\leq D} \leq \left\|M^{-\top}c\right\|_2.$$

Let

$$w := M^{-\top} c.$$

Equivalently,

$$M^\top w = c.$$

Writing this coordinatewise, for every α,

$$c_\alpha = \sum_{\beta \le \alpha} M_{\beta,\alpha} w_\beta = \sum_{\beta \le \alpha} w_\beta \sqrt{\frac{\beta!}{\alpha!}} \binom{\alpha}{\beta} \mathop{\mathbb{E}}_X \left[X^{\alpha-\beta} \right].$$

Multiplying by $\sqrt{\alpha!}$ gives

$$\mathop{\mathbb{E}}_X [x_1 X^\alpha] = \sum_{\beta \le \alpha} w_\beta \sqrt{\beta!} \binom{\alpha}{\beta} \mathop{\mathbb{E}}_X \left[X^{\alpha-\beta} \right].$$

Now define

$$\kappa_\beta := w_\beta \sqrt{\beta!}.$$

Then

$$\mathop{\mathbb{E}}_X [x_1 X^\alpha] = \sum_{\beta \le \alpha} \kappa_\beta \binom{\alpha}{\beta} \mathop{\mathbb{E}}_X \left[X^{\alpha-\beta} \right].$$

Solving this identity recursively gives

$$\kappa_\alpha = \mathop{\mathbb{E}}_X [x_1 X^\alpha] - \sum_{0 \le \beta < \alpha} \kappa_\beta \binom{\alpha}{\beta} \mathop{\mathbb{E}}_X \left[X^{\alpha-\beta} \right].$$

Finally,

$$\|w\|_2^2 = \sum_{|\alpha| \le D} w_\alpha^2 = \sum_{|\alpha| \le D} \frac{\kappa_\alpha^2}{\alpha!}.$$

Thus

$$\mathrm{corr}^2_{\le D} \le \sum_{|\alpha| \le D} \frac{\kappa_\alpha^2}{\alpha!},$$

as claimed. □

14.5 Application to Sparse PCA

We now apply the preceding theorem to the nonnegative sparse PCA model

$$\mathbb{P}_1 : \qquad Y = \lambda v v^\top + W,$$

where

$$v_i \overset{\text{i.i.d.}}{\sim} \begin{cases} 0 & \text{with probability } 1-\rho \\ \frac{1}{\sqrt{k}} & \text{with probability } \rho \end{cases} \qquad \rho := \frac{k}{n},$$

and

$$W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

The null model is

$$\mathbb{P}_2 : \qquad Y = W.$$

The predicted picture for this model was discussed earlier (see Figure 14.1).

To apply the Gaussian-additive theorem, it is cleaner to estimate the support indicator

$$x_i := \sqrt{k} v_i \in \{0, 1\},$$

so that

$$\mathbb{P}(x_i = 1) = \rho.$$

Then

$$\lambda v v^\top = \frac{\lambda}{k} x x^\top.$$

Thus the Gaussian additive signal has coordinates

$$X_{ij} := \frac{\lambda}{k} x_i x_j, \qquad 1 \le i, j \le n.$$

We estimate x_1 from $Y = X + W$.

For a multi-index $\alpha = (\alpha_{ij})_{1 \le i,j \le n}$, it is useful to interpret α as a weighted multigraph on vertex set $[n]$: the number α_{ij} is the multiplicity of the edge (i, j). We write

$$|\alpha| := \sum_{i,j} \alpha_{ij}.$$

Let $V(\alpha)$ be the set of vertices incident to at least one edge with positive multiplicity, and let $|V(\alpha)|$ be its size. When discussing connectedness, we ignore orientations and multiplicities.

For any multi-index γ,

$$X^\gamma = \prod_{i,j} \left(\frac{\lambda}{k} x_i x_j \right)^{\gamma_{ij}}.$$

Since $x_i \in \{0, 1\}$, powers of x_i are again equal to x_i whenever the exponent is positive. Therefore

$$\mathop{\mathbb{E}}_X [X^\gamma] = \left(\frac{\lambda}{k} \right)^{|\gamma|} \rho^{|V(\gamma)|}.$$

Similarly,

$$\mathop{\mathbb{E}}_X [x_1 X^\gamma] = \left(\frac{\lambda}{k} \right)^{|\gamma|} \rho^{|V(\gamma) \cup \{1\}|}.$$

The goal is to show that, if λ lies below the conjectural polynomial-time recovery threshold, then low-degree polynomials cannot correlate with x_1 substantially better than the trivial estimator.

14.5.1 A structural vanishing lemma

The first important observation is that many coefficients κ_α vanish exactly.

Lemma 14.4. *If α has a connected component that does not contain vertex 1, then*

$$\kappa_\alpha = 0.$$

The proof is most naturally expressed using cumulants. We first recall the relevant facts.

Definition 14.5. *Let $(Z_1, \dots, Z_m)$ be jointly distributed random variables. Their joint cumulant is*

$$\kappa(Z_1, \dots, Z_m) := \frac{\partial^m}{\partial t_1 \cdots \partial t_m} \log \mathbb{E}\left[\exp\left(\sum_{i=1}^m t_i Z_i\right)\right]\Big|_{t_1=\cdots=t_m=0}.$$

Equivalently,

$$\kappa(Z_1, \dots, Z_m) = \sum_{\pi \in \mathcal{P}([m])} (|\pi|-1)!(-1)^{|\pi|-1} \prod_{B \in \pi} \mathbb{E}\left[\prod_{i \in B} Z_i\right],$$

where $\mathcal{P}([m])$ denotes the set of partitions of $[m]$.

The basic property we need is that joint cumulants vanish across independent blocks: if the variables can be split into two nonempty independent groups, then the joint cumulant of all variables is zero.

The coefficients κ_α defined above are precisely joint cumulants. More explicitly,

$$\kappa_\alpha = \kappa\left(x_1,\ \{X_{ij} \text{ repeated } \alpha_{ij} \text{ times}\}_{i,j}\right).$$

This identity follows by comparing the recursive moment-cumulant relation

$$\kappa(Z_0, Z_1, \dots, Z_m) = \mathbb{E}\left[\prod_{i=0}^m Z_i\right] - \sum_{\emptyset \neq S \subseteq \{1,\dots,m\}} \kappa(Z_i : i \notin S)\mathbb{E}\left[\prod_{i \in S} Z_i\right]$$

with the recursion defining the coefficients κ_α.

Proof of Lemma 14.4. Now suppose α has a connected component C not containing vertex 1. The variables X_{ij} corresponding to edges inside C depend only on the support indicators $\{x_i : i \in C\}$. The remaining variables, together with x_1, depend only on support indicators outside C. Since the coordinates of x are independent, these two groups of random variables are independent. Hence the joint cumulant factors across two independent groups and must vanish. Therefore

$$\kappa_\alpha = 0.$$

□

This lemma is the reason connected graphs appear in the counting argument: only the connected component containing vertex 1 contributes to the low-degree correlation upper bound.

14.5.2 Bounding the cumulant coefficients

The next lemma gives a crude but sufficient bound on the nonzero coefficients.

Lemma 14.6. *We have*

$$\kappa_0 = \rho.$$

Moreover, for every α with $|\alpha| \geq 1$,

$$|\kappa_\alpha| \leq (|\alpha|+1)^{|\alpha|}\left(\frac{\lambda}{k}\right)^{|\alpha|} \rho^{|V(\alpha)|}.$$

Proof. The identity $\kappa_0 = \rho$ follows from

$$\kappa_0 = \mathbb{E}[x_1] = \rho.$$

We prove the general bound by induction on $|\alpha|$. From the recursive definition,

$$\kappa_\alpha = \mathbb{E}[x_1 X^\alpha] - \sum_{0 \le \beta < \alpha} \kappa_\beta \binom{\alpha}{\beta} \mathbb{E}\left[X^{\alpha-\beta}\right].$$

Taking absolute values,

$$|\kappa_\alpha| \le \left|\mathbb{E}[x_1 X^\alpha]\right| + \sum_{0 \le \beta < \alpha} |\kappa_\beta| \binom{\alpha}{\beta} \mathbb{E}\left[X^{\alpha-\beta}\right].$$

We already observed that

$$\mathbb{E}[X^\gamma] = \left(\frac{\lambda}{k}\right)^{|\gamma|} \rho^{|V(\gamma)|},$$

and

$$\mathbb{E}[x_1 X^\gamma] = \left(\frac{\lambda}{k}\right)^{|\gamma|} \rho^{|V(\gamma)\cup\{1\}|} \le \left(\frac{\lambda}{k}\right)^{|\gamma|} \rho^{|V(\gamma)|},$$

because $0 < \rho \le 1$.

The term $\beta = 0$ contributes

$$|\kappa_0| \mathbb{E}[X^\alpha] = \rho \left(\frac{\lambda}{k}\right)^{|\alpha|} \rho^{|V(\alpha)|} \le \left(\frac{\lambda}{k}\right)^{|\alpha|} \rho^{|V(\alpha)|}.$$

Together with the first moment term, these give at most

$$2 \left(\frac{\lambda}{k}\right)^{|\alpha|} \rho^{|V(\alpha)|}.$$

For the remaining terms, let $0 < \beta < \alpha$. By the induction hypothesis,

$$|\kappa_\beta| \le (|\beta| + 1)^{|\beta|} \left(\frac{\lambda}{k}\right)^{|\beta|} \rho^{|V(\beta)|}.$$

Also,

$$\mathbb{E}\left[X^{\alpha-\beta}\right] = \left(\frac{\lambda}{k}\right)^{|\alpha-\beta|} \rho^{|V(\alpha-\beta)|}.$$

Multiplying,

$$|\kappa_\beta| \mathbb{E}\left[X^{\alpha-\beta}\right] \le (|\beta| + 1)^{|\beta|} \left(\frac{\lambda}{k}\right)^{|\alpha|} \rho^{|V(\beta)|+|V(\alpha-\beta)|}.$$

Since every vertex incident to an edge of α is incident either to an edge of β or to an edge of $\alpha - \beta$,

$$V(\alpha) \subseteq V(\beta) \cup V(\alpha - \beta).$$

Thus

$$|V(\beta)| + |V(\alpha - \beta)| \ge |V(\alpha)|.$$

Because $\rho \le 1$, this implies

$$\rho^{|V(\beta)|+|V(\alpha-\beta)|} \le \rho^{|V(\alpha)|}.$$

Therefore

$$|\kappa_\alpha| \le \left(\frac{\lambda}{k}\right)^{|\alpha|} \rho^{|V(\alpha)|} \left(2 + \sum_{0<\beta<\alpha} (|\beta|+1)^{|\beta|} \binom{\alpha}{\beta}\right).$$

It remains to bound the combinatorial factor. If $|\beta| = \ell$, then

$$(|\beta|+1)^{|\beta|} = (\ell+1)^\ell.$$

Moreover,

$$\sum_{\beta \le \alpha:\ |\beta|=\ell} \binom{\alpha}{\beta} \le \binom{|\alpha|}{\ell},$$

by the multinomial interpretation of choosing ℓ units from the $|\alpha|$ total units of α. Hence

$$2 + \sum_{0<\beta<\alpha} (|\beta|+1)^{|\beta|} \binom{\alpha}{\beta} \le 2 + \sum_{\ell=1}^{|\alpha|-1} (\ell+1)^\ell \binom{|\alpha|}{\ell}.$$

A crude bound gives

$$2 + \sum_{\ell=1}^{|\alpha|-1} (\ell+1)^\ell \binom{|\alpha|}{\ell} \le \sum_{\ell=0}^{|\alpha|} (|\alpha|+1)^\ell \binom{|\alpha|}{\ell} = (|\alpha|+2)^{|\alpha|}.$$

Changing the harmless universal constant in the base, this is bounded by

$$(|\alpha|+1)^{|\alpha|}$$

up to a universal factor, which can be absorbed into the same type of estimate. Thus

$$|\kappa_\alpha| \le (|\alpha|+1)^{|\alpha|} \left(\frac{\lambda}{k}\right)^{|\alpha|} \rho^{|V(\alpha)|},$$

as claimed. □

14.5.3 Counting connected multigraphs

We now need to count the possible multi-indices α that survive the cumulant vanishing lemma.

Lemma 14.7. *Fix $d \ge 1$ and $0 \le h \le d$. The number of connected multi-indices α such that*

$$|\alpha| = d, \qquad 1 \in V(\alpha), \qquad |V(\alpha)| = d + 1 - h,$$

is at most

$$(Cdn)^d \left(\frac{Cd}{n}\right)^h$$

for a universal constant $C > 0$.

Proof. Let

$$r := |V(\alpha)| = d + 1 - h.$$

We construct such a connected multigraph by first choosing its vertex set and then choosing its edges. Since vertex 1 must belong to the graph, the remaining $r - 1 = d - h$ vertices can be chosen in at most

$$\binom{n}{d-h} \le n^{d-h}$$

ways. Once the vertex set is chosen, a connected multigraph with d edges contains a spanning tree on these r vertices, which has $r-1 = d-h$ edges, plus h additional edges. The number of possible spanning trees is at most $r^{r-2} \le d^d$, and the additional h edges can be chosen with repetition from at most $r^2 \le d^2$ possible ordered pairs. This gives at most $d^d d^{2h}$ choices after the vertex set is fixed.

Combining the estimates,

$$\#\{\alpha\} \le n^{d-h} d^{d+2h}.$$

Since $h \le d$, this is bounded by

$$(Cdn)^d \left(\frac{Cd}{n}\right)^h$$

after adjusting the universal constant C. This proves the desired bound. □

14.5.4 The low-degree lower bound

We now combine the previous estimates. By the general Gaussian-additive theorem,

$$\mathrm{corr}^2_{\le D} \le \sum_{|\alpha| \le D} \frac{\kappa_\alpha^2}{\alpha!}.$$

Since $\alpha! \ge 1$, we may upper-bound this by

$$\sum_{|\alpha| \le D} \kappa_\alpha^2.$$

By the vanishing lemma, only connected components containing vertex 1 contribute. The zero multi-index contributes

$$\kappa_0^2 = \rho^2.$$

For $|\alpha| = d \ge 1$, write

$$|V(\alpha)| = d + 1 - h.$$

Using the cumulant bound and the counting lemma,

$$\sum_{\substack{|\alpha|=d \\ 1 \in V(\alpha) \\ \alpha \text{ connected}}} \kappa_\alpha^2 \le \sum_{h=0}^{d} (Cdn)^d \left(\frac{Cd}{n}\right)^h (d+1)^{2d} \left(\frac{\lambda}{k}\right)^{2d} \rho^{2(d+1-h)}.$$

Since $\rho = k/n$, this becomes

$$\sum_{\substack{|\alpha|=d \\ 1 \in V(\alpha) \\ \alpha \text{ connected}}} \kappa_\alpha^2 \le \rho^2 \sum_{h=0}^{d} \left(Cd(d+1)^2 \frac{\lambda^2}{n}\right)^d \left(\frac{Cd}{\rho^2 n}\right)^h.$$

Equivalently, separating the tree-like part from the excess-edge part,

$$\sum_{\substack{|\alpha|=d \\ 1 \in V(\alpha) \\ \alpha \text{ connected}}} \kappa_\alpha^2 \le \rho^2 \sum_{h=0}^{d} \left(Cd(d+1)^2 \frac{\lambda^2}{n}\right)^{d-h} \left(Cd^2(d+1)^2 \frac{\lambda^2}{k^2}\right)^h.$$

Therefore

$$\mathrm{corr}^2_{\le D} \le \rho^2 + \rho^2 \sum_{d=1}^{D} \sum_{h=0}^{d} \left(CD(D+1)^2 \frac{\lambda^2}{n} \right)^{d-h} \left(CD^2(D+1)^2 \frac{\lambda^2}{k^2} \right)^{h}.$$

Assume now that

$$\lambda \le n^{-\varepsilon} \min\left\{k, \sqrt{n}\right\}$$

for some fixed $\varepsilon > 0$, and let

$$D \le n^{\delta}$$

for $\delta > 0$ sufficiently small depending on ε. Then

$$CD(D+1)^2 \frac{\lambda^2}{n} = o(1)$$

and

$$CD^2(D+1)^2 \frac{\lambda^2}{k^2} = o(1).$$

Consequently, the double sum is $o(1)$, and we obtain

$$\mathrm{corr}^2_{\le D} \le \rho^2(1 + o(1)).$$

Since the constant estimator already achieves correlation

$$\mathbb{E}[x_1] = \rho,$$

this bound says that low-degree polynomials do not improve over the trivial prior correlation. Therefore

$$\mathrm{MMSE}_{\le D} = \mathbb{E}[x_1^2] - \mathrm{corr}^2_{\le D} \ge \rho - \rho^2(1 + o(1)).$$

This matches the risk of the prior-only estimator $A(Y) \equiv \rho$ up to lower-order terms. Thus, in the regime

$$\lambda \le n^{-\varepsilon} \min\left\{k, \sqrt{n}\right\},$$

degree-D polynomial estimators cannot achieve nontrivial approximate recovery.

