Title
High-dimensional Principal Component Analysis

Author
Amini, Arash A.

Publication Date
2011

Peer reviewed|Thesis/dissertation

High-dimensional Principal Component Analysis

by

Arash Ali Amini

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Electrical Engineering & Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Martin Wainwright, Chair
Professor Peter Bickel
Professor Michael Jordan

Fall 2011

High-dimensional Principal Component Analysis

Abstract

High-dimensional Principal Component Analysis

by

Arash Ali Amini

Doctor of Philosophy in Engineering-Electrical Engineering & Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

University of California, Berkeley

Associate Professor Martin Wainwright, Chair

Advances in data acquisition and emergence of new sources of data, in recent years, have led to generation of massive datasets in many fields of science and engineering. These datasets are usually characterized by having high dimensions and low number of samples. Without appropriate modifications, classical tools of statistical analysis are not quite applicable in these "high-dimensional" settings. Much of the effort of contemporary research in statistics and related fields is to extend inference procedures, methodologies and theories to these new datasets. One widely used assumption which can mitigate the effects of dimensionality is the sparsity of the underlying parameters. In the first half of this thesis we consider principal component analysis (PCA), a classical dimension reduction procedure, in the high-dimensional setting with "hard" sparsity constraints. We will analyze the statistical performance of two modified procedures for PCA, a simple diagonal cut-off method and a more elaborate semidefinite programming relaxation (SDP). Our results characterize the statistical complexity of the two methods, in terms of the number of samples required for asymptotic recovery. The results show a trade-off between statistical and computational complexity. In the second half of the thesis, we consider PCA in function spaces (fPCA), an infinite-dimensional analog of PCA, also known as Karhunen–Loéve transform. We introduce a functional-theoretic framework to study effects of sampling in fPCA under smoothness constraints on functions. The framework generates high dimensional models with a different type of structural assumption, an "ellipsoid" condition, which can be thought of as a soft sparsity constraint. We provide a M-estimator to estimate principal component subspaces which takes the form of a regularized eigenvalue problem. We provide rates of convergence for the estimator and show minimax optimality. Along the way, some problems in approximation theory are also discussed.

Contents

List of Figures v

1 Introduction 1
- 1.1 Statistics on massive datasets . . . 1
- 1.2 Structured high-dimensional datasets . . . 4
- 1.3 Principal component analysis (PCA) and sparsity . . . 5
  - 1.3.1 Sparse PCA in high-dimensional setting . . . 6
- 1.4 Functional PCA and its sampling problem . . . 8
- 1.5 Organization of the thesis . . . 10

2 Background 11
- 2.1 Vectors, Matrices and their norms . . . 11
  - 2.1.1 Vector ℓ_p norms . . . 11
  - 2.1.2 Classes of matrices and spectral theorems . . . 12
  - 2.1.3 Matrix operator norms . . . 13
- 2.2 Matrix perturbation theory . . . 15
- 2.3 Concentration inequalities . . . 17
- 2.4 PCA and its SDP formulation . . . 22
  - 2.4.1 Classical consistency theory (fixed p, large n) . . . 24
  - 2.4.2 Random Matrix Theory . . . 25
  - 2.4.3 PCA inconsistency in high-dimensional setting . . . 29
- 2.5 Hilbert spaces and reproducing kernels . . . 30
- 2.6 Minimax lower bounds via Fano inequality . . . 35
  - 2.6.1 Decision theory and minimax criterion . . . 36
  - 2.6.2 Reduction by discretization and Bayesian averaging . . . 36
  - 2.6.3 Fano inequality . . . 37
  - 2.6.4 Bounds on mutual information . . . 39
  - 2.6.5 Symmetry and risk flatness . . . 40

3 High-dimensional sparse PCA **41**
3.1 Sparse spiked covariance model . . . 41
3.1.1 Model selection problem . . . 42
3.2 Two methods of support recovery and results . . . 43
3.2.1 Diagonal cut-off . . . 43
3.2.2 Semidefinite-programming relaxation . . . 44
3.2.3 Minimax lower bound . . . 46
3.3 Proof of Proposition 2 – diagonal cut-off . . . 47
3.4 Proof of Theorem 7(b) – SDP relaxation . . . 51
3.4.1 High-level proof outline . . . 51
3.4.2 Sufficient conditions for general noise matrices . . . 53
3.4.3 Noise in sample covariance – identity case . . . 55
3.4.4 Nonidentity noise covariance . . . 62
3.5 Proof of Theorem 8 – minimax lower bound . . . 65
3.6 Some results on ℓ_q sparsity . . . 66
3.6.1 Proof of Theorem 9 . . . 67
Appendix 3.A Large deviations for χ^2 variates . . . 69
Appendix 3.B Proof of Lemma 6 . . . 70
Appendix 3.C Proofs for §3.4.2 . . . 71
3.C.1 Proof of Lemma 8 . . . 71
3.C.2 Proof of Lemma 9 . . . 72
3.C.3 Proof of Lemma 10 . . . 73
Appendix 3.D Proof of Lemma 11 . . . 75
Appendix 3.E Proof of Lemma 16 . . . 78
Appendix 3.F Proof of Theorem 7(a) . . . 80
Appendix 3.G Proofs of §3.6 . . . 81
3.G.1 Proof of Lemma 19 . . . 81
3.G.2 Proof of Lemma 20 . . . 82