14.6 Discussion and transition

While the method to bound the low-degree MMSE presented in this chapter has been fully based on the very nice "cumulant" idea from [SW22], other very interesting approaches have been suggested recently, including a more direct analysis of the low-degree correlation as in [SW25], an expansion of the correlation on an "almost-orthonormal" basis as in [CGG+25], or a recent reduction approach between the low-degree MMSE and the (much more well-understood) low-degree lower bounds for detection [Li26].

Of course, the low-degree framework for estimation is a proxy for computational hardness, and not a formal lower bound against all polynomial-time algorithms. A natural continuation is therefore to understand when low-degree estimation lower bounds can be connected to other computational hardness approaches for estimation. In detection, as we discussed this relationship has been developed for instanse through the low-degree likelihood-ratio method and its connections with SQ algorithms and average-case reductions. For recovery problems, less is known but a recent work builds an equivalence

between the low-degree MMSE and the Franz-Parisi potential monotonicity criterion for a broad class of Gaussian additive models [TWZ26].

The next chapters study a different algorithmic paradigm for approximate recovery: posterior sampling by Markov chain Monte Carlo. Whereas low-degree lower bounds give evidence against broad classes of polynomial-time estimators, MCMC lower bounds focus on concrete sampling dynamics. Importantly, local dynamics which when run for polynomial-time are of high degree, i.e., at least naively they *cannot* be approximated by low-degree polynomials.

Chapter 15

MCMC Methods: The Basics

15.1 Introduction

In the previous chapter, we studied approximate recovery through the lens of low-degree estimators. The starting point was the Bayesian estimation problem

$$\theta \sim \mu, \qquad Y \sim \mathbb{P}_\theta,$$

where the goal is to construct an estimator $\mathcal{A}(Y)$ with small mean-squared error

$$\mathrm{MSE}(\mathcal{A}) = \mathop{\mathbb{E}}_{\theta\sim\mu,\, Y\sim\mathbb{P}_\theta} \left[\|\mathcal{A}(Y) - \theta\|_2^2\right].$$

The statistically optimal estimator is the posterior mean

$$\mathop{\mathbb{E}}_{\theta\sim\mu}[\theta \mid Y] = \sum_{\theta'\in S} \theta' \,\mathbb{P}\left(\theta' \mid Y\right),$$

where $S = \mathrm{supp}(\mu)$.

The computational difficulty is that, as in many high-dimensional problems, the support S can be very large. Thus, while the posterior mean is statistically optimal, computing it exactly may require summing over exponentially many possible parameters.

This chapter studies a natural computational strategy for avoiding this summation: sampling from the posterior. If we could efficiently sample

$$\theta'_1, \ldots, \theta'_m \overset{\text{i.i.d.}}{\sim} \mathbb{P}(\theta \mid Y),$$

then the empirical average

$$\frac{1}{m}\sum_{i=1}^{m} \theta'_i$$

would approximate the posterior mean. Markov chain Monte Carlo methods (MCMC) provide a general mechanism for sampling from a distribution known only up to normalization. The central question is whether the relevant Markov chains mix in polynomial time. In this chapter, we discuss methods to understand their mixing time.

15.2 Posterior Sampling and Estimation

Let

$$\theta \sim \mu, \qquad Y \sim \mathbb{P}_\theta.$$

Given Y, the posterior distribution is

$$\mathbb{P}\left(\theta' = v \mid Y\right) = \frac{\mu(v)\mathbb{P}(Y \mid v)}{\sum_{u \in S} \mu(u)\mathbb{P}(Y \mid u)}.$$

The denominator is the same normalization obstacle that appears in the posterior mean.

However, if we can somehow sample from the posterior, then we can estimate the posterior mean without computing the normalizing constant explicitly. Let

$$\theta'_1, \ldots, \theta'_m \overset{\text{i.i.d.}}{\sim} \mathbb{P}(\theta \mid Y),$$

and define

$$\widehat{\theta}(Y) := \frac{1}{m}\sum_{i=1}^{m} \theta'_i.$$

Conditional on Y,

$$\mathbb{E}\left[\widehat{\theta}(Y) \mid Y\right] = \mathbb{E}[\theta \mid Y].$$

Moreover, by the law of large numbers,

$$\widehat{\theta}(Y) \overset{m\to\infty}{\to} \mathbb{E}[\theta \mid Y].$$

Let us check explicitly how this affects the MSE. Write

$$m(Y) := \mathbb{E}[\theta \mid Y].$$

Then

$$\widehat{\theta}(Y) - \theta = \widehat{\theta}(Y) - m(Y) + m(Y) - \theta.$$

Therefore

$$\left\|\widehat{\theta}(Y) - \theta\right\|_2^2 = \left\|\widehat{\theta}(Y) - m(Y)\right\|_2^2 + \|m(Y) - \theta\|_2^2 + 2\left\langle \widehat{\theta}(Y) - m(Y), m(Y) - \theta \right\rangle.$$

Taking expectation, the cross term vanishes. Indeed,

$$\mathbb{E}\left[\left\langle \widehat{\theta}(Y) - m(Y), m(Y) - \theta \right\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle \widehat{\theta}(Y) - m(Y), m(Y) - \theta \right\rangle \mid Y, \theta\right]\right].$$

Conditional on Y, the samples θ'_i are independent posterior samples, so

$$\mathbb{E}\left[\widehat{\theta}(Y) - m(Y) \mid Y\right] = 0.$$

Hence the cross term is zero. Thus

$$\mathrm{MSE}\left(\widehat{\theta}\right) = \mathbb{E}\left[\left\|\widehat{\theta}(Y) - m(Y)\right\|_2^2\right] + \mathbb{E}\left[\|m(Y) - \theta\|_2^2\right].$$

The second term is the MMSE. The first term is the sampling error, which goes to zero as m grows. Therefore, efficient posterior sampling would give an efficient approximate recovery algorithm.

15.3 Sampling and MAP Estimation

Sampling can also be used for exact estimation. Exact recovery is typically formulated through the maximum a posteriori estimator

$$\mathcal{A}_{\mathrm{MAP}}(Y) = \arg\max_{\theta' \in S} \mathbb{P}\left(\theta' \mid Y\right).$$

If the posterior has a unique maximizer and a nontrivial gap, then sampling from a sufficiently low-temperature version of the posterior recovers the MAP estimator with high probability.

For $\beta > 0$, define the tilted posterior

$$\pi_\beta(\theta') = \frac{\mathbb{P}(\theta' \mid Y)^\beta}{\sum_{u \in S} \mathbb{P}(u \mid Y)^\beta}.$$

Large β amplifies posterior differences. The following elementary lemma makes this precise.

Lemma 15.1. *Assume that θ_{MAP} is the unique posterior maximizer and that*

$$\mathbb{P}(\theta_{\mathrm{MAP}} \mid Y) \geq 1.1\, \mathbb{P}(\theta' \mid Y) \qquad \text{for every } \theta' \neq \theta_{\mathrm{MAP}}.$$

Then, for $\beta \geq C \log |S|$ with C large enough,

$$\pi_\beta(\theta_{\mathrm{MAP}}) \geq 0.99.$$

Proof. Let

$$p_* := \mathbb{P}(\theta_{\mathrm{MAP}} \mid Y).$$

For every $\theta' \neq \theta_{\mathrm{MAP}}$, the assumption gives

$$\mathbb{P}(\theta' \mid Y) \leq \frac{p_*}{1.1}.$$

Therefore

$$\sum_{\theta' \in S} \mathbb{P}\left(\theta' \mid Y\right)^\beta \leq p_*^\beta + (|S| - 1)\left(\frac{p_*}{1.1}\right)^\beta = p_*^\beta \left(1 + (|S| - 1) 1.1^{-\beta}\right).$$

Hence

$$\pi_\beta(\theta_{\mathrm{MAP}}) = \frac{p_*^\beta}{\sum_{\theta'} \mathbb{P}(\theta' \mid Y)^\beta} \geq \frac{1}{1 + (|S| - 1) 1.1^{-\beta}}.$$

Thus it is enough to choose β so that

$$(|S| - 1) 1.1^{-\beta} \leq \frac{1}{99}.$$

This holds whenever

$$\beta \geq \frac{\log(99|S|)}{\log 1.1}.$$

Thus $\beta \geq C \log |S|$ for sufficiently large C implies

$$\pi_\beta(\theta_{\mathrm{MAP}}) \geq 0.99.$$

□

Remark 15.2. *Clearly, 1.1 is an illustrative constant. The same result follows for $\mathbb{P}(\theta_{\mathrm{MAP}} \mid Y) \geq (1 + \delta)\mathbb{P}(\theta' \mid Y)$ for every $\theta' \neq \theta_{\mathrm{MAP}}$, $\delta > 0$.*

This low-temperature distribution is closely related to *hill climbing*. If we want to maximize a function F over a discrete state space S, then the distribution

$$\pi_\beta(v) \propto e^{\beta F(v)}$$

puts increasingly large mass on maximizers of F as β grows. As we will see, a Metropolis chain targeting π_β behaves like a noisy hill-climbing algorithm: moves that increase F are always accepted, while moves that decrease F are accepted with probability exponentially small in the energy loss.

15.4 Markov Chains

Now that sampling is well-motivated, we turn to the performance of a natural MCMC strategy.

A discrete-time Markov chain on a finite state space Ω is a stochastic process

$$X_0, X_1, X_2, \dots$$

with transition matrix

$$P(a,b) := \mathbb{P}(X_{t+1} = b \mid X_t = a).$$

The chain is homogeneous if P does not depend on t.

Definition 15.3. *A probability distribution π on Ω is stationary for P if*

$$\pi(b) = \sum_{a \in \Omega} \pi(a) P(a,b) \qquad \text{for every } b \in \Omega.$$

Equivalently, viewing π as a row vector,

$$\pi P = \pi.$$

A very useful sufficient condition for stationarity is detailed balance.

Definition 15.4. *The transition matrix P is reversible with respect to π if*

$$\pi(a)P(a,b) = \pi(b)P(b,a) \qquad \text{for every } a, b \in \Omega.$$

Lemma 15.5. *If P satisfies detailed balance with respect to π, then π is stationary.*

Proof. For every $b \in \Omega$,

$$\sum_{a \in \Omega} \pi(a)P(a,b) = \sum_{a \in \Omega} \pi(b)P(b,a) = \pi(b) \sum_{a \in \Omega} P(b,a) = \pi(b),$$

because the rows of P sum to one. □

We also need the standard irreducibility and aperiodicity assumptions.

Definition 15.6. *The chain is irreducible if for every $a, b \in \Omega$ there exists $t \geq 1$ such that*

$$P^t(a,b) > 0.$$

Definition 15.7. *For $x \in \Omega$, define*

$$T(x) := \left\{ t \geq 1 : P^t(x,x) > 0 \right\}.$$

The period of x is $\gcd T(x)$. The chain is aperiodic if every state has period 1.

Theorem 15.8. *If P is irreducible and aperiodic on a finite state space, then there exists a unique stationary distribution π. Moreover, for every initial state $x_0 \in \Omega$,*

$$\lim_{t\to\infty} \mathrm{TV}\left(P^t(x_0,\cdot),\pi\right) = 0.$$

Proof. See [LPW06, Theorem 4.9.]. □

This theorem says that the chain eventually samples from π. For algorithms, however, the important question is the rate of convergence.

Definition 15.9. *The mixing time is*

$$t_{\mathrm{mix}} := \max_{x_0\in\Omega} \inf\left\{t \geq 0 : \mathrm{TV}(P^t(x_0,\cdot),\pi) \leq \frac{1}{4}\right\}.$$

Polynomial-time MCMC requires t_{mix} to be polynomial in the problem dimension.

15.4.1 Terminology from statistical physics

Before introducing the Metropolis process, let us briefly explain the terminology that will be used throughout the chapter. Much of the language of MCMC comes from statistical physics, where one studies probability distributions of the form

$$\pi_\beta(v) = \frac{e^{\beta F(v)}}{Z_\beta}, \qquad Z_\beta := \sum_{u\in\Omega} e^{\beta F(u)}.$$

Here $F(v)$ is often called the *energy* or *Hamiltonian*, although in physics one more commonly writes the Gibbs measure as $e^{-\beta H(v)}$, where H is the physical energy. In our convention, we are maximizing F, so $F = -H$ relative to the usual physics notation.

The parameter $\beta > 0$ is called the *inverse temperature*. Thus:

$$\text{large } \beta \quad\Longleftrightarrow\quad \text{low temperature},$$

and

$$\text{small } \beta \quad\Longleftrightarrow\quad \text{high temperature}.$$

When β is *large*, the measure π_β strongly favors states with large value of $F(v)$. In the limit $\beta \to \infty$, it concentrates on the maximizers of F. This is why large-β MCMC resembles hill climbing.

When β is *small*, the measure is flatter and does not only depend on the best value of F in a region. The number of states in that region also matters. This counting contribution is called *entropy*. Thus, at high temperature, a region with many moderately good states may have more mass than a region with very few excellent states.

15.5 The Metropolis Process

The Metropolis process, which we introduce here, originated in the work [MRR+53] and was later generalized in [Has70].

Let π be a probability distribution on Ω of the form

$$\pi(v) = \frac{e^{F(v)}}{\sum_{u\in\Omega} e^{F(u)}}.$$

The key point is that the Metropolis algorithm only needs ratios

$$\frac{\pi(b)}{\pi(a)} = e^{F(b)-F(a)},$$

so it does not require knowing the normalizing constant.

Let Ψ be a base Markov chain on Ω with symmetric transitions:

$$\Psi(a,b) = \Psi(b,a).$$

For $b \neq a$, define

$$P(a,b) = \Psi(a,b) \min\left\{1, \frac{\pi(b)}{\pi(a)}\right\}.$$

Then set

$$P(a,a) = 1 - \sum_{b \neq a} P(a,b).$$

This is the Metropolis chain associated with Ψ and target distribution π.

Claim 15.10. *The distribution π is stationary for the Metropolis chain.*

Proof. It is enough to prove detailed balance. For $a \neq b$,

$$\pi(a)P(a,b) = \pi(a)\Psi(a,b) \min\left\{1, \frac{\pi(b)}{\pi(a)}\right\} = \Psi(a,b) \min\left\{\pi(a), \pi(b)\right\}.$$

Using symmetry of the proposal chain,

$$\Psi(a,b) \min\left\{\pi(a), \pi(b)\right\} = \Psi(b,a) \min\left\{\pi(b), \pi(a)\right\} = \pi(b)P(b,a).$$

Thus detailed balance holds for all $a \neq b$. The diagonal case is automatic. Therefore π is stationary. □

In Bayesian estimation, the target distribution is usually the posterior

$$\pi(v) = \mathbb{P}(v \mid Y) \propto \mu(v)\mathbb{P}(Y \mid v).$$

Since ratios of posterior probabilities can often be computed efficiently, the Metropolis process gives a natural sampling algorithm.

15.6 Example: Sparse PCA

Consider nonnegative sparse PCA:

$$Y = \lambda\theta\theta^\top + W,$$

where

$$\theta \in \Omega := \left\{v \in \left\{0, \frac{1}{\sqrt{k}}\right\}^n : \|v\|_0 = k\right\},$$

and

$$W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

The prior is uniform on Ω.

For a candidate $v \in \Omega$, the posterior is proportional to

$$\exp\left(-\frac{1}{2}\left\|Y - \lambda vv^\top\right\|_F^2\right).$$

Since $\left\|vv^\top\right\|_F^2 = \|v\|_2^4 = 1$ is constant over $v \in \Omega$, the posterior satisfies

$$\mathbb{P}(v \mid Y) \propto \exp\left(\lambda v^\top Y v\right).$$

Thus the posterior is

$$\pi_\lambda(v) \propto e^{\lambda v^\top Y v}.$$

For exact recovery, one can consider instead the low-temperature distribution

$$\pi_\beta(v) \propto e^{\beta v^\top Y v}.$$

As $\beta \to \infty$, this concentrates on maximizers of $v^\top Y v$, i.e. on the MAP estimator

$$\mathcal{A}_{\mathrm{MAP}}(Y) = \arg\max_{v \in \Omega} v^\top Y v.$$

A natural proposal chain on Ω swaps one coordinate inside the support with one outside the support. Thus v' is a neighbor of v if

$$d_H\left(v, v'\right) = 2.$$

There are $k(n-k)$ such neighbors. The base chain is

$$\Psi\left(v, v'\right) = \frac{1}{k(n-k)} \qquad \text{if } d_H\left(v, v'\right) = 2.$$

The Metropolis chain is therefore

$$P\left(v, v'\right) = \frac{1}{k(n-k)} \min\left\{1, \exp\left(\beta\left(\left(v'\right)^\top Y v' - v^\top Y v\right)\right)\right\}$$

for neighboring v, v', with the diagonal probability chosen so that rows sum to one.

15.7 Conclusion and transition

The purpose of this chapter was to explain why Markov chain Monte Carlo methods arise naturally in Bayesian estimation problems. The statistically optimal estimator for squared error loss is the posterior mean

$$\mathbb{E}[\theta \mid Y],$$

but computing this expectation directly may require summing over an exponentially large parameter space. Posterior sampling offers a way around this obstacle: if one can sample efficiently from the posterior distribution, then empirical averages of posterior samples approximate the posterior mean.

We then introduced the Metropolis process as a general method for sampling from a distribution known only up to normalization. This is particularly well suited to Bayesian inference, because posterior ratios are often much easier to compute than the posterior normalizing constant. In the sparse PCA example, the posterior distribution takes the Gibbs form

$$\pi_\lambda(v) \propto e^{\lambda v^\top Y v},$$

and its low-temperature variants are

$$\pi_\beta(v) \propto e^{\beta v^\top Y v}.$$

The natural Metropolis chain proposes local support swaps and accepts or rejects them according to the change in the objective $v^\top Y v$.

The central question is therefore no longer whether the posterior is statistically informative, but whether the Markov chain can sample from it efficiently. The next chapter is devoted to lower bounds on this mixing time. We will see that in the hard regimes of sparse PCA and planted clique, the relevant Gibbs landscapes contain bottlenecks: regions with substantial stationary mass but very small boundary mass. These bottlenecks force local Markov chains to mix slowly. For a deeper dive on Markov chain theory we direct the reader to classical textbooks, such as [LPW06].

Chapter 16

MCMC Lower Bounds: Overlap Gap Property and Bottlenecks

16.1 Introduction

The previous chapter introduced posterior sampling as a computational route to Bayesian estimation and described the Metropolis process as a natural algorithm for sampling from posterior-type distributions. We now study when this strategy can be proven to fail, providing further evidence of hardness in hard regimes.

The main obstruction is geometric. A Markov chain may have the correct stationary distribution and may converge to it eventually, but convergence can take a very long time if the state space contains a bottleneck. Informally, a bottleneck is a set A that has non-negligible stationary probability but whose boundary has extremely small stationary probability. If the chain starts inside A, it is unlikely to leave quickly. Consequently, the chain cannot mix rapidly.

This chapter develops this bottleneck method and applies it to the sparse PCA Metropolis chain from the previous chapter. The state space is

$$\Omega = \left\{ v \in \left\{ 0, \frac{1}{\sqrt{k}} \right\}^n : \|v\|_0 = k \right\},$$

and the Gibbs distribution is

$$\pi_\beta(v) \propto e^{\beta v^\top Y v}.$$

The sparse PCA analysis has two complementary parts. At low temperature, the bottleneck comes from an overlap-gap property, a geometric obstruction that has played a central role in the study of algorithmic barriers for estimation problems; see, for instance, [GZ22] for sparse high-dimensional linear regression and [AWZ23] for sparse PCA. Informally, the chain must move from low-overlap states to states correlated with the planted vector, but any local path between these regions has to cross intermediate overlap levels where even the best available energy is significantly lower. At higher temperature, the bottleneck is entropic: there are overwhelmingly many low-overlap states and very few states with significant overlap with the truth.

After sparse PCA, we briefly discuss the analogous picture for planted clique. There, the Metropolis chain runs over cliques of varying sizes, and bottlenecks arise from the geometry of the clique-overlap profile. This shows that MCMC lower bounds provide a different kind of computational evidence from low-degree or SQ lower bounds: they identify concrete geometric obstructions faced by natural local sampling algorithms.

16.2 Bottlenecks and Mixing Lower Bounds

The main lower-bound technique is to show that the Markov chain has a bottleneck: a set with non-negligible stationary mass but very small probability flow leaving it.

Definition 16.1. *Let P be a Markov chain with stationary distribution π. For $A \subseteq \Omega$, define its* conductance ratio *or* escape probability

$$\Phi(A) := \frac{\sum_{a \in A,\, b \notin A} \pi(a) P(a,b)}{\pi(A)}.$$

We say that A is a T-bottleneck if

$$\pi(A) \le \frac{1}{2}$$

and

$$\Phi(A) \le \frac{1}{T}.$$

Theorem 16.2. *If a Markov chain has a T-bottleneck, then*

$$t_{\mathrm{mix}} \ge \frac{T}{4}.$$

Proof. See [LPW06, Theorem 7.3]. □

Constants are irrelevant for our purposes. If $T = n^{\omega(1)}$, then the chain does not mix in polynomial time.

16.2.1 A toy bottleneck

Consider a graph on 2^n vertices made of two large pieces, each of size 2^{n-1}, connected by a single edge. Suppose the graph is approximately Δ-regular, and consider the simple random walk with

$$P(a,b) = \frac{1}{\Delta}$$

for neighboring vertices. The stationary distribution is uniform.

Let A be one of the two halves. Then

$$\pi(A) = \frac{1}{2}.$$

There is only one edge crossing from A to A^c, so

$$\sum_{a \in A,\, b \notin A} \pi(a) P(a,b) = \frac{1}{2^n} \cdot \frac{1}{\Delta}.$$

Therefore

$$\Phi(A) = \frac{2}{\Delta 2^n}.$$

The corresponding bottleneck scale is exponential in n. The walk mixes slowly because it spends an exponentially long time in one half before finding the unique bridge.

This toy example, depicted in Figure 16.1, explains the general principle: fast mixing requires every large set to have enough *boundary*.

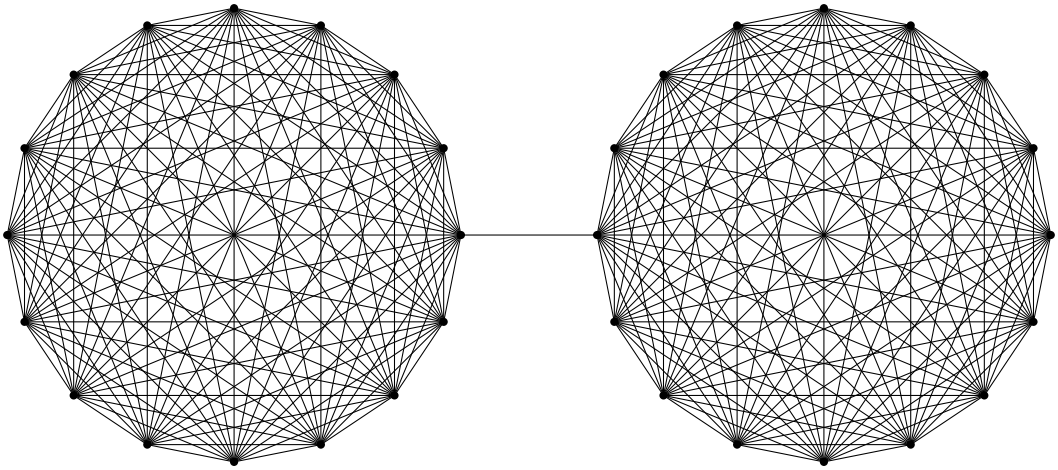

Figure 16.1: Two copies of K_{16} joined by one edge.

16.3 Bottlenecks for Sparse PCA

For sparse PCA, the relevant state space is

$$\Omega = \left\{ v \in \left\{ 0, \frac{1}{\sqrt{k}} \right\}^n : \|v\|_0 = k \right\}.$$

Neighbors differ by one support swap. For $A \subseteq \Omega$, define the vertex boundary

$$\partial A := \{ v \in A : \exists w \notin A \text{ with } d_H(v, w) = 2 \}.$$

For the Metropolis chain,

$$\Phi(A) = \frac{1}{\pi_\beta(A)} \sum_{a \in A,\, b \notin A} \pi_\beta(a) P(a, b).$$

Since the maximum degree of the proposal graph is $k(n-k)$ and each transition probability is at most $1/(k(n-k))$, we have

$$\Phi(A) \leq \frac{\pi_\beta(\partial A)}{\pi_\beta(A)}.$$

Thus, to prove slow mixing, it is enough to find a set A with

$$\pi_\beta(A) \leq \frac{1}{2}$$

and

$$\frac{\pi_\beta(\partial A)}{\pi_\beta(A)} \ll \frac{1}{\mathrm{poly}(n)}.$$

The rest of the sparse PCA analysis is devoted to constructing such a set.

16.4 The Overlap Gap Property for Estimation tasks

Let θ be the true spike. For $v \in \Omega$, the overlap with θ takes values

$$\langle v, \theta \rangle \in \left\{ 0, \frac{1}{k}, \frac{2}{k}, \ldots, 1 \right\}.$$

Indeed, $\langle v, \theta \rangle = \ell/k$ if the support of v intersects the support of θ in exactly ℓ coordinates.

Define the energy

$$H_Y(v) := v^\top Y v.$$

For $\ell = 0, 1, \ldots, k$, define

$$\Gamma(\ell) := \max_{v \in \Omega:\ \langle v, \theta \rangle = \ell/k} H_Y(v).$$

Then

$$\max_{v \in \Omega} H_Y(v) = \max_{0 \le \ell \le k} \Gamma(\ell).$$

In the exact-recovery regime $\lambda \gg \sqrt{k}$, the MAP estimator is θ with high probability. Equivalently,

$$\arg\max_{0 \le \ell \le k} \Gamma(\ell) = k$$

with high probability.

The overlap gap property is the existence of a *hard* overlap level: an intermediate overlap ℓ where the best energy is significantly worse than the best energy at overlap 0. This creates a valley between a low-overlap region and the planted region. This framework for Markov chain lower bounds has been originally developed in [GZ22; GZ24].

Theorem 16.3 (Overlap gap implies a bottleneck)**.** *Assume that for some $\ell \in \{0, \ldots, k-1\}$ and some $\eta > 0$,*

$$\Gamma(\ell) \le \Gamma(0) - \eta.$$

Assume also that for β large enough that mass of π_β is concentrated near the MAP region, so that

$$\pi_\beta(\{v : \langle v, \theta \rangle \le \ell/k\}) \le \frac{1}{2}.$$

Then, if

$$\beta\eta \gg k \log n,$$

the Metropolis chain for π_β has a superpolynomial bottleneck.

Proof. Let

$$A := \{v \in \Omega : \langle v, \theta \rangle \le \ell/k\}.$$

A single support-swap changes the overlap by at most $1/k$. Therefore any path from A to A^c must pass through states with overlap exactly ℓ/k. Hence

$$\partial A \subseteq \{v \in \Omega : \langle v, \theta \rangle = \ell/k\}.$$

We bound

$$\frac{\pi_\beta(\partial A)}{\pi_\beta(A)} = \frac{\sum_{v \in \partial A} e^{\beta H_Y(v)}}{\sum_{v \in A} e^{\beta H_Y(v)}}.$$

The numerator is bounded by

$$\sum_{v \in \partial A} e^{\beta H_Y(v)} \le |\Omega| e^{\beta \Gamma(\ell)}.$$

For the denominator, since A contains all states of overlap 0, there exists a state with energy $\Gamma(0)$, so

$$\sum_{v \in A} e^{\beta H_Y(v)} \ge e^{\beta \Gamma(0)}.$$

Therefore

$$\frac{\pi_\beta(\partial A)}{\pi_\beta(A)} \le |\Omega| e^{\beta(\Gamma(\ell)-\Gamma(0))} \le \binom{n}{k} e^{-\beta\eta}.$$

Using

$$\binom{n}{k} \le \left(\frac{en}{k}\right)^k \le e^{k\log n},$$

for n large enough, we get

$$\frac{\pi_\beta(\partial A)}{\pi_\beta(A)} \le \exp(k\log n - \beta\eta).$$

If $\beta\eta \gg k\log n$, this is $1/\mathrm{poly}(n)$, or smaller. Thus A is a bottleneck, and the mixing time is superpolynomial. □

16.5 Overlap Gap Property for Estimation in Sparse PCA

We now show that the overlap gap exists in the hard sparse PCA regime. This chapter follows the work [AWZ23].

Assume

$$\sqrt{k} \ll \lambda \ll k, \qquad k \ll \sqrt{n}.$$

The first inequality places us above the information-theoretic threshold for exact recovery, while the second says that we remain below the conjectures efficient recovery threshold.

For v with overlap ℓ/k with θ,

$$H_Y(v) = v^\top \left(\lambda\theta\theta^\top + W\right) v = \lambda \langle v,\theta\rangle^2 + v^\top W v = \lambda\frac{\ell^2}{k^2} + v^\top W v.$$

Let

$$N_\ell := |\{v \in \Omega : \langle v,\theta\rangle = \ell/k\}| = \binom{k}{\ell}\binom{n-k}{k-\ell}.$$

For a fixed v, the random variable $v^\top W v$ is a centered Gaussian with variance of order one. Hence the maximum over N_ℓ such variables is heuristically of order $\sqrt{2\log N_\ell}$, as long as $\ell \ll k$ (and hence the variables have low enough correlation). This gives the heuristic

$$\Gamma(\ell) \approx \lambda\frac{\ell^2}{k^2} + \sqrt{2\log N_\ell}.$$

At overlap 0,

$$N_0 = \binom{n-k}{k},$$

so

$$\Gamma(0) \approx \sqrt{2\log\binom{n-k}{k}}.$$

The important point is that N_ℓ is much smaller than N_0 when $k \ll \sqrt{n}$ and $\ell > 0$. Indeed,

$$\frac{N_\ell}{N_0} = \binom{k}{\ell}\frac{\binom{n-k}{k-\ell}}{\binom{n-k}{k}}.$$

For ℓ not too large,

$$\binom{k}{\ell}\frac{\binom{n-k}{k-\ell}}{\binom{n-k}{k}} \approx \frac{1}{\ell!}\left(\frac{k^2}{n}\right)^\ell.$$

Since $k^2/n \ll 1$, this ratio decays exponentially in ℓ.

Thus increasing the overlap has two competing effects:

$$\text{signal gain:} \quad \lambda \frac{\ell^2}{k^2},$$

but

$$\text{entropy loss:} \quad \sqrt{2\log N_\ell} - \sqrt{2\log N_0} < 0.$$

At a suitable intermediate ℓ, the entropy loss dominates the signal gain, creating an overlap gap. The above heuristic picture leads to the following theorem.

Theorem 16.4. *Assume*

$$\sqrt{k} \ll \lambda \ll k$$

and $k \ll \sqrt{n}$. Then, with high probability, there exists ℓ such that

$$\Gamma(\ell) \le \Gamma(0) - c\frac{k}{\lambda}\sqrt{\log n}$$

for some constant $c > 0$.

Proof. For fixed ℓ, a union bound for Gaussian maxima gives, with high probability,

$$\Gamma(\ell) \le \lambda \frac{\ell^2}{k^2} + \sqrt{2\log N_\ell} + \mathcal{O}(1).$$

Similarly, by a very delicate second-moment lower bound for sparse Gaussian maxima [AWZ23, Theorem 4.2.],

$$\Gamma(0) \ge \sqrt{2\log N_0} - 100\sqrt{\log n}$$

with high probability.

Therefore

$$\Gamma(\ell) - \Gamma(0) \le \lambda \frac{\ell^2}{k^2} + \sqrt{2\log N_\ell} - \sqrt{2\log N_0} + 100\sqrt{\log n}\mathcal{O}(1).$$

Now

$$\log N_\ell - \log N_0 = \log \binom{k}{\ell} + \log \frac{\binom{n-k}{k-\ell}}{\binom{n-k}{k}}.$$

Using $k \ll \sqrt{n}$ and $\ell \ll k$, this satisfies

$$\log N_\ell - \log N_0 \le -c_1 \ell \log n$$

for some constant $c_1 > 0$, in the polynomial sparsity regime $k = n^\alpha$ with $\alpha < 1/2$. Since

$$\log N_0 = \Theta(k \log n),$$

we obtain

$$\sqrt{2\log N_\ell} - \sqrt{2\log N_0} = \frac{2(\log N_\ell - \log N_0)}{\sqrt{2\log N_\ell} + \sqrt{2\log N_0}} \le -c_2 \ell \sqrt{\frac{\log n}{k}}.$$

Hence

$$\Gamma(\ell) - \Gamma(0) \le \lambda \frac{\ell^2}{k^2} - c_2 \ell \sqrt{\frac{\log n}{k}} + 100\sqrt{\log n} + \mathcal{O}(1).$$

Choose

$$\ell_* = \left\lfloor c_3 \frac{k^{3/2}}{\lambda} \right\rfloor$$

for $c_3 > 0$ sufficiently small. Because $\lambda \gg \sqrt{k}$, we have $\ell_* \ll k$, and because $\lambda \ll k$, we have $\ell_* \to \infty$. Substituting this value gives

$$\lambda \frac{\ell_*^2}{k^2} = \mathcal{O}\left(\frac{k}{\lambda}\right),$$

while

$$\ell_* \sqrt{\frac{\log n}{k}} = \Theta\left(\frac{k}{\lambda}\sqrt{\log n}\right).$$

Since $\sqrt{\log n} \to \infty$, the negative entropy term dominates the positive signal term. Moreover, because $\lambda \ll k$,

$$\frac{k}{\lambda}\sqrt{\log n} \gg \sqrt{\log n}.$$

Thus the additive $100\sqrt{\log n}$ term is also dominated. Therefore, for $\ell = \ell_*$,

$$\Gamma(\ell) - \Gamma(0) \le -c\frac{k}{\lambda}\sqrt{\log n}$$

with high probability. □

Combining this theorem with the bottleneck theorem yields slow mixing for sufficiently large β. In particular, taking

$$\eta = \Theta\left(\frac{k}{\lambda}\sqrt{\log n}\right),$$

the condition

$$\beta\eta \gg k \log n$$

becomes

$$\beta \gg \lambda\sqrt{\log n}.$$

Up to logarithmic factors, this says that low-temperature MCMC has slow mixing throughout the regime

$$\sqrt{k} \ll \lambda \ll k.$$

16.5.1 Why the overlap profile controls the large-β regime

The previous argument was based on the function

$$\Gamma(\ell) := \max_{v \in \Omega:\ \langle v, \theta\rangle = \ell/k} v^\top Y v.$$

Let us pause to explain why this function is the right object to study when β is large.