4 Approximation properties of operator norms **83**
4.1 Introduction . . . 83
4.1.1 Notation . . . 85
4.2 Background and setup . . . 86
4.2.1 Hilbert spaces . . . 86
4.2.2 Linear operators, semi-norms and examples . . . 88
4.3 Main result and some consequences . . . 89
4.3.1 General upper bounds on $R_\Phi(\varepsilon)$ . . . 89
4.3.2 Some illustrative examples . . . 91
4.4 Proof of Theorem 10 . . . 95
4.5 Conclusion . . . 99
Appendix 4.A Analysis of random time sampling . . . 99

4.A.1 Proof of Corollary 4 . . . 100
4.A.2 Proof of Lemma 25 . . . 101
Appendix 4.B Proof of Lemma 24 . . . 103
Appendix 4.C Relationship between $R_\Phi(\varepsilon)$ and $\underline{T}_\Phi(\varepsilon)$ . . . 105
Appendix 4.D The 2×2 subproblem . . . 107
Appendix 4.E Details of Fourier truncation example . . . 107
Appendix 4.F A quadratic inequality . . . 109

5 Sampled Functional PCA in RKHS **110**
5.1 Background and problem set-up . . . 112
5.1.1 Reproducing Kernel Hilbert Spaces . . . 112
5.1.2 Functional model and observations . . . 113
5.1.3 Approximation-theoretic quantities . . . 115
5.2 M-estimator and implementation . . . 116
5.2.1 M-estimator . . . 117
5.2.2 Implementation details . . . 119
5.3 Main results . . . 121
5.3.1 Subspace-based estimation rates (for $\widehat{\mathfrak{Z}}$) . . . 121
5.3.2 Function-based estimation rates (for $\widehat{\mathfrak{F}}$) . . . 124
5.3.3 Lower bounds . . . 127
5.4 Proof of subspace-based rates . . . 128
5.4.1 Preliminaries . . . 128
5.4.2 Proof of Theorem 11 . . . 130
5.5 Proof of functional rates . . . 133
5.6 Proof of minimax lower bounds . . . 135
5.6.1 Preliminary results . . . 135
5.6.2 Proof of Theorem 13 . . . 135
5.6.3 Proof of Theorem 14 . . . 137
5.7 Discussion . . . 138
Appendix 5.A A special kernel . . . 139
Appendix 5.B Auxiliary lemmas . . . 140
5.B.1 Proof of Lemma 29 . . . 140
5.B.2 Proof of Lemma 30 . . . 140
5.B.3 Proof of Lemma 31 . . . 141
5.B.4 Proof of Lemma 33 . . . 141
5.B.5 Proof of inequality (5.60) . . . 142
Appendix 5.C Proofs for Theorem 11 . . . 142
5.C.1 Derivation of the bound (5.50) . . . 142
5.C.2 Proof of Lemma 36 . . . 143
5.C.3 Proof of Lemma 37 . . . 144
5.C.4 Proof of Lemma 38 . . . 145

5.C.5 Proof of Lemma 39 146
Appendix 5.D Proofs for Theorem 12 146
5.D.1 Proof of Lemma 40 146
5.D.2 Proof of Lemma 41 148
Appendix 5.E Proofs for Theorems 13 and 14 148
5.E.1 Proof of Lemma 42 148
5.E.2 Proof of Lemma 43 149
5.E.3 Proof of Lemma 44 150
Appendix 5.F Suprema involving Gaussian products 150
Appendix 5.G Bounding an operator norm of a Gaussian matrix 153
Appendix 5.H A uniform law 153
Appendix 5.I Some useful matrix-theoretic inequalities 155

List of Figures

1.1 Gene expression data . . . 3
1.2 PCA toy example . . . 6

2.1 ℓ_p norm balls . . . 12
2.2 Marcheko-Pastur density . . . 26

3.1 Success probability plots for diagonal cut-off . . . 45
3.2 Success probability plots SDP . . . 47

4.1 Geometry of Fourier truncation . . . 91
4.2 Sparse periodic Ψ matrices . . . 94
4.3 Geometry of the proof of (4.17) . . . 99
4.4 Geometry of the 2×2 subproblem . . . 108

5.1 Regularized fPCA for time samples. . . . 120

Acknowledgments

I would like to thank my adviser Martin Wainwright for his support, encouragement, advice and doing the heavy lifting in many parts of this thesis and for fueling my interest in statistical theory and academic life in general. I would also like to thank the members of my qualification exam, Peter Bickel, Michael Jordan and Peter Bartlett for the encouragement and advice. I learned a lot from them through their courses, books and papers and they too shaped my interest in the subject.