Recall that the Gibbs measure is

$$\pi_\beta(v) = \frac{e^{\beta v^\top Y v}}{Z_\beta}, \qquad Z_\beta := \sum_{u \in \Omega} e^{\beta u^\top Y u}.$$

For each overlap level ℓ, define

$$\Omega_\ell := \left\{ v \in \Omega : \langle v, \theta\rangle = \frac{\ell}{k} \right\}.$$

Then

$$\pi_\beta(\Omega_\ell) = \frac{1}{Z_\beta} \sum_{v \in \Omega_\ell} e^{\beta v^\top Y v}.$$

By definition of $\Gamma(\ell)$, we always have

$$e^{\beta\Gamma(\ell)} \le \sum_{v\in\Omega_\ell} e^{\beta v^\top Y v} \le |\Omega_\ell| e^{\beta\Gamma(\ell)}.$$

Therefore

$$\frac{e^{\beta\Gamma(\ell)}}{Z_\beta} \le \pi_\beta(\Omega_\ell) \le \frac{|\Omega_\ell| e^{\beta\Gamma(\ell)}}{Z_\beta}.$$

Since

$$|\Omega_\ell| \le |\Omega| = \binom{n}{k},$$

we get the rough but useful estimate

$$\pi_\beta(\Omega_\ell) \in \left[\frac{e^{\beta\Gamma(\ell)}}{Z_\beta}, \frac{\binom{n}{k} e^{\beta\Gamma(\ell)}}{Z_\beta}\right].$$

Equivalently,

$$\frac{1}{\beta}\log \pi_\beta(\Omega_\ell) = \Gamma(\ell) - \frac{1}{\beta}\log Z_\beta + \mathcal{O}\left(\frac{k\log(en/k)}{\beta}\right).$$

Thus, when

$$\beta \gg k\log\frac{en}{k},$$

up to the relevant energy scale, the Gibbs mass of an overlap level is governed mainly by $\Gamma(\ell)$. In words, for large β, the measure π_β is very *peaky*: it gives most of its mass to configurations with nearly maximal value of $v^\top Y v$ inside each overlap level.

This explains why a non-monotonicity in the overlap profile creates a bottleneck. Suppose, for instance, that for some ℓ,

$$\Gamma(\ell) \le \Gamma(0) - \eta.$$

Then

$$\pi_\beta(\Omega_\ell) \le \binom{n}{k} e^{-\beta\eta} \frac{e^{\beta\Gamma(0)}}{Z_\beta}.$$

Since

$$\frac{e^{\beta\Gamma(0)}}{Z_\beta} \le \pi_\beta(\Omega_0),$$

this shows that the mass of the intermediate overlap level is exponentially smaller than the mass of the zero-overlap level whenever

$$\beta\eta \gg k\log\frac{en}{k}.$$

This is precisely the large-β mechanism behind the bottleneck argument.

Applying the overlap-gap estimate from the previous subsection, we have with high probability an overlap level satisfying

$$\Gamma(\ell) \le \Gamma(0) - \eta, \qquad \eta \asymp \frac{k}{\lambda}\sqrt{\log n}.$$

Hence the large-β argument gives slow mixing once

$$\beta\eta \gg k\log n.$$

Equivalently, up to logarithmic factors,

$$\beta \gg \lambda\sqrt{\log n}.$$

An important point is that this does not yet prove slow mixing at the posterior temperature

$$\beta = \lambda.$$

Indeed, the argument above relies on the approximation

$$\pi_\beta(\Omega_\ell) \approx \frac{e^{\beta\Gamma(\ell)}}{Z_\beta},$$

which is accurate only when β is large enough to suppress the entropy of each overlap level. At smaller β, the Gibbs measure is flatter, and the number of configurations at each overlap becomes essential.

This motivates the next section. In the high-temperature regime, the bottleneck is no longer purely energetic; it is entropic. Instead of looking only at the maximum energy $\Gamma(\ell)$, we must also use the fact that there are far fewer configurations with positive overlap with the planted vector.

16.6 A High-Temperature Bottleneck

The previous overlap-gap property argument is most useful when β is large. In that regime, the Gibbs measure

$$\pi_\beta(v) \propto e^{\beta v^\top Y v}$$

is strongly influenced by the best energy at each overlap level. Thus it is natural to approximate the mass of an overlap level by

$$\pi_\beta(\{v : \langle v, \theta\rangle = \ell/k\}) \approx \frac{e^{\beta\Gamma(\ell)}}{Z}.$$

However, this approximation is no longer accurate when β is small. In this regime we describe a different technique presented in [AWZ23], and leveraging a very nice symmetry idea from [AGJ20].

Theorem 16.5. *For $v \in \Omega$, define the overlap*

$$R(v) := k\langle v, \theta\rangle = |\operatorname{supp}(v) \cap \operatorname{supp}(\theta)|.$$

Assume that

$$k \ll \sqrt{n}, \qquad \beta\lambda \ll \frac{k^2}{(\log n)^2}.$$

Then, with high probability over the randomness of the instance, there exists an overlap level $\ell \in \{0, 1, \dots, k\}$ such that, for

$$A_\ell := \{v \in \Omega : R(v) \le \ell\},$$

we have

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \le \exp\left(-c \min\left\{k, \frac{k^2}{\beta\lambda}\right\}\right),$$

up to harmless logarithmic factors.

Consequently, if in addition

$$\pi_\beta(A_\ell) \le \frac{1}{2},$$

then A_ℓ is a bottleneck for the Metropolis chain, and the mixing time is at least

$$\exp\left(c \min\left\{k, \frac{k^2}{\beta\lambda}\right\}\right),$$

again up to logarithmic factors.

To prove this result, we need the following lemma:

Lemma 16.6. *For every fixed $A \subseteq \Omega$,*

$$\mathop{\mathbb{E}}_{W}[\widetilde{\pi}_\beta(A)] = \frac{|A|}{|\Omega|}.$$

Proof. It is enough to prove the claim for a singleton $A = \{v\}$ and then sum over $v \in A$.

Fix $u, v \in \Omega$. There exists a permutation matrix U such that

$$U^\top U = I_n, \qquad Uu = v, \qquad U\Omega = \Omega.$$

Since the Gaussian noise matrix is invariant in distribution under conjugation,

$$U^\top W U \overset{d}{=} W.$$

Therefore

$$\mathop{\mathbb{E}}_{W}\left[\frac{e^{\beta v^\top W v}}{\sum_{z\in\Omega} e^{\beta z^\top W z}}\right] \underset{U\Omega=\Omega}{\overset{Uu=v}{=}} \mathop{\mathbb{E}}_{W}\left[\frac{e^{\beta (Uu)^\top W U u}}{\sum_{z\in\Omega} e^{\beta (Uz)^\top W U z}}\right] = \mathop{\mathbb{E}}_{W}\left[\frac{e^{\beta u^\top \left(U^\top W U\right) u}}{\sum_{z\in\Omega} e^{\beta z^\top \left(U^\top W U\right) z}}\right]$$
$$\underset{=}{\overset{U^\top W U \overset{d}{=} W}{=}} \mathop{\mathbb{E}}_{W}\left[\frac{e^{\beta u^\top W u}}{\sum_{z\in\Omega} e^{\beta z^\top W z}}\right].$$

Thus the expectation is the same for every $v \in \Omega$. Since

$$\sum_{v\in\Omega} \widetilde{\pi}_\beta(v) = 1$$

for every W, summing expectations over all v gives

$$|\Omega| \mathop{\mathbb{E}}_{W}[\widetilde{\pi}_\beta(v)] = 1.$$

Hence

$$\mathop{\mathbb{E}}_{W}[\widetilde{\pi}_\beta(v)] = \frac{1}{|\Omega|}.$$

Summing over $v \in A$ proves the claim. □

Proof of Theorem 16.5. Let

$$R(v) := k\langle v, \theta\rangle.$$

Thus $R(v)$ is the number of coordinates shared by the supports of v and θ, and

$$R(v) \in \{0, 1, \dots, k\}.$$

For $\ell = 0, 1, \dots, k$, define

$$G(\ell) := |\{v \in \Omega : R(v) = \ell\}|.$$

Since one must choose ℓ coordinates from the support of θ and $k - \ell$ coordinates outside it,

$$G(\ell) = \binom{k}{\ell}\binom{n-k}{k-\ell}.$$

In particular,

$$\frac{G(\ell+1)}{G(\ell)} = \frac{(k-\ell)^2}{(\ell+1)(n-2k+\ell+1)}.$$

When $k \ll \sqrt{n}$, this ratio is already very small at $\ell = 0$:

$$\frac{G(1)}{G(0)} = \frac{k^2}{n-2k+1} = o(1).$$

Indeed, it is easy to show that under the uniform measure on Ω, the overlap with θ is overwhelmingly concentrated near zero.

For $1 \le \ell \le k$ and $k \ll \sqrt{n}$, the upper tail satisfies

$$\frac{|\{v : R(v) \ge \ell\}|}{|\Omega|} \le 2\left(\frac{Ck^2}{n\ell}\right)^{\ell}$$

for a universal constant $C > 0$.

To see this, notice the ratio formula above gives

$$\frac{G(r)}{G(0)} = \prod_{j=0}^{r-1} \frac{G(j+1)}{G(j)} \le \prod_{j=0}^{r-1} \frac{Ck^2}{(j+1)n} = \frac{1}{r!}\left(\frac{Ck^2}{n}\right)^r.$$

Using $r! \ge (r/e)^r$,

$$\frac{G(r)}{G(0)} \le \left(\frac{Cek^2}{nr}\right)^r.$$

Since $G(0) \le |\Omega|$ and the terms decrease geometrically in the regime $k \ll \sqrt{n}$, summing over $r \ge \ell$ gives

$$\frac{|\{v : R(v) \ge \ell\}|}{|\Omega|} \le 2\left(\frac{Ck^2}{n\ell}\right)^{\ell}.$$

We now compare the planted Gibbs measure with the pure-noise Gibbs measure

$$\widetilde{\pi}_\beta(v) := \frac{e^{\beta v^\top W v}}{\sum_{u\in\Omega} e^{\beta u^\top W u}}.$$

The following symmetry lemma, which can be found in the proof of [AWZ23, Theorem 3.8.], is crucial.

By Markov's inequality, for every fixed $A \subseteq \Omega$,

$$\mathbb{P}_W\left(\widetilde{\pi}_\beta(A) \ge \frac{|A|}{|\Omega|}\log n\right) \le \frac{1}{\log n}.$$

In particular, with probability at least $1 - 1/\log n$,

$$\widetilde{\pi}_\beta(A) \le \frac{|A|}{|\Omega|}\log n.$$

Now fix an overlap level ℓ and define

$$A_\ell := \{v \in \Omega : R(v) \le \ell\}.$$

A single support-swap changes $R(v)$ by at most one, so

$$\partial A_\ell \subseteq \{v : R(v) = \ell\}.$$

We want to bound

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)}.$$

Since

$$Y = \lambda\theta\theta^\top + W,$$

we have

$$v^\top Y v = \lambda\langle v, \theta\rangle^2 + v^\top W v = \lambda\frac{R(v)^2}{k^2} + v^\top W v.$$

Therefore, for $v \in \partial A_\ell$, where $R(v) = \ell$,

$$e^{\beta v^\top Y v} = e^{\beta\lambda\ell^2/k^2} e^{\beta v^\top W v}.$$

Hence

$$\sum_{v\in\partial A_\ell} e^{\beta v^\top Y v} \le e^{\beta\lambda\ell^2/k^2} \sum_{v:R(v)=\ell} e^{\beta v^\top W v}.$$

On the other hand,

$$\sum_{v\in A_\ell} e^{\beta v^\top Y v} \ge \sum_{v\in A_\ell} e^{\beta v^\top W v},$$

because the signal contribution is nonnegative. Thus

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \le e^{\beta\lambda\ell^2/k^2} \frac{\sum_{v:R(v)=\ell} e^{\beta v^\top W v}}{\sum_{v:R(v)\le\ell} e^{\beta v^\top W v}}.$$

Bounding the numerator by the larger set $\{R(v) \ge \ell\}$ gives

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \le e^{\beta\lambda\ell^2/k^2} \frac{\widetilde{\pi}_\beta(\{v : R(v) \ge \ell\})}{1 - \widetilde{\pi}_\beta(\{v : R(v) > \ell\})}.$$

Whenever

$$\widetilde{\pi}_\beta(\{v : R(v) \ge \ell\}) \le \frac{1}{2},$$

we obtain the simpler bound

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \le 2e^{\beta\lambda\ell^2/k^2}\widetilde{\pi}_\beta(\{v : R(v) \ge \ell\}).$$

Using the symmetry lemma and the overlap-counting estimate, with probability at least $1 - 1/\log n$,

$$\widetilde{\pi}_\beta(\{v : R(v) \ge \ell\}) \le 2\left(\frac{Ck^2}{n\ell}\right)^\ell \log n.$$

Consequently,

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \le 4\log n\, \exp\left(\beta\lambda\frac{\ell^2}{k^2} + \ell\log\frac{Ck^2}{n\ell}\right).$$

We now choose ℓ to make the exponent negative. Suppose first that $\beta\lambda \gg k$ and

$$\beta\lambda \ll \frac{k^2}{(\log n)^2}.$$

Set

$$\ell = \left\lfloor \frac{k^2}{2\beta\lambda} \right\rfloor.$$

Then $\ell \to \infty$, and the condition above ensures $\ell \gg (\log n)^2$. Also, since $\beta\lambda \gg k$, we have $\ell \ll k$.

For this choice,

$$\beta\lambda \frac{\ell^2}{k^2} \leq \frac{k^2}{4\beta\lambda}.$$

Meanwhile,

$$\ell \log \frac{n\ell}{Ck^2} = \ell \log \frac{n}{C\beta\lambda} \gg \ell,$$

because $k \ll \sqrt{n}$ and $\beta\lambda \ll k^2/(\log n)^2$. Therefore

$$\beta\lambda \frac{\ell^2}{k^2} + \ell \log \frac{Ck^2}{n\ell} \leq -c\frac{k^2}{\beta\lambda}$$

for some constant $c > 0$. Hence

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \leq \exp\left(-c\frac{k^2}{\beta\lambda}\right)$$

up to logarithmic factors. Since

$$\frac{k^2}{\beta\lambda} \gg (\log n)^2,$$

this is superpolynomially small.

There is also a simpler argument when $\beta\lambda \lesssim k$. In that case, take

$$\ell = \lfloor c_0 k \rfloor$$

for a sufficiently small constant $c_0 > 0$. Then

$$\beta\lambda \frac{\ell^2}{k^2} = \mathcal{O}(k),$$

whereas the entropy term is

$$-\ell \log \frac{n\ell}{Ck^2} \leq -c_1 k \log \frac{n}{k^2}.$$

Since $k \ll \sqrt{n}$, this is negative and dominates the signal term. Thus again

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \leq \exp(-c_2 k)$$

or better.

Therefore the high-temperature mechanism gives, in the relevant regime,

$$\frac{\pi_\beta(\partial A_\ell)}{\pi_\beta(A_\ell)} \leq \exp\left(-c \min\left\{k, \frac{k^2}{\beta\lambda}\right\}\right),$$

ignoring logarithmic factors. □

This should be interpreted as an entropic bottleneck. At low temperature, bottlenecks arise because the energy landscape has an overlap gap. At high temperature, the chain is flatter, but the state space itself is highly imbalanced: there are overwhelmingly many low-overlap states and very few intermediate-overlap states. The chain still has to cross those intermediate-overlap levels in order to approach the planted vector.

16.7 Subexponential-Time Predictions

The two bottleneck mechanisms discussed above are complementary.

First, the low-temperature overlap-gap argument gives an energetic bottleneck. In that regime, the bottleneck comes from the fact that every path from low-overlap states to the planted state must cross an overlap level whose best energy is lower by an amount of order k/λ, up to logarithmic factors. Thus, ignoring logarithmic factors, this gives

$$\frac{\pi_\beta(\partial A)}{\pi_\beta(A)} \lesssim \exp\left(-c\frac{\beta k}{\lambda}\right).$$

Second, the high-temperature argument gives an entropic bottleneck. There, the obstruction is that there are very few configurations with nontrivial overlap with the planted vector. Optimizing the overlap level gives, again ignoring logarithmic factors,

$$\frac{\pi_\beta(\partial A)}{\pi_\beta(A)} \lesssim \exp\left(-c\min\left\{\frac{k^2}{\beta\lambda}, k\right\}\right).$$

The minimum with k reflects the fact that the entropic bottleneck saturates once the chosen overlap level is of order k.

Combining both mechanisms, it can be proven that for every β one expects to be able to choose a bottleneck set A such that

$$\frac{\pi_\beta(\partial A)}{\pi_\beta(A)} \le \exp\left(-c\max\left\{\frac{\beta k}{\lambda}, \min\left\{\frac{k^2}{\beta\lambda}, k\right\}\right\}\right),$$

up to logarithmic factors (see Figure 16.2).

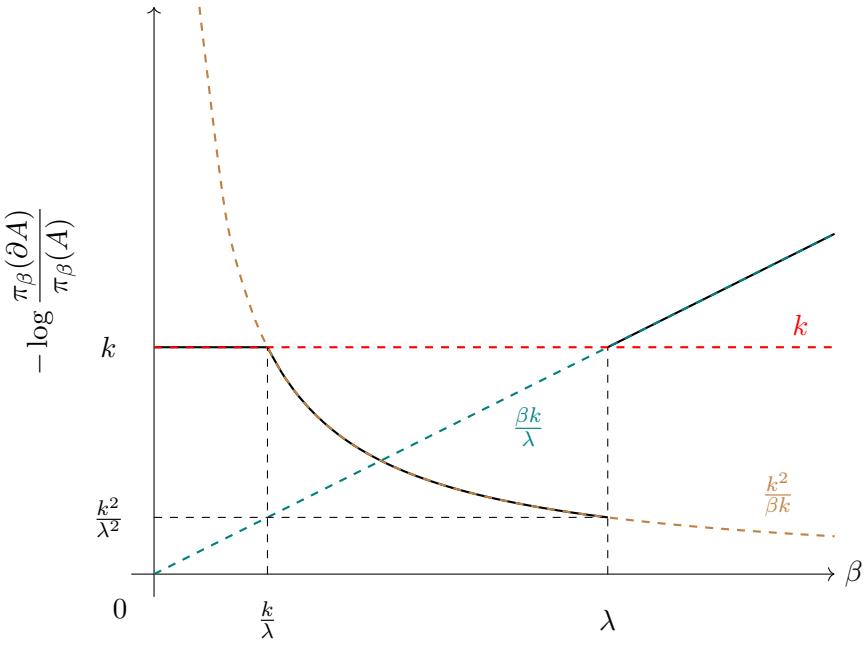

Figure 16.2: Log behavior of bottleneck size of MCMC recovery for Sparse PCA

The weakest lower bound, and hence the conjecturally fastest mixing temperature, occurs at the posterior temperature

$$\beta = \lambda.$$

At this value, the high-temperature contribution gives

$$\min\left\{\frac{k^2}{\lambda^2}, k\right\}.$$

In the regime $\sqrt{k} \ll \lambda \ll k$, this is

$$\frac{k^2}{\lambda^2}.$$

Thus, if the above method was optimal, the posterior Metropolis chain would satisfy

$$t_{\mathrm{mix}} \approx \exp\left(c\frac{k^2}{\lambda^2}\right),$$

up to logarithmic factors.

This interestingly matches the runtime scale suggested by the low-degree picture [DKW+23]: in the hard recovery regime

$$\sqrt{k} \ll \lambda \ll k,$$

the conjectural optimal runtime is

$$\exp\left(\Theta\left(\frac{k^2}{\lambda^2}\right)\right).$$

16.8 Planted Clique

We now briefly discuss an analogous MCMC picture for planted clique.

Recall that in planted clique we observe

$$G \sim \mathcal{G}_{n,\frac{1}{2}} \cup \mathrm{Clique}(S), \qquad S \sim \mathrm{Unif}\left(\binom{V(G)}{k}\right).$$

Exact recovery is impossible if $k < 2\log_2 n$ and easy when $k \gg \sqrt{n}$.

If one works directly with the posterior over k-subsets, then

$$\pi(C) \propto \mathbf{1}\{C \text{ is a } k\text{-clique in } G\}.$$

When $\log n \ll k$, the planted clique is typically the unique k-clique, so this posterior is essentially degenerate. This makes the fixed-size posterior essentially degenerate, and therefore not a useful landscape for studying local Markov-chain dynamics..

In [Jer92], Jerrum proposed instead to run MCMC over cliques of all sizes, with Gibbs distribution

$$\pi_\beta(C) \propto e^{\beta|C|}, \qquad C \text{ a clique in } G.$$

This distribution favors large cliques but still allows the chain to explore cliques of different sizes. Jerrum proved that when $k \ll \sqrt{n}$, this chain has a bottleneck for suitable β, providing early evidence for the planted clique conjecture.

More recent work [CMZ25] proves a stronger statement for this type of Metropolis dynamics: if

$$k = n^\alpha \qquad \text{with} \qquad 0 < \alpha < 1,$$

then for every value of the inverse temperature β there exists a bottleneck. Hence the corresponding Metropolis chain mixes slowly across the entire range of polynomial planted clique sizes.

This statement should be interpreted carefully. When $k \gg \sqrt{n}$, planted clique is algorithmically easy by spectral or degree-based methods. The result does not say that planted clique is computationally hard in that regime. Rather, it says that this particular local Metropolis chain over cliques is not an optimal algorithm, even in a regime where other polynomial-time algorithms succeed. In fact, in [GJX25] proved that modified local dynamics succeed in polynomial-time for $\alpha > 1/2$.

16.8.1 The overlap profile

Let S be the planted clique. For a clique C in G, define its overlap with S by $|C \cap S|$. Let

$$\Gamma(\ell) := \max_{C:\, |C\cap S|=\ell} |C|.$$

Thus $\Gamma(\ell)$ is the size of the largest clique with exactly ℓ planted vertices.

At zero overlap, the graph behaves like an Erdős-Rényi random graph, so

$$\Gamma(0) \approx 2\log_2 n.$$

For small overlap $\ell = \gamma \log_2 n$, it turns out that

$$\Gamma(\ell) \approx \left(1 + \sqrt{(1-\gamma)^2 + 2\alpha\gamma}\right) \log_2 n.$$

In particular, if $\alpha < 1$, then at a suitable intermediate overlap there is a strict drop below $\Gamma(0)$ (see Figure 16.3):

$$\Gamma(\ell) - \Gamma(0) \le -c \log n.$$

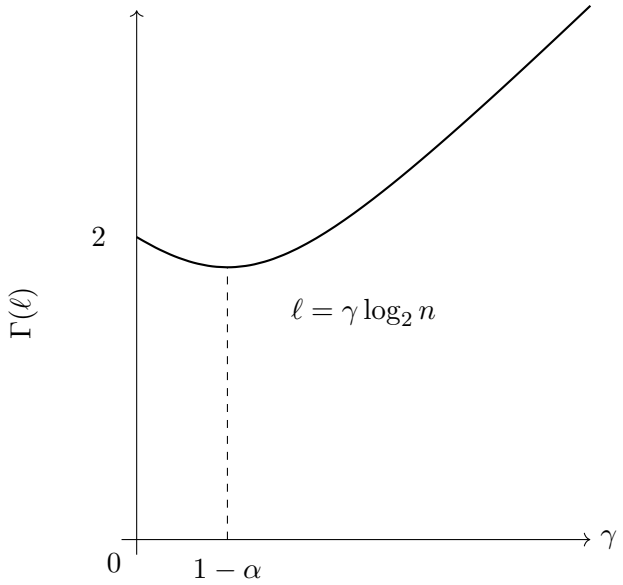

Figure 16.3: Overlap Gap Property for Planted Clique

For large β, this creates the same kind of bottleneck as in sparse PCA: to move from low-overlap random cliques to the planted clique, the chain must pass through overlap levels where the best available clique size is smaller by order $\log n$. This is the reason for the Metropolis failure for large $\beta > 0$. [CMZ25] proved it for all β.

16.8.2 A first-moment derivation of $\Gamma(\gamma \log_2 n)$

Let

$$\ell = \gamma \log_2 n, \qquad t = \xi \log_2 n.$$

We estimate the expected number of cliques of size t with overlap ℓ with the planted clique. There are

$$\binom{k}{\ell}\binom{n-k}{t-\ell}$$

ways to choose the vertices. The edges inside the ℓ planted vertices are automatically present, but all other edges must appear randomly. Therefore the probability that the chosen set is a clique is

$$2^{\binom{\ell}{2}-\binom{t}{2}}.$$

Hence

$$\mathbb{P}(\Gamma(\ell) \geq t) \leq \binom{k}{\ell}\binom{n-k}{t-\ell} 2^{\binom{\ell}{2}-\binom{t}{2}}.$$

Using $k = n^\alpha$, $\ell = \gamma \log_2 n$, and $t = \xi \log_2 n$, the exponent at scale $(\log_2 n)^2$ is

$$\alpha\gamma + \xi - \gamma + \frac{\gamma^2 - \xi^2}{2}.$$

Thus the first moment tends to zero when

$$\alpha\gamma + \xi - \gamma + \frac{\gamma^2 - \xi^2}{2} < 0.$$

Solving for ξ gives

$$\xi > 1 + \sqrt{(1-\gamma)^2 + 2\alpha\gamma}.$$

This yields the upper bound

$$\Gamma(\gamma \log_2 n) \lesssim \left(1 + \sqrt{(1-\gamma)^2 + 2\alpha\gamma}\right) \log_2 n.$$

16.9 Conclusion

This chapter studied MCMC lower bounds through the geometry of the posterior landscape. The key idea is that slow mixing follows from bottlenecks: sets of non-negligible stationary mass whose boundary has very small stationary mass. For local Metropolis chains, such bottlenecks often arise because moving from an uninformative *low-overlap* region of the state space to the *high-overlap* region requires crossing a low-probability barrier. An interesting and wide open direction is to prove tight uppper bounds for MCMC dynamics. A later work in fact revealed an intriguing *geometric* rule between the location of the exact low-temperature MCMC threshold (where some "canonical" low-temperature MCMC method actually succeeds in polynomial-time), the low-degree polynomial threshold, and the information-theoretic threshold [CSZ25].

References

[AS92] N. Alon and J. H. Spencer. *The Probabilistic Method.* 1st. John Wiley & Sons, 1992. ISBN: 978-0-471-53588-1.

[AGJ20] G. B. Arous, R. Gheissari, and A. Jagannath. "Algorithmic thresholds for tensor PCA". In: *The Annals of Probability* 48.4 (2020), pp. 2052–2087.

[AWZ23] G. B. Arous, A. S. Wein, and I. Zadik. "Free Energy Wells and Overlap Gap Property in Sparse PCA." In: *Communications on Pure & Applied Mathematics* 76.10 (2023).

[BBP05] J. Baik, G. Ben Arous, and S. Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *The Annals of Probability* 33.5 (2005), pp. 1643–1697.

[BEH+22] A. S. Bandeira, A. El Alaoui, S. Hopkins, T. Schramm, A. S. Wein, and I. Zadik. "The Franz-Parisi criterion and computational trade-offs in high dimensional statistics". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33831–33844.

[BDM+16] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová. "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems.* NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 424–432. ISBN: 9781510838819.

[BN11] F. Benaych-Georges and R. R. Nadakuditi. "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". In: *Advances in Mathematics* 227.1 (2011), pp. 494–521. ISSN: 0001-8708. DOI: https://doi.org/10.1016/j.aim.2011.02.007.

[BR13] Q. Berthet and P. Rigollet. "Optimal detection of sparse principal components in high dimension". In: *The Annals of Statistics* 41.4 (2013), pp. 1780–1815. DOI: 10.1214/13-AOS1127.

[BB19] M. Brennan and G. Bresler. "Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness". In: *Conference on Learning Theory.* PMLR. 2019, pp. 469–470.

[BBH18] M. Brennan, G. Bresler, and W. Huleihel. "Reducibility and Computational Lower Bounds for Problems with Planted Sparse Structure". In: *Proceedings of the 31st Conference On Learning Theory.* Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 48–166.

[BBH19] M. Brennan, G. Bresler, and W. Huleihel. "Universality of Computational Lower Bounds for Submatrix Detection". In: *Proceedings of Machine Learning Research* 99 (2019). Publisher Copyright: © 2019 M. Brennan, G. Bresler & W. Huleihel.; 32nd Conference on Learning Theory, COLT 2019 ; Conference date: 25-06-2019 Through 28-06-2019, pp. 417–468. ISSN: 2640-3498.

[BBH+21] M. S. Brennan, G. Bresler, S. Hopkins, J. Li, and T. Schramm. "Statistical Query Algorithms and Low Degree Tests Are Almost Equivalent". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 774–774.

[BRS+21] J. Bruna, O. Regev, M. J. Song, and Y. Tang. "Continuous lwe". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 694–707.

[BHJ+25] R.-D. Buhai, J.-T. Hsieh, A. Jain, and P. K. Kothari. "The quasi-polynomial low-degree conjecture is false". In: *arXiv preprint arXiv:2505.17360* (2025).

[CGG+25] A. Carpentier, S. M. Giancola, C. Giraud, and N. Verzelen. "Low-degree lower bounds via almost orthonormal bases". In: *arXiv preprint arXiv:2509.09353* (2025).

[CMZ25] Z. Chen, E. Mossel, and I. Zadik. "Almost-Linear Planted Cliques Elude the Metropolis Process". In: *Random Structures & Algorithms* 66.2 (2025), e21274. DOI: https://doi.org/10.1002/rsa.21274.

[CSZ25] Z. Chen, C. Sheehan, and I. Zadik. *On the Low-Temperature MCMC threshold: the cases of sparse tensor PCA, sparse regression, and a geometric rule*. 2025. arXiv: 2408.00746 [math.ST].

[DK22] I. Diakonikolas and D. Kane. "Non-gaussian component analysis via lattice basis reduction". In: *Conference on Learning Theory*. PMLR. 2022, pp. 4535–4547.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 73–84.

[DKW+23] Y. Ding, D. Kunisky, A. S. Wein, and A. S. Bandeira. "Subexponential-Time Algorithms for Sparse PCA". In: *Found. Comput. Math.* 24.3 (Jan. 2023), pp. 865–914. ISSN: 1615-3375. DOI: 10.1007/s10208-023-09603-0.

[DH21] R. Dudeja and D. Hsu. "Statistical query lower bounds for tensor PCA". In: *J. Mach. Learn. Res.* 22.1 (Jan. 2021). ISSN: 1532-4435.

[FGR+17] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. "Statistical Algorithms and a Lower Bound for Detecting Planted Cliques". In: *Journal of the ACM* 64 (Apr. 2017), pp. 1–37. DOI: 10.1145/3046674.

[FGV17] V. Feldman, C. Guzmán, and S. Vempala. "Statistical query algorithms for mean vector estimation and stochastic convex optimization". In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '17. Barcelona, Spain: Society for Industrial and Applied Mathematics, 2017, pp. 1265–1277.

[FVR+22] O. Y. Feng, R. Venkataramanan, C. Rush, and R. J. Samworth. "A unifying tutorial on approximate message passing". In: *Foundations and Trends in Machine Learning* 15.4 (2022), pp. 335–536.

[FP07] D. Féral and S. Péché. "The Largest Eigenvalue of Rank One Deformation of Large Wigner Matrices". In: *Communications in Mathematical Physics* 272.1 (2007), pp. 185–228. ISSN: 1432-0916. DOI: 10.1007/s00220-007-0209-3.

[FS26] D. Fu and Y. Sohn. "Low-degree estimation thresholds in planted hypergraphs and tensor PCA". In: *arXiv preprint arXiv:2605.30113* (2026).

[GZ22] D. Gamarnik and I. Zadik. "Sparse high-dimensional linear regression. Estimating squared error and a phase transition". In: *The Annals of Statistics* 50.2 (2022), pp. 880–903.

[GZ24] D. Gamarnik and I. Zadik. "The landscape of the planted clique problem: Dense subgraphs and the overlap gap property". In: *The Annals of Applied Probability* 34.4 (2024), pp. 3375–3434.

[GLS+15] D. Gavinsky, S. Lovett, M. Saks, and S. Srinivasan. "A tail bound for read-k families of functions". In: *Random Structures & Algorithms* 47.1 (2015), pp. 99–108. doi: https://doi.org/10.1002/rsa.20532.

[GJX25] R. Gheissari, A. Jagannath, and Y. Xu. "Finding planted cliques using gradient descent". In: *SIAM Journal on Mathematics of Data Science* 7.2 (2025), pp. 643–669.

[GSV05] D. Guo, S. Shamai, and S. Verdu. "Mutual information and minimum mean-square error in Gaussian channels". In: *IEEE Trans. Inf. Theor.* 51.4 (Apr. 2005), pp. 1261–1282. issn: 0018-9448. doi: 10.1109/TIT.2005.844072.

[Has70] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109. issn: 00063444, 14643510.

[HS24] S. Hirahara and N. Shimizu. "Planted Clique Conjectures Are Equivalent". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Vancouver, BC, Canada: Association for Computing Machinery, 2024, pp. 358–366. isbn: 9798400703836. doi: 10.1145/3618260.3649751.