Special thanks to my parents for their support and help throughout the years, for their kind words and well wishing, for their love and patience, and for abiding by my absence. Special thanks to my wife for her support, encouragement and patience and for lifting much of the burden of life during the final stages of this thesis. Special thanks to my friends, especially Amin Aminzadeh Gohari, Ali Ghazizadeh, Omid Etesami (and numerous others) for their support and help, for the many interesting discussions that we had and without whom life away from home would have been much more difficult. Special thanks to staff of EE department, especially Graduate Affairs, for their helpfulness, their patience and professionalism.

I would also like to thank numerous others who I have learned from during the years, the many great teachers that I have had and the many great friends (most of them scattered around the globe) whose support and well wishing has had an immense impact even from a distance. I wish I could name you all.

Chapter 1

Introduction

1.1 Statistics on massive datasets

The classical goal of statistical data analysis is to extract information (or more ambitiously generate knowledge) from raw datasets, a process which is usually called "statistical inference". This rather general statement is meant to encompass the many problems considered in statistical analysis and its wide area of applications. In particular, the type of information one seeks and the type of data at hand could be quite varied. Among the early datasets considered are for example the results of experimental studies (say drug effectiveness studies) and polling results (categorical data) [40]. Among the more recent examples are:

- the medical and astronomical images (e.g., results of MRI, images of far away galaxies, images of the surface of the Earth),
- the signal received by a cellphone, a radar, etc. [92],
- the firing of a collection of neurons,
- gene expression data [87],
- paths traced by hurricanes [24],
- reading of nodes in a sensor network,
- epidemic graphs,
- documents on the Internet (and the hyperlink graph),
- paths between nodes on the Internet (routing graphs) [72, 38, 12],
- financial time series (e.g., share prices of a collection companies over time),
- voting history of members of a parliament/senate,
- databases of handwritten data, faces, shapes of body organs or other complex objects,

and so on. The trait shared by these very different datasets which makes them targets of statistical analysis is that they are "unorganized" and plagued by noise (of various sources, e.g., measurement noise, noise inherent in the generating process, etc.). More precisely, one suspects that they contain some organization (or regularity or information) obscured by noise and one hopes to uncover it by means of analysis. It is also worth noting that these datasets

are modeled through a myriad of mathematical objects: vectors, functions, matrices, graphs, probability densities, etc.

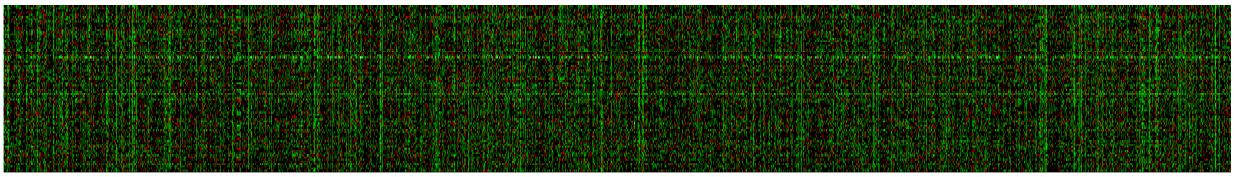
The information one seeks to obtain could be the value of an unknown parameter suspected to have influence on the observed dataset, e.g., the sequence of symbols transmitted in the case of the cellphone signal; it could be a decision to be made, e.g., whether the treatment was effective in an experimental study, whether a target is present in radar detection; it could be an effective (i.e., low complexity) representation of the data revealing patterns and facilitating interpretation and visualization; it could be a classification or clustering task, i.e., grouping similar observations into classes, e.g., classifying a face as male/female, a document according to its content, detecting objects in a scene and so on; it could be a prediction problem, e.g. given the path of a hurricane so far what is the most likely path it is going to take in the future; it could be revealing relationships or connections, e.g., what is a reasonable association graph for members of the parliament given their votes, what is the connectivity graph for nodes of a network?

The statistical aspects[1] of these problems are usually modeled by a family of probability distributions $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ indexed by a *parameter* θ taking values in some parameter space Θ. Each $\mathbb{P}_\theta$ is a probability distribution over the observation space $\mathcal{X}$, which could be a space of vectors, functions or other more fancy mathematical objects. For example, in the voting problem mentioned, Θ can be the space of $\{-1, 0, 1\}$-valued $N_s \times N_b$ matrices and Θ a subspace of the space of all graphs on N_s nodes—where N_s is the number of senators and N_b is the number of bills. It is usually assumed that we have access to n *independent* samples $X := \{x_i\}_{i=1}^n \subset \mathcal{X}$ drawn from $\mathbb{P}_{\theta^*}$, where $\theta^* \in \Theta$ is the unknown *true* parameter or the state of nature. The goal is then to *infer* θ^* from X; that is, to obtain a function of X, say $\widehat{\theta} = \widehat{\theta}(X)$, which serves as an *estimate* of θ^*. We can then compare it to the true value in some appropriate error metric, say $r(\widehat{\theta}, \theta^*) := [\mathbb{E}_{\theta^*} d^2(\widehat{\theta}, \theta^*)]^{1/2}$, providing an assessment of the quality of inference (cf. §2.6.1 and [16, 62]).