[HW21] J. Holmgren and A. S. Wein. "Counterexamples to the Low-Degree Conjecture". In: *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*. Ed. by J. R. Lee. Vol. 185. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 75:1–75:9. isbn: 978-3-95977-177-1. doi: 10.4230/LIPIcs.ITCS.2021.75.

[Hop18] S. Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

[HL19] S. B. Hopkins and J. Li. "How Hard is Robust Mean Estimation?" In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 1649–1682.

[HKK+26] J.-T. Hsieh, D. M. Kane, P. K. Kothari, J. Li, S. Mohanty, and S. Tiegel. "Rigorous implications of the low-degree heuristic". In: *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*. 2026, pp. 1763–1770.

[Jer92] M. Jerrum. "Large Cliques Elude the Metropolis Process". In: *Random Structures & Algorithms* 3.4 (1992), pp. 347–359. doi: https://doi.org/10.1002/rsa.3240030402.

[JV26] H. Jia and A. Vijayaraghavan. "Low-degree method fails to predict robust subspace recovery". In: *arXiv preprint arXiv:2603.02594* (2026).

[Kea98] M. Kearns. "Efficient noise-tolerant learning from statistical queries". In: *J. ACM* 45.6 (Nov. 1998), pp. 983–1006. issn: 0004-5411.

[Kun22] D. Kunisky. "Lecture notes on sum-of-squares optimization". In: *Lecture Notes, Yale University [http://www. kunisky. com/static/teaching/2022spring-sos/sos-notes.pdf]* (2022).

[KWB19] D. Kunisky, A. S. Wein, and A. S. Bandeira. *Notes on Computational Hardness of Hypothesis Testing: Predictions using the Low-Degree Likelihood Ratio*. 2019. arXiv: 1907.11636 [math.ST].

[LL18] M. Lelarge and M. Léo. "Fundamental limits of symmetric low-rank matrix estimation". In: *Probability Theory and Related Fields* 173.3-4 (2018), pp. 859–929. issn: 0178-8051. doi: 10.1007/s00440-018-0845-x.

[LM19] M. Lelarge and L. Miolane. *Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting*. 2019. arXiv: 1907.03792 [cs.LG].

[LPW06] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

[LZZ25] S. Li, I. Zadik, and M. Zampetakis. "On the Hardness of Learning One Hidden Layer Neural Networks". In: *Algorithmic Learning Theory*. PMLR. 2025, pp. 700–701.

[Li26] Z. Li. "Algorithmic contiguity from low-degree heuristic II: Predicting detection-recovery gaps". In: *arXiv preprint arXiv:2604.17410* (2026).

[MW15] Z. Ma and Y. Wu. "Computational barriers in minimax submatrix detection". In: *The Annals of Statistics* (2015), pp. 1089–1116.

[MRR+53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. *Equation of state calculations by fast computing machines*. Tech. rep. Los Alamos Scientific Lab., Los Alamos, NM (United States); Univ. of Chicago, IL (United States), Mar. 1953. doi: 10.2172/4390578.

[NZ20a] J. Niles-Weed and I. Zadik. *The All-or-Nothing Phenomenon in Sparse Tensor PCA*. 2020. arXiv: 2007.11138 [math.ST].

[NZ20b] J. Niles-Weed and I. Zadik. "The all-or-nothing phenomenon in sparse tensor PCA". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. isbn: 9781713829546.

[SW22] T. Schramm and A. S. Wein. "Computational barriers to estimation from low-degree polynomials". In: *The Annals of Statistics* 50.3 (2022), pp. 1833–1858. doi: 10.1214/22-AOS2179.

[SW25] Y. Sohn and A. S. Wein. "Sharp phase transitions in estimation with low-degree polynomials". In: *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*. 2025, pp. 891–902.

[TWZ26] K. Tsirkas, L. Wang, and I. Zadik. "The monotonicity of the Franz-Parisi potential is equivalent with Low-degree MMSE lower bounds". In: *arXiv preprint arXiv:2603.20070* (2026).

[ZSW+22] I. Zadik, M. J. Song, A. S. Wein, and J. Bruna. "Lattice-based methods surpass sum-of-squares in clustering". In: *Conference on Learning Theory*. PMLR. 2022, pp. 1247–1248.

[ZK16] L. Zdeborová and F. Krzakala. "Statistical physics of inference: Thresholds and algorithms". In: *Advances in Physics* 65.5 (2016), pp. 453–552.

Appendix A

Deferred proofs

A.1 Proof of Theorem 3.2

Proof of Theorem 3.2. For $0 \le i \le k-1$, let

$$X_i := \#\left\{A \in \binom{V(G)}{k} : \binom{A}{2} \subseteq E(G) \text{ and } |A \cap P| = i\right\}.$$

Thus, X_i counts the k-cliques other than P that intersect P in exactly i vertices.

For a fixed set A with $|A \cap P| = i$, the $\binom{i}{2}$ edges contained in $A \cap P$ are present deterministically. All remaining $\binom{k}{2} - \binom{i}{2}$ required edges are independent Bernoulli random variables with parameter $1/2$. Hence

$$\mathbb{E}[X_i] = \binom{k}{i}\binom{n-k}{k-i}2^{-\binom{k}{2}+\binom{i}{2}}.$$

Let

$$X := \sum_{i=0}^{k-1} X_i.$$

Then X is the number of k-cliques in G other than P. Substituting $j = k - i$, we obtain

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{k-1}\binom{k}{i}\binom{n-k}{k-i}2^{-\binom{k}{2}+\binom{i}{2}} \\
&= \sum_{j=1}^{k}\binom{k}{j}\binom{n-k}{j}2^{-\binom{k}{2}+\binom{k-j}{2}} \\
&= \sum_{j=1}^{k}\binom{k}{j}\binom{n-k}{j}2^{\frac{j(j-2k+1)}{2}}.
\end{aligned}$$

We first prove the upper-threshold assertion. Suppose that

$$k \ge (2+\varepsilon)\log_2 n.$$

Using

$$\binom{a}{b} \le \left(\frac{ae}{b}\right)^b$$

and $n - k \le n$, we get

$$\mathbb{E}[X] \le \sum_{j=1}^{k} \left(\frac{e^2 kn}{j^2} 2^{\frac{j-2k+1}{2}} \right)^j.$$

Set

$$B_j := \frac{e^2 kn}{j^2} 2^{\frac{j-2k+1}{2}}, \qquad 1 \le j \le k.$$

Regarding j as a real variable,

$$\log B_j = \log(e^2 kn) - 2\log j + \frac{j-2k+1}{2}\log 2,$$

and therefore

$$\frac{d^2}{dj^2} \log B_j = \frac{2}{j^2} > 0.$$

Thus, $\log B_j$ is convex on $[1, k]$, so its maximum is attained at one of the endpoints. Consequently,

$$\max_{1 \le j \le k} B_j = \max\{B_1, B_k\}.$$

At the endpoints,

$$B_1 = e^2 kn\, 2^{1-k}.$$

Since $k \le n$ and

$$2^{-k} \le n^{-(2+\varepsilon)},$$

we have

$$B_1 \le 2e^2 n^2 n^{-(2+\varepsilon)} = 2e^2 n^{-\varepsilon} = o(1).$$

Similarly,

$$B_k = \frac{e^2 n}{k} 2^{\frac{1-k}{2}} \le \sqrt{2} e^2 n\, 2^{-k/2} \le \sqrt{2} e^2 n^{-\varepsilon/2} = o(1).$$

Let

$$M := \max\{B_1, B_k\}.$$

Then $M = o(1)$, and hence

$$\mathbb{E}[X] \le \sum_{j=1}^{k} M^j \le \frac{M}{1-M} = o(1).$$

By Markov's inequality,

$$\mathbb{P}(X > 0) \le \mathbb{E}[X] = o(1).$$

Therefore, P is the unique k-clique with high probability.

We now prove the lower-threshold assertion. Suppose that

$$k \le (2 - \varepsilon)\log_2 n.$$

It is enough to show that, with high probability, there is a k-clique disjoint from P. Let

$$m := n - k.$$

Then X_0 counts the k-cliques contained in $V(G) \setminus P$, and

$$\mathbb{E}[X_0] = \binom{m}{k} 2^{-\binom{k}{2}}.$$

Since

$$\binom{m}{k} \geq \left(\frac{m}{k}\right)^k,$$

we obtain

$$\log_2 \mathbb{E}[X_0] \geq k\left(\log_2 m - \log_2 k - \frac{k-1}{2}\right).$$

Here $k = O(\log n)$, so $m \sim n$ and $\log_2 k = O(\log\log n)$. Therefore,

$$\begin{aligned}\log_2 \mathbb{E}[X_0] &\geq k\left(\log_2 n - \log_2 k - \frac{k-1}{2} + o(1)\right)\\ &\geq k\left(\frac{\varepsilon}{2}\log_2 n - O(\log\log n)\right) \longrightarrow \infty.\end{aligned}$$

In particular,

$$\mathbb{E}[X_0] \longrightarrow \infty.$$

We apply the second moment method. For two k-subsets $C_1, C_2 \subseteq V(G) \setminus P$ with $|C_1 \cap C_2| = i$, the event that both sets induce cliques requires

$$2\binom{k}{2} - \binom{i}{2}$$

distinct random edges. Thus,

$$\mathbb{E}\left[X_0^2\right] = \binom{m}{k}\sum_{i=0}^{k}\binom{k}{i}\binom{m-k}{k-i}2^{-2\binom{k}{2}+\binom{i}{2}}.$$

Dividing by

$$\mathbb{E}[X_0]^2 = \binom{m}{k}^2 2^{-2\binom{k}{2}},$$

we get

$$\frac{\mathbb{E}\left[X_0^2\right]}{\mathbb{E}[X_0]^2} = \sum_{i=0}^{k}\frac{\binom{k}{i}\binom{m-k}{k-i}}{\binom{m}{k}}2^{\binom{i}{2}}.$$

The term corresponding to $i = 0$ is at most 1:

$$\frac{\binom{m-k}{k}}{\binom{m}{k}} \leq 1.$$

The term corresponding to $i = k$ is

$$\frac{2^{\binom{k}{2}}}{\binom{m}{k}} = \frac{1}{\mathbb{E}[X_0]} = o(1).$$

For $1 \leq i \leq k-1$, we have

$$\begin{aligned}\frac{\binom{m-k}{k-i}}{\binom{m}{k}} &\leq \frac{\binom{m-i}{k-i}}{\binom{m}{k}}\\ &= \frac{k(k-1)\cdots(k-i+1)}{m(m-1)\cdots(m-i+1)}\\ &\leq \left(\frac{k}{m}\right)^i.\end{aligned}$$

It follows that

$$\begin{aligned}\frac{\binom{k}{i}\binom{m-k}{k-i}}{\binom{m}{k}}2^{\binom{i}{2}} &\le \binom{k}{i}\left(\frac{k}{m}\right)^i 2^{i(i-1)/2} \\ &\le \left(\frac{k^2}{m}\right)^i 2^{i(i-1)/2} \\ &\le \left(\frac{k^2}{m}2^{(k-1)/2}\right)^i.\end{aligned}$$

Set

$$Q := \frac{k^2}{m}2^{(k-1)/2}.$$

Because $k \le (2-\varepsilon)\log_2 n$, we have

$$2^{k/2} \le n^{1-\varepsilon/2}.$$

Since $m \sim n$ and $k = O(\log n)$,

$$Q \le \frac{k^2}{m}n^{1-\varepsilon/2} = O\left(\frac{(\log n)^2}{n^{\varepsilon/2}}\right) = o(1).$$

Consequently,

$$\sum_{i=1}^{k-1}\frac{\binom{k}{i}\binom{m-k}{k-i}}{\binom{m}{k}}2^{\binom{i}{2}} \le \sum_{i=1}^{k-1}Q^i \le \frac{Q}{1-Q} = o(1).$$

We conclude that

$$\frac{\mathbb{E}\left[X_0^2\right]}{\mathbb{E}[X_0]^2} \le 1 + \frac{1}{\mathbb{E}[X_0]} + \frac{Q}{1-Q} = 1 + o(1).$$

On the other hand, by nonnegativity of the variance,

$$\mathbb{E}\left[X_0^2\right] \ge \mathbb{E}[X_0]^2.$$

Therefore,

$$\frac{\mathbb{E}\left[X_0^2\right]}{\mathbb{E}[X_0]^2} \longrightarrow 1.$$

By the Paley-Zygmund inequality,

$$\mathbb{P}(X_0 > 0) \ge \frac{\mathbb{E}[X_0]^2}{\mathbb{E}\left[X_0^2\right]} \longrightarrow 1.$$

Thus, with high probability there is a k-clique disjoint fromc P, and hence P is not the unique k-clique. □

A.2 Non-optimality of PCA for the Rademacher prior

Assume that

$$P_0 = \mathrm{Unif}\left(\{-1, 1\}\right).$$

Then

$$\begin{aligned}i(c) &= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log \mathop{\mathbb{E}}_{x\sim P_0}\left[\exp\left(\sqrt{c}\,xy - \frac{cx^2}{2}\right)\right]\right] \\ &= \frac{c}{2} - \mathop{\mathbb{E}}_{y}\left[\log\left(e^{-\frac{c}{2}}\cosh\left(\sqrt{c}\,y\right)\right)\right] \\ &= c - \mathop{\mathbb{E}}_{y}\left[\log\cosh\left(\sqrt{c}\,y\right)\right] \\ &= c - \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\log\cosh\left(c + \sqrt{c}\,Z\right)\right],\end{aligned}$$

where the last equality follows from the symmetries of $\cosh$ and P_0.

Then the replica-symmetric formula yields not a closed expression, but a fixed-point relation that allows to end up proving the non-optimality of PCA. Particularly, it can be proven that

will later allow us to prove the non-optimality of PCA: let

$$\phi_t(q) := \frac{tq}{2}\left(1-\frac{q}{2}\right) - tq + \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\log\cosh\left(tq+\sqrt{tq}\,Z\right)\right]$$

$$\begin{aligned}
\phi_t'(q) &= -\frac{t}{2} - \frac{tq}{2} + \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh\left(tq+\sqrt{tq}\,Z\right)\left(t+\frac{t}{2\sqrt{tq}}Z\right)\right] \\
&= -\frac{t}{2} - \frac{tq}{2} + t\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh\left(tq+\sqrt{tq}\,Z\right)\right] + \frac{t}{2\sqrt{tq}}\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[Z\tanh\left(tq+\sqrt{tq}\,Z\right)\right] \\
&= -\frac{t}{2} - \frac{tq}{2} + t\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh\left(tq+\sqrt{tq}\,Z\right)\right] + \frac{t}{2}\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[1-\tanh^2\left(tq+\sqrt{tq}\,Z\right)\right] \\
&= -\frac{tq}{2} + t\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh\left(tq+\sqrt{tq}\,Z\right)\right] - \frac{t}{2}\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh^2\left(tq+\sqrt{tq}\,Z\right)\right] \\
&= -\frac{tq}{2} + t\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh^2\left(tq+\sqrt{tq}\,Z\right)\right] - \frac{t}{2}\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh^2\left(tq+\sqrt{tq}\,Z\right)\right] \\
&= \frac{t}{2}\left(\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh^2\left(tq+\sqrt{tq}\,Z\right)\right] - q\right),
\end{aligned}$$

where Gaussian integration by parts has been used in the third equality and the fifth inequality has used the following result.

Fact A.1. *Let $a>0$ and let*

$$u = a + \sqrt{a}\,Z, \qquad Z\sim\mathcal{N}(0,1).$$

Then

$$\mathop{\mathbb{E}}_{Z}[\tanh(u)] = \mathop{\mathbb{E}}_{Z}\left[\tanh^2(u)\right].$$

Proof. Let f be the density of u. Thus

$$f(x) = \frac{1}{\sqrt{2\pi a}}\exp\left(-\frac{(x-a)^2}{2a}\right) \quad\Longrightarrow\quad \frac{f(x)}{f(-x)} = \exp\left(\frac{(-x-a)^2-(x-a)^2}{2a}\right) = e^{2x}$$

Hence, $f(x) = e^{2x}f(-x)$. Using that $\tanh$ is odd and $\tanh^2$ is even, we can write

$$\mathop{\mathbb{E}}_{Z}[\tanh(u)] = \int_0^\infty \tanh(x)\big(f(x)-f(-x)\big)\mathrm{d}x, \quad \mathop{\mathbb{E}}_{Z}\left[\tanh^2(u)\right] = \int_0^\infty \tanh^2(x)\big(f(x)+f(-x)\big)\mathrm{d}x.$$

Using $f(x) = e^{2x}f(-x)$, it is enough to check that

$$\tanh(x)(e^{2x}-1) = \tanh^2(x)(e^{2x}+1).$$

But this follows immediately from

$$\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}.$$

Therefore the two integrals are equal. □

So, the stationary points of $\phi_t(q)$ (and thus $q^*(t)$, unless $q^*(t)=0$) satisfy the fixed-point equation

$$q(t) = \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh^2\left(tq(t)+\sqrt{tq(t)}\,Z\right)\right] \overset{\text{Fact A.1}}{=} \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh\left(tq(t)+\sqrt{tq(t)}\,Z\right)\right].$$

Let's now see PCA is not optimal given $t > 1$, i.e.

$$1 - q^*(t)^2 < 1 - \left(1 - \frac{1}{t}\right)^2 \quad \equiv \quad q^*(t) > 1 - \frac{1}{t}, \quad t > 1.$$

Set

$$Z \sim \mathcal{N}(0,1), \quad U := t - 1 + \sqrt{t-1}Z, \quad F_t(q) := \mathop{\mathbb{E}}_Z\left[\tanh\left(tq + \sqrt{tq}\, Z\right)\right]$$

$$\implies \quad F_t\left(1 - \frac{1}{t}\right) := \mathop{\mathbb{E}}_Z\left[\tanh(U)\right], \qquad \phi_t'(q) = \frac{t}{2}(F_t(q) - q).$$

By Gaussian integration by parts and Fact A.1,

$$\begin{aligned}
\mathop{\mathbb{E}}_Z\left[\left(tq + \sqrt{tq}\, Z\right)\tanh\left(tq + \sqrt{tq}\, Z\right)\right] &= tq\mathop{\mathbb{E}}_Z\left[\tanh\left(tq + \sqrt{tq}\, Z\right)\right] + \sqrt{tq}\mathop{\mathbb{E}}_Z\left[Z\tanh\left(tq + \sqrt{tq}\, Z\right)\right] \\
&= tq\mathop{\mathbb{E}}_Z\left[\tanh\left(tq + \sqrt{tq}\, Z\right)\right] + tq\mathop{\mathbb{E}}_Z\left[\operatorname{sech}^2\left(tq + \sqrt{tq}\, Z\right)\right] \\
&= tq\mathop{\mathbb{E}}_Z\left[\tanh^2\left(tq + \sqrt{tq}\, Z\right)\right] + tq\mathop{\mathbb{E}}_Z\left[\operatorname{sech}^2\left(tq + \sqrt{tq}\, Z\right)\right] \\
&= tq. \qquad \left(\tanh^2(x) + \operatorname{sech}^2(x) = 1\right)
\end{aligned}$$

Using $q = 1 - \frac{1}{t}$, Cauchy-Schwarz gives

$$(t-1)^2 = \mathop{\mathbb{E}}_Z\left[U\tanh(U)\right]^2 < \mathop{\mathbb{E}}_Z\left[U^2\right]\mathop{\mathbb{E}}_Z\left[\tanh^2(U)\right].$$

The inequality is strict because equality in Cauchy-Schwarz would force $\tanh(U)$ to be a linear function of U almost surely, which is impossible since U is a non-degenerate Gaussian and $\tanh$ is not linear.

Since

$$\mathop{\mathbb{E}}_Z\left[U^2\right] = (t-1)^2 + t - 1 = t(t-1),$$

we obtain

$$\mathop{\mathbb{E}}_Z\left[\tanh^2(U)\right] > 1 - \frac{1}{t} \quad \overset{\text{Fact A.1}}{\implies} \quad F_t\left(1 - \frac{1}{t}\right) = \mathop{\mathbb{E}}_Z[\tanh(U)] = \mathop{\mathbb{E}}_Z\left[\tanh^2(U)\right] > 1 - \frac{1}{t}.$$

On the other hand,

$$F_t(1) < 1,$$

because $\tanh\left(t + \sqrt{t}Z\right) < 1$ almost surely. Since F_t is continuous, the function

$$F_t(q) - q = \frac{2}{t}\phi_t'(q)$$

is positive at $1 - \frac{1}{t}$ and negative at 1. Hence there exists a $F_t(q)$-fixed point in $\left(1 - \frac{1}{t}, 1\right)$ that corresponds to

$$\arg\max_{q \in \left[1-\frac{1}{t}, 1\right]} \phi_t(q) \in \left(1 - \frac{1}{t}, 1\right) \quad \implies \quad q^*(t) > 1 - \frac{1}{t}.$$

Finally, let's see PCA is optimal for $0 \le t \le 1$, i.e.

$$q^*(t) = 0, \quad 0 \le t \le 1.$$

As $\phi_t'(q) = \frac{t}{2}(F_t(q) - q)$, it suffices to prove $F_t(q) \le qt \le q$.

For every real x, we have

$$\tanh^2(x) \le x\tanh(x),$$

because $\tanh(x)$ has the same sign as x and $|\tanh(x)| \le |x|$. The inequality is strict for $x \neq 0$. Since U is a non-degenerate Gaussian, it is nonzero with probability one. Hence, reusing a previous equality,

$$F_t(q) \overset{\text{Fact A.1}}{=} \mathop{\mathbb{E}}_{Z}\left[\tanh^2\left(tq + \sqrt{tq}\,Z\right)\right] < \mathop{\mathbb{E}}_{Z}\left[\left(tq + \sqrt{tq}\,Z\right)\tanh\left(tq + \sqrt{tq}\,Z\right)\right] = qt.$$

Thus, PCA is only optimal for $0 \le t < 1$; although it succeeds above $t = 1$, it does not attain the information-theoretic minimum mean-squared error.

A.3 Proof of Lemma 6.1

Proof of Lemma 6.1. Note:

$$\mathbb{P}\left(\max_i Z_i \ge 10\sqrt{\log M}\right) = \mathbb{P}\left(\bigcup_i \left\{Z_i \ge 10\sqrt{\log M}\right\}\right)$$

Then, by the union bound:

$$\le M\mathbb{P}\left(Z_i \ge 10\sqrt{\log M}\right)$$

Note for any $t > 0$:

$$\mathbb{P}(Z_i \ge t) = \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\mathrm{d}x \le \frac{1}{t}\sqrt{\frac{2}{\pi}}e^{-t^2/2}$$

hence for the original quantity, we have:

$$\le M\frac{1}{10\sqrt{\log M}}\sqrt{\frac{2}{\pi}}e^{-\left(10\sqrt{\log M}\right)^2/2}$$

$$= \frac{\sqrt{\frac{2}{\pi}}}{10\sqrt{\log M}} M^{1-50} \le 0.001, \text{ if M is "large enough."}$$

So $\max_i Z_i \le 10\sqrt{\log M}$, with probabili ty 0.999, if M is "large enough," as desired. □

A.4 Proof of Proposition 6.4

Proof of Proposition 6.4. Let

$$\mathcal{E} := \left\{\frac{k}{2} \le \|\theta\|_0 \le 2k\right\}.$$

Since $\|\theta\|_0 \sim \mathrm{Bin}(n, \rho)$ and $k = \rho n$, Chernoff's bound gives

$$\mathbb{P}_\theta(\mathcal{E}^c) \le e^{-c_\rho n}$$

for some constant $c_\rho > 0$. Let $\widetilde{\mu}$ denote the law of θ conditioned on $\mathcal{E}$, and let $\widetilde{\mathbb{P}}_1$ be the corresponding planted distribution. Then

$$\mathrm{TV}\left(\mathbb{P}_1, \widetilde{\mathbb{P}}_1\right) \le \mathbb{P}_\theta(\mathcal{E}^c) = o(1).$$

Thus it is enough to show that $\widetilde{\mathbb{P}}_1$ and $\mathbb{P}_2$ are bounded away from perfect distinguishability.

We prove the second-moment bound

$$\left\|\frac{\widetilde{\mathbb{P}}_1}{\mathbb{P}_2}\right\|_2^2 = \mathcal{O}(1).$$

By the Gaussian additive model identity,

$$\left\| \frac{\widetilde{\mathbb{P}}_1}{\mathbb{P}_2} \right\|_2^2 = \mathop{\mathbb{E}}_{\theta,\theta' \sim \widetilde{\mu}} \left[\exp\left(\frac{\lambda^2}{2} \left\langle \theta, \theta' \right\rangle^2 \right) \right].$$

Let

$$S = \mathrm{supp}(\theta), \qquad S' = \mathrm{supp}(\theta'), \qquad R = |S \cap S'|.$$

Condition on

$$|S| = \ell, \qquad |S'| = \ell',$$

where, by construction of $\widetilde{\mu}$,

$$\frac{k}{2} \le \ell, \ell' \le 2k.$$

Given $R = r$, the signs on the intersection are independent Rademacher variables, so

$$\left\langle \theta, \theta' \right\rangle = \frac{1}{k} \sum_{i=1}^{r} \varepsilon_i,$$

where $\varepsilon_1, \dots, \varepsilon_r$ are i.i.d. Rademacher random variables. Write

$$T_r := \sum_{i=1}^{r} \varepsilon_i$$

and

$$a := \frac{\lambda^2}{2k^2}.$$

Since

$$\lambda \le 0.01 \sqrt{n\rho \log \frac{1}{\rho}} = 0.01 \sqrt{k \log \frac{1}{\rho}},$$

we have, writing

$$L := \log \frac{1}{\rho},$$

that

$$a \le 5 \cdot 10^{-5} \frac{L}{k}.$$

Therefore it suffices to prove, uniformly over $\ell, \ell' \in [k/2, 2k]$, that

$$\mathop{\mathbb{E}}_{R} \left[\mathop{\mathbb{E}}_{\varepsilon} \left[e^{a T_R^2} \right] \right] = \mathcal{O}(1).$$

Set

$$r_* := \left\lfloor \frac{1}{4a} \right\rfloor.$$

We split the expectation according to whether $R \le r_*$ or $R > r_*$.

First assume $r \le r_*$. Then $2ar \le 1/2$. Let $G \sim N(0,1)$ be independent of the signs. Using

$$e^{a T_r^2} = \mathop{\mathbb{E}}_{G} \left[e^{\sqrt{2a} G T_r} \;\middle|\; T_r \right],$$

we obtain

$$\begin{aligned}\mathop{\mathbb{E}}_{\varepsilon}\left[e^{aT_r^2}\right] &= \mathop{\mathbb{E}}_{G}\left[\prod_{i=1}^{r}\mathop{\mathbb{E}}_{\varepsilon_i}\left[e^{\sqrt{2a}G\varepsilon_i}\right]\right] \\ &= \mathop{\mathbb{E}}_{G}\left[\cosh\left(\sqrt{2a}G\right)^r\right] \\ &\le \mathop{\mathbb{E}}_{G}\left[e^{arG^2}\right] \\ &= \frac{1}{\sqrt{1-2ar}} \\ &\le e^{2ar}.\end{aligned}$$

Thus

$$\mathbb{E}\left[e^{aT_R^2}\mathbf{1}_{\{R\le r_*\}}\right] \le \mathbb{E}\left[e^{2aR}\right].$$

Conditional on ℓ, ℓ', the random variable R is hypergeometric. Its moment-generating function is bounded by the corresponding binomial moment-generating function, and therefore

$$\mathbb{E}\left[e^{2aR} \mid \ell, \ell'\right] \le \left(1 + \frac{\ell'}{n-\ell}\left(e^{2a}-1\right)\right)^{\ell}.$$

Since $\ell, \ell' \le 2k = 2\rho n$ and $\rho < 1/4$, we have

$$\frac{\ell'}{n-\ell} \le \frac{2\rho}{1-2\rho}.$$

Hence

$$\mathbb{E}\left[e^{2aR} \mid \ell, \ell'\right] \le \exp\left(\frac{4\rho k}{1-2\rho}\left(e^{2a}-1\right)\right).$$

Since

$$a \le 5\cdot 10^{-5}\frac{L}{k},$$

we have $e^{2a} - 1 = \mathcal{O}\left(\frac{L}{k}\right)$, and therefore

$$\frac{4\rho k}{1-2\rho}\left(e^{2a}-1\right) = \mathcal{O}(1).$$

Thus the contribution from $R \le r_*$ is $\mathcal{O}(1)$.

We now treat the contribution from $R > r_*$. If $r > r_*$, then the crude bound $|T_r| \le r$ gives

$$\mathop{\mathbb{E}}_{\varepsilon}\left[e^{aT_r^2}\right] \le e^{ar^2}.$$

Also, for the hypergeometric overlap,

$$\mathbb{P}\left(R = r \mid \ell, \ell'\right) \le \binom{\ell}{r}\left(\frac{\ell'}{n-\ell}\right)^r \le \left(\frac{C_\rho \rho k}{r}\right)^r,$$

where $C_\rho > 0$ depends only on ρ. Therefore

$$\mathbb{P}\left(R = r \mid \ell, \ell'\right)\mathop{\mathbb{E}}_{\varepsilon}\left[e^{aT_r^2}\right] \le \exp\left(r\log\frac{C_\rho\rho k}{r} + ar^2\right).$$

Write $r = qk$. Since $r \le 2k$, we have $q \in (0, 2]$. Moreover,

$$ar^2 \le 5\cdot 10^{-5} Lkq^2.$$

Hence the exponent is at most

$$k\left[q\log\frac{C_\rho\rho}{q}+5\cdot10^{-5}Lq^2\right].$$

If $r_*\geq 2k$, then the event $R>r_*$ is empty. Otherwise, $r_*\leq 2k$, which implies

$$\frac{1}{4a}\leq 3k$$

for all large n, and hence

$$ak\geq\frac{1}{12}.$$

Since $ak\leq 5\cdot10^{-5}L$, this can only happen when

$$L\geq\frac{1}{12\cdot5\cdot10^{-5}}.$$

In this case, for every $r>r_*$ we also have, for all large n,

$$q=\frac{r}{k}\geq\frac{1}{8ak}\geq\frac{1}{8\cdot5\cdot10^{-5}L}.$$

Thus

$$q\geq\frac{2500}{L}.$$

Using $q\leq 2$, we get

$$5\cdot10^{-5}Lq^2\leq10^{-4}Lq.$$

Therefore

$$\begin{aligned}q\log\frac{C_\rho\rho}{q}+5\cdot10^{-5}Lq^2&\leq q\left(\log C_\rho-L-\log q+10^{-4}L\right)\\&=q\left(\log C_\rho-\left(1-10^{-4}\right)L-\log q\right).\end{aligned}$$

Since $q\geq 2500/L$, we have

$$-\log q\leq\log\frac{L}{2500}.$$

For the range of L in which the large-overlap event is nonempty, the quantity

$$\log C_\rho+\log\frac{L}{2500}-\left(1-10^{-4}\right)L$$

is negative. In particular, there exists a constant $c'_\rho>0$ such that

$$q\log\frac{C_\rho\rho}{q}+5\cdot10^{-5}Lq^2\leq-c'_\rho q.$$

Consequently,

$$\mathbb{P}\left(R=r\mid\ell,\ell'\right)\mathop{\mathbb{E}}_{\varepsilon}\left[e^{aT_r^2}\right]\leq e^{-c'_\rho r}.$$

Summing over $r>r_*$ gives

$$\sum_{r>r_*}\mathbb{P}\left(R=r\mid\ell,\ell'\right)\mathop{\mathbb{E}}_{\varepsilon}\left[e^{aT_r^2}\right]\leq\sum_{r>r_*}e^{-c'_\rho r}=o(1),$$

because $r_*=\Theta_\rho(k)$ and $k=\rho n\to\infty$.