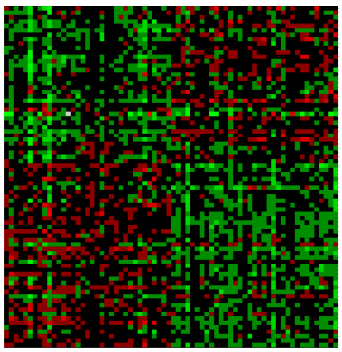
What makes the inference possible? Intuitively, since the "n" samples agree in the regular part (θ^*) and differ in the random part (the noise due to sampling), one can hope that due to independence, the noise can be averaged out as n grows large, revealing the underlying regularity. In mathematical terms, one hopes that the collection $\mathcal{P}^n := \{\mathbb{P}_\theta^{\otimes n} :\in \Theta\}$ of n-fold product measures generated by the family $\mathcal{P}$ becomes well-separated (in some appropriate metric for probability distributions) as $n \to \infty$, allowing for perfect identification of $\theta^* \in \Theta$ that generated the data X.

This is indeed the case in the *classical setting* where Θ is a subset of some finite-dimensional vector space of fixed dimension p, e.g., $\Theta \subset \mathbb{R}^p$. More specifically, asymptotic theory [94] shows that when p is kept constant while $n \to \infty$, the error $r(\widehat{\theta}, \theta^*)$ goes to

[1] Not all aspects are covered by the decision-theoretic model discussed here. For example, in the context of sensor networks, there is the problem of communication between nodes: most nodes have access to their own data, limited processing power and few communication links; there is a need for distribution of information and computation. This raises interesting questions about the rate of information dissemination in the network and its interaction with statistical rates.



(a) Entire dataset



(b) An enlarged subset

Figure 1.1: Gene expression data from a breast cancer study. The top plot is the 78×4918 matrix of expression levels, the rows represent patients (or tumor samples), the columns represent different genes suspected to influence treatment. The bottom plot shows the subset of data corresponding to 70 genes among those found to be a good signature for treatment success.

zero for many reasonable estimators $\widehat{\theta}$, i.e., they are *consistent* estimators of θ^*. In fact, one shows that the optimal convergence rate is $\mathcal{O}(n^{-1/2})$. The estimators that achieve this rate are called $\sqrt{n}$-*consistent*; they are asymptotically efficient[2] or optimal in a statistical sense. This sort of asymptotics is well-suited for approximating problems of classical statistics, where $n \gg p$. As an example consider a polling problem, where we ask a population of roughly $n \approx 1000$, which of the $p \approx 5$ candidates they are going to vote for in an election.

The situation is quite the opposite for most modern datasets mentioned earlier. Fig. 1.1(a) shows an example of gene expression data from a breast cancer study conducted by van't Veer et. al [95]. The observed dataset, X, collects the expression levels of $p = 4918$ genes in tumor samples from $n = 78$ patients, organized as a $n \times p$ matrix. Out of the 78 patients, 34 exhibited recurrence in a 5-year period after treatment; that is, the samples may be thought of as coming from two populations, say, $\mathbb{P}_{\theta_1}$, modeling gene sequences in patients that exhibited recurrence and $\mathbb{P}_{\theta_2}$, modeling those in patents that did not. The goal was to find a subset of genes that could reliably differentiate between the two groups and hence could be used as a predictor of treatment success. Here, we have $n \ll p$ and consequently classical asymptotic results are poor approximations for this problem. Fig. 1.1(b) shows the part of X corresponding to a subset of genes of size $n_1 = 70$ which are part of the signature found

[2] An estimator is statistically efficient roughly means that among all estimators it requires the least number of samples to achieve a certain level of error.

by the study to be a good predictor. Even in this reduced dataset, one has $n \approx p$, rendering classical asymptotics useless.

1.2 Structured high-dimensional datasets

For problems in which $n \ll p$, a more appropriate form of asymptotics is obtained by letting both n and p go to infinity simultaneously. This will be referred to as the *high-dimensional setting*. A rough calculation shows that the error $r(\widehat{\theta}, \theta^*)$ will be of the order $\mathcal{O}(\sqrt{p/n})$ for optimal estimators. Thus, as long as $p/n \to 0$ as $(n, p) \to \infty$, we cannot hope for consistency (cf. §2.4.3).

But $p/n \to 0$ is not a good model for our problems. We would like to let $p/n \to \gamma > 0$ or even $p/n \to \infty$ and still be able to consistently estimate the parameter. The key is to impose additional structure on the parameter, so that the parameter space, Θ, has some low *effective* dimension. A simple such structure which is both natural and popular is *sparsity*. Returning to our example in which $\Theta \subset \mathbb{R}^p$, we could further impose

$$\Theta \subset \left\{\theta \in \mathbb{R}^p : \ |\{j : \ [\theta]_j \neq 0\}| \leq s\right\} \tag{1.1}$$

for some $s \ll p$. (Here and elsewhere, $[\theta]_j$ applied to a vector θ is its j-th entry and $|A|$ applied to a set A is its cardinality.) In other words, all $\theta \in \Theta$ have at most "s" nonzero entries. In yet other words, all $\theta \in \Theta$ lie in the union of $\binom{p}{s}$ coordinate subspaces, each of which is s-dimensional. A rough estimate then suggests that if $\frac{s \log \binom{p}{s}}{n} \approx \frac{s^2 \log p}{n}$ goes to zero, we can hope for consistency. Rigorous results of this sort will appear in Chapter 3 when we discuss high-dimensional sparse PCA.

Assumption (1.1) is sometimes referred as a *hard sparsity* constraint suggesting that there are also *soft* measures of sparsity; notable among them is an ℓ_q^p-ball constraint,

$$\Theta \subset \left\{\theta \in \mathbb{R}^p : \ \sum_{j=1}^{p} \left|[\theta]_j\right|^q \leq C_q\right\} \tag{1.2}$$

for some $q < 1$. This model allows for a more graceful drop of entries of θ to zero. Although our focus will be mostly on hard sparsity, some results related to model (1.2) will be discussed in Chapter 3 in the context of PCA.

Another structure is an *ellipsoid* condition which arises naturally when the parameter is a function living in some class of functions and one imposes a *smoothness* condition on the class. When the function class is infinite-dimensional, the natural space for the parameter θ is usually the space of infinite sequences $\theta = (\theta_1, \theta_2, \dots)$ with some constraint, say, on their energy $\sum_{j=1}^{\infty} \theta_j^2 < \infty$; this space is referred to as ℓ_2 sequence space. The ellipsoid condition is then,

$$\Theta \subset \left\{\theta \in \ell_2 : \ \sum_{j=1}^{\infty} \frac{[\theta]_j^2}{\mu_j} \leq C_\mu\right\} \tag{1.3}$$

for a null positive sequence $\mu = (\mu_1, \mu_2, \dots)$: $\mu_j > 0$ and $\mu_j \to 0$ as $j \to \infty$. Since for large j, μ_j is small, so is the corresponding $[\theta]_j$. Hence, the ellipsoid condition too can be thought of as a soft sparsity assumption. Results for models of the form (1.3) will appear in Chapter 5 in the context of our discussion of functional PCA.

To summarize, all these models impose conditions implying that despite the true parameter living in a high-dimensional or even infinite-dimensional space, its truly significant part is of relatively small size (low-dimensional). This is somehow the essence of sparsity and it is a plausible assumption for many real-world datasets. As mentioned earlier, a sparse vector lies in a union of low-dimensional (linear) subspaces. One possible direction of generalization is to allow it to lie on a (globally nonlinear) submanifold of low dimension or on a union of such submanifolds.

1.3 Principal component analysis (PCA) and sparsity

As the title suggests, our focus in this thesis will be on principal component analysis (PCA) in the high-dimensional setting (and in function spaces). PCA is a classical method for reducing dimension, say from a subset of $\mathbb{R}^p$ to some subset of $\mathbb{R}^d$ where $d \ll p$, and is frequently used to obtain a low-dimensional representation of a dataset. It operates by projecting the data onto the d directions of maximal variance, captured by eigenvectors of the $p \times p$ population covariance matrix Σ. Of course, in practice, one does not have access to the population covariance, but instead must rely on a "noisy" version of the form

$$\widehat{\Sigma} = \Sigma + \Delta \tag{1.4}$$

where $\Delta = \Delta_n$ denotes a random noise matrix, typically arsing from having only a finite number n of samples. We usually take $\widehat{\Sigma}$ to be the sample covariance, which for a dataset $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$ with population mean zero, is given by $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$. Fig. 1.2 shows a toy example with $p = 3$ and $d = 1$ and $n = 200$.

As with any statistical procedure, a natural question when using the noisy version $\widehat{\Sigma}$ is the issue of consistency, i.e., under what conditions the sample eigenvectors (i.e., based on $\widehat{\Sigma}$) are consistent estimators of their population analogues. In the classical theory of PCA, the model dimension p is viewed as fixed, and asymptotic statements are established as the number of observations n tends to infinity. With this scaling, the influence of the noise matrix Δ dies off, so that sample eigenvectors and eigenvalues are consistent estimators of their population analogues [7] (cf. §2.4.1). However, such "fixed p, large n" scaling is inappropriate, as discussed in §1.2, for many contemporary applications in science and engineering (e.g., financial time series, astronomical imaging, sensor networks), in which the model dimension p is comparable or even larger than the number of observations n. This type of high-dimensional scaling causes dramatic breakdowns in standard PCA and related eigenvector methods, as shown by classical and ongoing work in random matrix theory [43, 55, 58] (cf. 2.4.2).