Combining the small-overlap and large-overlap estimates, uniformly over all

$$\ell,\ell'\in[k/2,2k],$$

we conclude that

$$\left\| \frac{\widetilde{\mathbb{P}}_1}{\mathbb{P}_2} \right\|_2^2 = \mathcal{O}(1).$$

Let

$$\widetilde{L}(Y) := \frac{\mathrm{d}\widetilde{\mathbb{P}}_1}{\mathrm{d}\mathbb{P}_2}(Y).$$

Then

$$\mathop{\mathbb{E}}_{\mathbb{P}_2}\left[\widetilde{L}\right] = 1, \qquad \mathop{\mathbb{E}}_{\mathbb{P}_2}\left[\widetilde{L}^2\right] \le C_\rho$$

for some finite constant C_ρ. By Paley-Zygmund,

$$\mathbb{P}_2\left(\widetilde{L} \ge \frac{1}{2}\right) \ge \frac{1}{4C_\rho}.$$

Therefore

$$\int \min\left\{\mathrm{d}\widetilde{\mathbb{P}}_1, \mathrm{d}\mathbb{P}_2\right\} = \mathop{\mathbb{E}}_{\mathbb{P}_2}\left[\min\left\{\widetilde{L}, 1\right\}\right] \ge \frac{1}{2}\mathbb{P}_2\left(\widetilde{L} \ge \frac{1}{2}\right) \ge \frac{1}{8C_\rho}.$$

Hence

$$\mathrm{TV}\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) \le 1 - \frac{1}{8C_\rho}.$$

Finally,

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) \le \mathrm{TV}\left(\mathbb{P}_1, \widetilde{\mathbb{P}}_1\right) + \mathrm{TV}\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) \le o(1) + 1 - \frac{1}{8C_\rho}.$$

Thus $\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2)$ is bounded away from 1. Therefore strong detection is impossible. □

A.5 Proof of Lemma 6.6

Proof of Lemma 6.6. Write

$$\langle \theta_1, \theta_2 \rangle = \sum_{i=1}^n X_i, \qquad X_i := (\theta_1)_i (\theta_2)_i.$$

The random variables $X_1, \dots, X_n$ are independent and centered. Moreover,

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

Therefore, for any $s \in \mathbb{R}$,

$$\mathbb{E}\left[e^{sX_i}\right] = 1 - \rho^2 + \rho^2 \cosh\left(\frac{s}{k}\right) = 1 + \rho^2\left(\cosh\left(\frac{s}{k}\right) - 1\right).$$

Let

$$S := \langle \theta_1, \theta_2 \rangle.$$

For $s > 0$, Chernoff's bound and $\log(1+x) \le x$ gives

$$\mathbb{P}(S \ge t) \le \exp(-st)\left(1 + \rho^2\left(\cosh\left(\frac{s}{k}\right) - 1\right)\right)^n \le \exp\left(-st + n\rho^2\left(\cosh\left(\frac{s}{k}\right) - 1\right)\right).$$

We now use the local expansion of $\cosh$. For every $\eta > 0$, there exists $a_\eta > 0$ such that, for all $|a| \le a_\eta$,

$$\cosh(a) - 1 \le \frac{1+\eta/2}{2} a^2.$$

Choose

$$s = nt \quad \Longrightarrow \quad \frac{s}{k} = \frac{nt}{k} = \frac{t}{\rho}.$$

Thus, if $t \le c_\eta \rho$ with $c_\eta \le a_\eta$, then $s/k \le a_\eta$, and so

$$\cosh\left(\frac{s}{k}\right) - 1 \le \frac{1+\eta/2}{2} \frac{t^2}{\rho^2}.$$

Substituting this into the Chernoff bound gives

$$\mathbb{P}(S \ge t) \le \exp\left(-nt^2 + n\rho^2 \frac{1+\eta/2}{2} \frac{t^2}{\rho^2}\right) = \exp\left(-\frac{1-\eta/2}{2} nt^2\right).$$

After decreasing c_η if necessary, we may replace $1-\eta/2$ by $1-\eta$. The symmetry of S around 0 (because each X_i is symmetric) and the union bound prove the desired claim. □

A.6 Proof of Proposition 7.11

Proof of Proposition 7.11. We work under the null model $\mathbb{P}_2 = \mathcal{G}_{n,\frac{1}{2}}$. For each edge e, let

$$\chi_e := \begin{cases} 1 & e \in E(G), \\ -1 & e \notin E(G). \end{cases}$$

Then the functions

$$\chi_F := \prod_{e \in F} \chi_e, \qquad F \subseteq \binom{[n]}{2},$$

form an orthonormal basis for $L^2(\mathbb{P}_2)$.

For a fixed planted clique $S \in \binom{[n]}{v}$, the likelihood ratio is

$$L_S(G) = 2^{\binom{k}{2}} \mathbf{1}\{S \text{ is a clique in } G\}.$$

Since

$$2\mathbf{1}\{e \in E(G)\} = 1 + \chi_e,$$

we may write

$$L_S(G) = \prod_{e \in \binom{S}{2}} (1 + \chi_e) = \sum_{F \subseteq \binom{S}{2}} \chi_F.$$

Averaging over the uniformly random planted clique S, the full likelihood ratio is

$$L(G) := \frac{\mathbb{P}_1(G)}{\mathbb{P}_2(G)} = \sum_F a_F \chi_F, \qquad a_F = \mathbb{P}_S\left(F \subseteq \binom{S}{2}\right).$$

If $v(F)$ denotes the number of vertices incident to at least one edge of F, then

$$a_F = \frac{\binom{n-v(F)}{k-v(F)}}{\binom{n}{k}} = \frac{(k)_{v(F)}}{(n)_{v(F)}} \le \left(\frac{k}{n}\right)^{v(F)}.$$

Therefore,

$$\|L_{\le D}\|_2^2 = \sum_{|F|\le D} a_F^2 = 1 + \sum_{1\le|F|\le D} a_F^2.$$

It remains to show that the second term is $o(1)$.

Group edge sets F according to

$$d := |F|, \qquad v := v(F).$$

The number of choices of F with d edges and v non-isolated vertices is at most

$$\binom{n}{v}\binom{\binom{v}{2}}{d}.$$

Hence

$$\begin{aligned}\sum_{1\le|F|\le D} a_F^2 &\le \sum_{d=1}^{D}\sum_{v=2}^{2d}\binom{n}{v}\binom{\binom{v}{2}}{d}\left(\frac{k}{n}\right)^{2v} \\ &\le \sum_{d=1}^{D}\sum_{v=2}^{2d}\left(\frac{ek^2}{vn}\right)^{v}\binom{\binom{v}{2}}{d}.\end{aligned}$$

Equivalently, summing first over v, it is enough to bound

$$\sum_{v=2}^{2D}\left(\frac{ek^2}{vn}\right)^{v}\sum_{d=1}^{\min\left\{D,\binom{v}{2}\right\}}\binom{\binom{v}{2}}{d}.$$

We split the range of v into two parts.

First suppose

$$v \le 3\sqrt{D}.$$

Then

$$\sum_{d=1}^{\min\left\{D,\binom{v}{2}\right\}}\binom{\binom{v}{2}}{d} \le 2^{\binom{v}{2}} \le \exp\left(Cv^2\right)$$

for a universal constant $C > 0$. Therefore the contribution of such v is at most

$$\exp\left(v\log\frac{ek^2}{vn} + Cv^2\right).$$

Since $\frac{k^2}{n} \le n^{-2\delta}$,

$$v\log\frac{ek^2}{vn} \le -2\delta v\log n + \mathcal{O}(v\log v).$$

Because

$$v \le 3\sqrt{D} = 3(\log n)^{(1+\varepsilon)/2} = o(\log n),$$

we have

$$\mathcal{O}(v\log v) + Cv^2 = o(v\log n).$$

Thus, uniformly in this range,

$$v\log\frac{ek^2}{vn} + Cv^2 \le -\delta v\log n$$

for all large n. Hence the total contribution of the range $v \le 3\sqrt{D}$ is

$$\sum_{v=2}^{\lfloor 3\sqrt{D}\rfloor} e^{-\delta v\log n} = e^{-2\delta\log n}\frac{1-\exp\left(-\left(\left\lfloor 3\sqrt{D}\right\rfloor - 1\right)\delta\log n\right)}{1-e^{-\delta\log n}} = o(1).$$

Now suppose

$$v > 3\sqrt{D}.$$

Then $\binom{v}{2} > D$ for all large n, and we use the standard binomial bound

$$\sum_{d=1}^{D} \binom{\binom{v}{2}}{d} \le (D+1)\left(\frac{ev^2}{D}\right)^D.$$

Since $v \le 2D$, this is at most

$$\exp(CD\log D)$$

for a universal constant $C > 0$. Hence the contribution of a fixed v in this range is at most

$$\exp\left(v\log\frac{ek^2}{vn} + CD\log D\right).$$

Again,

$$v\log\frac{ek^2}{vn} \le -2\delta v\log n + \mathcal{O}(v\log v).$$

Moreover,

$$\mathcal{O}(v\log v) = o(v\log n),$$

because $v \le 2D = 2(\log n)^{1+\varepsilon}$, and

$$CD\log D = o(v\log n)$$

uniformly for $v > 3\sqrt{D}$. Indeed,

$$\frac{D\log D}{v\log n} \le \frac{D\log D}{3\sqrt{D}\log n} = \mathcal{O}\left((\log n)^{(\varepsilon-1)/2}\log\log n\right) = o(1),$$

since $\varepsilon < 1$. Therefore, for all large n,

$$v\log\frac{ek^2}{vn} + CD\log D \le -\delta v\log n.$$

Thus the contribution of the range $v > 3\sqrt{D}$ is at most

$$\begin{aligned}\sum_{v=\lfloor 3\sqrt{D}\rfloor+1}^{2D} e^{-\delta v\log n} &= \exp\left(-\left(\left\lfloor 3\sqrt{D}\right\rfloor+1\right)\delta\log n\right)\frac{1-\exp\left(-\left(2D-\left\lfloor 3\sqrt{D}\right\rfloor\right)\delta\log n\right)}{1-e^{-\delta\log n}}\\ &= o(1).\end{aligned}$$

□

A.7 Proof of Proposition 9.1

Proof of Proposition 9.1. We first prove the χ^2 bound for the conditioned prior. By the Gaussian additive χ^2 formula,

$$\chi^2\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) + 1 = \mathop{\mathbb{E}}_{x,x'\sim\widetilde{\mu}}\left[\exp\left(m\left\langle x, x'\right\rangle^2\right)\right].$$

It is enough to show that the expectation on the right-hand side is $1 + o(1)$. Let

$$\rho := \frac{k}{n}.$$

For two independent unconditioned Bernoulli spikes x, x', define

$$R := \left|\mathrm{supp}(x) \cap \mathrm{supp}\left(x'\right)\right|.$$

Then

$$R \sim \mathrm{Bin}(n, \rho^2),$$

because a coordinate belongs to both supports with probability ρ^2. Conditional on $R = r$,

$$\left\langle x, x' \right\rangle = \frac{1}{k} \sum_{i=1}^{r} \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_r$ are i.i.d. Rademacher random variables. Since $\mathbb{P}(\mathcal{E}) \to 1$, conditioning on $\mathcal{E}$ changes probabilities only by a factor $1 + o(1)$. Moreover, on the event $\mathcal{E}$ for both x and x', we have $R \leq 2k$. Therefore, setting

$$t := \frac{m}{k^2},$$

we have

$$\chi^2\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) + 1 \leq (1 + o(1)) \sum_{r=0}^{2k} \mathbb{P}(R = r) \mathop{\mathbb{E}}_{\varepsilon} \left[\exp\left(t \left(\sum_{i=1}^{r} \varepsilon_i\right)^2\right)\right].$$

Thus it suffices to show that

$$\sum_{r=0}^{2k} \mathbb{P}(R = r) \left(\mathop{\mathbb{E}}_{\varepsilon} \left[\exp\left(t \left(\sum_{i=1}^{r} \varepsilon_i\right)^2\right)\right] - 1\right) = o(1).$$

Let

$$r_0 := \left\lfloor \frac{1}{4t} \right\rfloor = \left\lfloor \frac{k^2}{4m} \right\rfloor.$$

Since $m = o(k \log(n/k))$ and $k = n^{\Omega(1)}$, we have $m = o(k^2)$, and hence $r_0 \to \infty$. We first consider $r \leq r_0$. Let

$$T_r := \sum_{i=1}^{r} \varepsilon_i.$$

Since $tr \leq 1/4$, the standard Gaussian comparison gives

$$\mathop{\mathbb{E}}_{\varepsilon}\left[e^{tT_r^2}\right] \leq \frac{1}{\sqrt{1 - 2tr}} \leq 1 + 4tr.$$

Indeed, if $G \sim N(0,1)$ is independent of the signs, then

$$e^{tT_r^2} = \mathop{\mathbb{E}}_{G}\left[e^{\sqrt{2t} G T_r} \mid T_r\right],$$

and therefore

$$\mathop{\mathbb{E}}_{\varepsilon}\left[e^{tT_r^2}\right] = \mathop{\mathbb{E}}_{G}\left[\cosh\left(\sqrt{2t}G\right)^r\right] \leq \mathop{\mathbb{E}}_{G}\left[e^{trG^2}\right] = \frac{1}{\sqrt{1 - 2tr}}.$$

Thus the small-overlap contribution is bounded by

$$\sum_{r \leq r_0} \mathbb{P}(R = r)\, 4tr \leq 4t\, \mathbb{E}[R].$$

Since $R \sim \mathrm{Bin}(n, \rho^2)$,

$$\mathbb{E}[R] = n\rho^2 = \frac{k^2}{n}.$$

Hence

$$4t\,\mathbb{E}[R] = 4\frac{m}{k^2}\frac{k^2}{n} = \frac{4m}{n} = o(1),$$

because $m = o(k\log(n/k))$ and $k = o(n)$ imply $m = o(n)$. We now consider the large-overlap contribution $r > r_0$. For every $r \geq 1$, the binomial overlap satisfies

$$\mathbb{P}(R = r) = \binom{n}{r}\rho^{2r}(1-\rho^2)^{n-r} \leq \left(\frac{en\rho^2}{r}\right)^r = \left(\frac{ek^2}{nr}\right)^r.$$

Also, since $R \leq 2k$ on the conditioned event, for $r \leq 2k$ we have

$$tT_r^2 \leq tr^2 \leq \frac{2m}{k}r.$$

Therefore, for $r > r_0$,

$$\mathbb{P}(R = r)\,\mathop{\mathbb{E}}_{\varepsilon}\left[e^{tT_r^2}\right] \leq \left(\frac{ek^2}{nr}\right)^r e^{2mr/k}.$$

Since $r > r_0 = k^2/(4m)$, we have

$$\frac{k^2}{r} \leq 4m.$$

Hence

$$\mathbb{P}(R = r)\,\mathop{\mathbb{E}}_{\varepsilon}\left[e^{tT_r^2}\right] \leq \left(\frac{4e\,m}{n}e^{2m/k}\right)^r.$$

Set

$$a_n := \frac{4e\,m}{n}e^{2m/k}.$$

Because

$$m = o\left(k\log\frac{n}{k}\right),$$

we have

$$e^{2m/k} = \exp\left(o\left(\log\frac{n}{k}\right)\right) = \left(\frac{n}{k}\right)^{o(1)}.$$

Moreover,

$$\frac{m}{n} = o\left(\frac{k}{n}\log\frac{n}{k}\right).$$

Thus

$$a_n = o\left(\frac{k}{n}\log\frac{n}{k}\right)\left(\frac{n}{k}\right)^{o(1)} = o(1).$$

For all large n, $a_n \leq 1/2$. Hence

$$\sum_{r>r_0}^{2k}\mathbb{P}(R = r)\mathop{\mathbb{E}}_{\varepsilon}\left[e^{tT_r^2}\right] \leq \sum_{r>r_0}^{2k} a_n^r \leq \frac{a_n^{r_0+1}}{1-a_n} = o(1),$$

because $r_0 \to \infty$. Combining the small-overlap and large-overlap estimates gives

$$\chi^2\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) \to 0.$$

It remains to relate the conditioned and unconditioned Bernoulli priors. By a Chernoff bound,

$$\mathbb{P}(\|x\|_0 > 2k) \leq e^{-ck}.$$

Therefore

$$\mathrm{TV}\left(\mathbb{P}_1, \widetilde{\mathbb{P}}_1\right) \leq e^{-ck} = o(1).$$

Since $\chi^2\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) \to 0$, we also have

$$\mathrm{TV}\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) \to 0.$$

By the triangle inequality,

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) \le \mathrm{TV}\left(\mathbb{P}_1, \widetilde{\mathbb{P}}_1\right) + \mathrm{TV}\left(\widetilde{\mathbb{P}}_1, \mathbb{P}_2\right) \to 0.$$

Thus weak detection is impossible for the original Bernoulli model in the same regime. □

A.8 Proof of Lemma 14.2

Proof of Lemma 14.2. Recall the generating function for the normalized Hermite polynomials:

$$\exp\left(tx - \frac{t^2}{2}\right) = \sum_{k=0}^{\infty} \frac{t^k}{\sqrt{k!}} \widehat{h}_k(x).$$

Substituting $x = \mu + z$ gives

$$\exp\left(t(\mu+z) - \frac{t^2}{2}\right) = e^{t\mu} \exp\left(tz - \frac{t^2}{2}\right).$$

Expanding both factors,

$$e^{t\mu} = \sum_{r=0}^{\infty} \frac{(t\mu)^r}{r!}, \qquad \exp\left(tz - \frac{t^2}{2}\right) = \sum_{\ell=0}^{\infty} \frac{t^\ell}{\sqrt{\ell!}} \widehat{h}_\ell(z).$$

Therefore

$$\sum_{k=0}^{\infty} \frac{t^k}{\sqrt{k!}} \widehat{h}_k(\mu+z) = \sum_{r,\ell \ge 0} \frac{t^{r+\ell} \mu^r}{r!\sqrt{\ell!}} \widehat{h}_\ell(z).$$

Comparing the coefficient of t^k, with $r = k - \ell$, we get

$$\frac{1}{\sqrt{k!}} \widehat{h}_k(\mu+z) = \sum_{\ell=0}^{k} \frac{\mu^{k-\ell}}{(k-\ell)!\sqrt{\ell!}} \widehat{h}_\ell(z).$$

Multiplying by $\sqrt{k!}$ gives

$$\widehat{h}_k(\mu+z) = \sum_{\ell=0}^{k} \frac{\sqrt{k!}}{(k-\ell)!\sqrt{\ell!}} \mu^{k-\ell} \widehat{h}_\ell(z).$$

Since

$$\frac{\sqrt{k!}}{(k-\ell)!\sqrt{\ell!}} = \sqrt{\frac{\ell!}{k!}} \binom{k}{\ell},$$

the claimed identity follows.

Taking expectation over $Z \sim \mathcal{N}(0,1)$, all terms with $\ell \ge 1$ vanish by orthogonality to constants, while $\widehat{h}_0 \equiv 1$. Thus

$$\mathop{\mathbb{E}}_{Z}\left[\widehat{h}_k(\mu+Z)\right] = \frac{\mu^k}{\sqrt{k!}}.$$

□