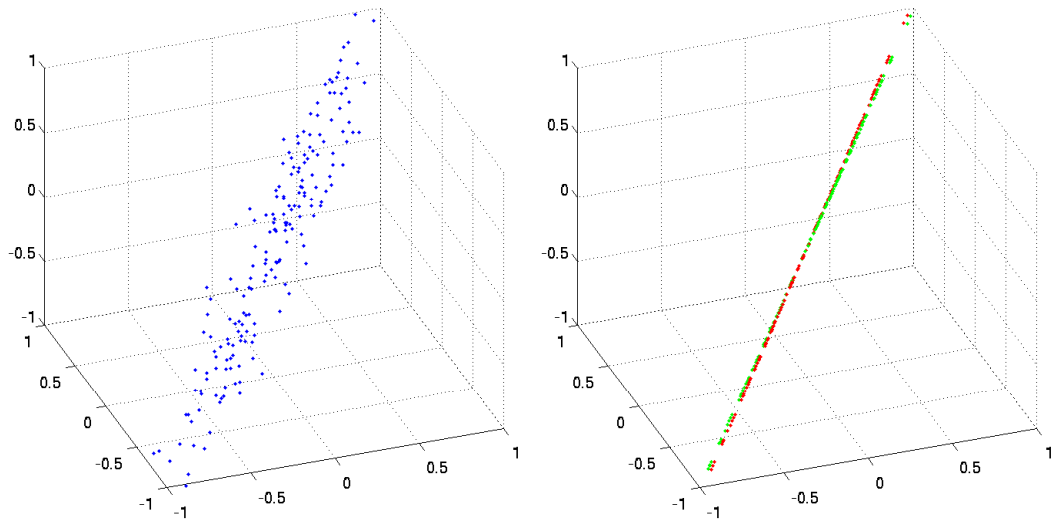


Figure 1.2: PCA toy example. (a) is the scatter-plot of some $n = 200$ points in $\mathbb{R}^3$ generated from a spiked covariance model with one component (cf. Chapter 3). The data clearly concentrates along a line through the origin (i.e., a 1-dimensional subspace). (b) shows the data after projection onto two 1-dimensional subspaces, generated by the true principal component (PC), the green points, and by the sample PC based on the sample covariance, the red points. Qualitatively, the sample version is a very good estimate of the true PC, as one expects for a classical scaling $n = 100 \gg 3 = p$.

As discussed in §1.2, it is possible to add a natural structural assumption, namely sparsity, to help mitigate the curse of dimensionality. Various types of sparse models have been studied in past statistical work. There is a substantial and on-going line of work on subset selection and sparse regression models [e.g., 26, 37, 69, 90, 100], focusing in particular on the behavior of various ℓ_1-based relaxation methods. Other work has tackled the problem of estimating sparse covariance matrices in the high-dimensional setting, using thresholding methods [17, 39] as well as ℓ_1-regularization methods [31, 108].

1.3.1 Sparse PCA in high-dimensional setting

A related problem—and the primary focus of Chapter 3—is recovering sparse eigenvectors from high-dimensional data. While related to sparse covariance estimation, the sparse eigenvector problem presents a different set of challenges; indeed, a covariance matrix may have a sparse eigenvector with neither it (nor its inverse) being a sparse matrix. Various researchers have proposed methods for extracting sparse eigenvectors, a problem often referred to as sparse principal component analysis (SPCA). Some of these methods are based on greedy or nonconvex optimization procedures (e.g., [59, 109, 71]), whereas others are based on various types of ℓ_1-regularization [110, 32]. Zou et al. [110] develop a method based on transforming the PCA problem to a regression problem, and then applying the Lasso (ℓ_1-regularization). Johnstone and Lu [58] proposed a two-step method, using an initial pre-

processing step to select relevant variables followed by ordinary PCA in the reduced space. Under a particular ℓ_q-ball sparsity model, they proved ℓ_2-consistency of their procedure as long as p/n converges to a constant. In recent work, d'Asprémont et al. [32] have formulated a direct semidefinite programming (SDP) relaxation of the sparse eigenvector problem, and developed fast algorithms for solving it, but have not provided high-dimensional consistency results. The elegant work of Paul and Johnstone [73, 75] studies estimation of eigenvectors satisfying weak ℓ_q-ball sparsity assumptions for $q \in (0, 2)$. We discuss connections to this work at more length below.

In Chapter 3, we study the *model selection* problem for sparse eigenvectors. More precisely, we consider a spiked covariance model [55], in which the maximal eigenvector z^* of the population covariance $\Sigma_p \in \mathbb{R}^{p \times p}$ is k-sparse, meaning that it has nonzero entries on a subset $S(z^*)$ with cardinality k, and our goal is to recover this support set exactly. In order to do so, we have access to a matrix $\widehat{\Sigma}$, representing a noisy version of the population covariance, as in equation (1.4). Although our theory is somewhat more generally applicable, the most natural instantiation of $\widehat{\Sigma}$ is as a sample covariance matrix based on n i.i.d. samples drawn from the population. We analyze this setup in the high-dimensional regime, in which all three parameters—the number of observations n, the ambient dimension p and the sparsity index k—are allowed to tend to infinity simultaneously. Our primary interest is in the following question:

> Using a given inference procedure, under what conditions on the scaling of triplet (n, p, k) is it possible, or conversely impossible, to recover the support set of the maximal eigenvector z^* with probability one?

We provide a detailed analysis of two procedures for recovering sparse eigenvectors:

(a) a simple *diagonal thresholding (or cut-off)* method, used as a pre-processing step by Johnstone and Lu [58], and

(b) a *semidefinite programming (SDP)* relaxation for sparse PCA, recently developed by d'Aspremont et al. [32].

Under the k-sparsity assumption on the maximal eigenvector, we prove that the success/failure probabilities of these two methods have qualitatively different scaling in terms of the triplet (n, p, k). For the diagonal thresholding method, we prove that its success/failure is governed by the rescaled sample size

$$\theta_{\mathrm{dia}}(n, p, k) \quad := \quad \frac{n}{k^2 \log(p - k)}, \tag{1.5}$$

meaning that it succeeds with probability one for scalings of the triplet (n, p, k) such that θ_{dia} is above some critical value and, conversely, fails with probability one when this ratio falls below some critical value (cf. Proposition 2). We then establish performance guarantees for

the SDP relaxation [32]: in particular, for the same class of models, we show that it always has a unique rank-one solution that specifies the correct signed support once $\theta_{\text{dia}}(n, p, k)$ is sufficiently large, and moreover, that for sufficiently large values of the rescaled sample size

$$\theta_{\text{sdp}}(n, p, k) \quad := \quad \frac{n}{k \log(p-k)}, \tag{1.6}$$

if there exists a rank-one solution, then it specifies the correct signed support (cf. Theorem 7). The proof of this result is based on random matrix theory, concentration of measure, and Gaussian comparison inequalities (cf. § 2.4.2 and §2.3 for related background). Our final contribution, regarding the spiked covariance model with hard sparsity constraint, is to use information-theoretic arguments to show that no method can succeed in recovering the signed support for the spiked identity covariance model if the order parameter $\theta_{\text{sdp}}(n, p, k)$ lies below some critical value (cf. Theorem 8). One consequence is that the given scaling (1.6) for the SDP relaxation is sharp, meaning the SDP relaxation also fails once θ_{sdp} drops below a critical threshold. Moreover, it shows that under the rank-one condition, the SDP is in fact statistically optimal, i.e., it requires only the necessary number of samples (up to a constant factor) to succeed.

Computational vs. statistical efficiency

Our results, in Chapter 3, highlight some interesting trade-offs between computational and statistical costs in high-dimensional inference. On one hand, the statistical efficiency of SDP relaxation is substantially greater than the diagonal thresholding method, requiring $\mathcal{O}(1/k)$ fewer observations to succeed. However, the computational complexity of SDP is also larger by roughly a factor $\mathcal{O}(p^3)$: an implementation due to d'Asprémont et al. [32] has complexity $\mathcal{O}(np + p^4 \log p)$ as opposed to the $\mathcal{O}(np + p \log p)$ complexity of the diagonal thresholding method. Moreover, our information-theoretic analysis shows that the best possible method—namely, one based on an exhaustive search over all $\binom{p}{k}$ subsets, with exponential complexity—does not have substantially greater statistical efficiency than the SDP relaxation. The following table summarizes these results:

Method	Computational complexity	Statistical complexity
Diagonal cut-off	$\mathcal{O}(np + p \log p)$	$\mathcal{O}\big(k^2 \log(p-k)\big)$
SDP relaxation	$\mathcal{O}(np + p^4 \log p)$	$\mathcal{O}\big(k \log(p-k)\big)$

1.4 Functional PCA and its sampling problem

The second half of this thesis considers mainly the functional version of PCA, or fPCA for short. fPCA lies within the broader field of functional data analysis (FDA), that is, statistical analysis of data which can be modeled as functions. FDA is an established field in

statistics with a great number of practical applications [80, 81]. When the data is available as finely sampled curves, say in time, it is common to treat it as a collection of continuous-time curves or functions, each being observed in totality. These datasets are then termed "functional" and various statistical procedures applicable in finite-dimension are extended to be applicable to them, among which is the principal component analysis (PCA). By the infinite-dimensional (or nonparametric) nature of function spaces, however, new phenomena might also be expected. In particular, statistical analysis in infinite-dimension may provide some insights into aspects of the "high-dimensional" (parametric) setting which has been the focus of much recent work in theoretical statistics, as discussed in §1.2 and §1.3.

If one thinks of continuity as a mathematical abstraction of reality, then treating functional data as continuous curves is arguably a valid modeling device. However, in practice, one is faced with finite computational resources and is forced to implement a (finite-dimensional) approximation of true functional procedures by some sort of truncation of functions, say in frequency-domain. It is then important to understand the effects of this truncation on the statistical performance of the procedure. In other situations, for example in longitudinal data analysis [36], a continuous curve model is justified as a hidden underlying generating process to which one has access only through sparsely sampled, corrupted by noise perhaps, measurements in time. Studying how the time-sampling affects the estimation of the underlying functions in the presence of noise has some elements in common with that of the frequency-domain problem mentioned above.

The generalized sampling

The aim of the second half of this thesis, which constitutes Chapter 5, is to study effects of "sampling" on fPCA in smooth function spaces. We take a functional-theoretic approach to sampling by treating the sampling procedure as a (continuous) linear operator. This provides us with a notion of sampling general enough to treat both the frequency-truncation and time-sampling in the context of a unified framework. We take as our smooth function space a Hilbert subspace $\mathcal{H}$ of $L^2[0,1]$ and denote the sampling operator by $\Phi : \mathcal{H} \to \mathbb{R}^m$. We assume that there are functions $x_i(t)$, $t \in [0,1]$ in $\mathcal{H}$ for $i = 1, \ldots, n$, generated i.i.d. from a probabilistic model (to be discussed). We then observe the collection $\{\Phi x_i\}_{i=1}^n \subset \mathbb{R}^m$ in noise. We refer to the index n as the number of *statistical samples*, and to the index m as the number of *functional samples*.

We analyze a natural M-estimator which takes the form of a regularized PCA in $\mathbb{R}^m$ and provide rates of convergence in terms of n and m. The eigen-decay of two operators govern the rates, the product of Φ and its adjoint Φ^* and the product of the map embedding $\mathcal{H}$ in L^2 and its adjoint. These eigenvalues will determine ellipsoid models of the form (1.3) discussed 1.2. Our focus will be on the setting where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), in which case the two eigen-decays are intimately related through the kernel function $(s,t) \mapsto \mathbb{K}(s,t)$. In such cases, the two components of the rate interact and give rise to optimal values for the number of functional samples (m) in terms of the number

of statistical samples (n) or vice versa. This has practical appeal in cases where obtaining either type of samples is costly.

A particular rate of eigenvalue decay which concerns is the polynomial-α decay,

$$\mu_j \asymp \frac{1}{j^{2\alpha}}, \quad \text{for } j = 1, 2, \ldots \tag{1.7}$$

where $\{\mu_j\}$ are the relevant eigenvalues and $\alpha > 1/2$. For this type of decay, the rates of convergence for the two examples of time sampling and frequency truncation will be worked out in detail by specializing our general results to these two operators. Under suitable conditions we obtain the following rates,

time sampling	frequency truncation
$\left(\frac{1}{mn}\right)^{\frac{2\alpha}{2\alpha+1}} + \left(\frac{1}{m}\right)^{2\alpha}$	$\left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + \left(\frac{1}{m}\right)^{2\alpha}$

which are then shown to be minimax optimal. The interplay between the two types of sample (statistical versus functional) is clear from the table.

In deriving the rates of convergence in function space, more precisely in the L^2 norm, we encounter some problems which are approximation-theoretic in nature. In particular, we will need to understand how well the (semi)norm defined by $\|f\|_\Phi := \|\Phi f\|_2 = (\sum_{j=1}^m [\Phi f]_j)^{1/2}$ approximates the L^2 norm $\|f\|_{L^2} := (\int_0^1 f^2(t)dt)^{1/2}$. This type of approximation problem will be discussed as an interlude in Chapter 4, which also serves as an introduction to the Hilbert space setup for fPCA in Chapter 5.

1.5 Organization of the thesis

We start with some general background material in Chapter 2. This chapter reviews some aspects of vector-matrix analysis and its infinite-dimensional extensions (i.e., functional analysis) and sets the notation for the subsequent chapters. There are also material on concentration inequalities and Fano's inequality which are commonly used to establish upper and lower bounds on performance of estimators. Section 2.4 of this chapter contains a detailed introduction to PCA and some discussion of classical random matrix theory and high-dimensional effects.

In Chapter 3, we study sparse PCA in high-dimensional setting by analyzing diagonal cut-off and SDP relaxation methods and providing minimax lower bounds. This material appears in in [3]. Section 3.6 contains some unpublished analysis of SDP under the ℓ_q sparsity assumption. Chapter 4 is devoted to the study of the approximation problem related to $\|\cdot\|_\Phi$ and $\|\cdot\|_{L^2}$; the material appears in [4]. Chapter 5 contains the analysis of the sampling problem for functional PCA whose material appear in [5].

Chapter 2

Background

This chapter serves a two-fold purpose: to review some of the concepts and tools frequently used throughout the thesis and to establish our notation. We try to present some of the topics in their simpler forms. More refined versions will appear in the subsequent chapters and in appendices as needed.

We start with our conventions regarding matrix norms and special classes of matrices. We assume that the reader is familiar with linear algebra, basic matrix operations and finite-dimensional operator theory (e.g. [15, 54]).

2.1 Vectors, Matrices and their norms

2.1.1 Vector ℓ_p norms

The standard n-dimensional Euclidean space is denoted as $\mathbb{R}^n$. For a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and $p \in (0, \infty]$, we use

$$\|x\|_p := \begin{cases} \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} & p < \infty \\ \max_{1 \le i \le n} |x_i| & p = \infty \end{cases}, \tag{2.1}$$

to denote its ℓ_p (quasi)norm. To emphasize the (quasi)norm, we will denote $\mathbb{R}^n$ equipped with $\|\cdot\|_p$ as ℓ_p^n. Strictly speaking, $\|\cdot\|_p$ is a norm for $p \in [1, \infty)$ and only a quasinorm for $p \in (0, 1)$, as it does not satisfy the triangle inequality in the later case [61]. For simplicity, we will omit the "quasi" prefix and call $\|\cdot\|_p$ a norm for all $p \in (0, \infty]$. Fig. 2.1 illustrates the unit ball of ℓ_p^n, defined as

$$B_p^n := \left\{ x \in \mathbb{R}^n \ : \ \|x\|_p \le 1 \right\}, \tag{2.2}$$

for some values of p (and $n = 2$). Note that the unit ball is not convex for $p \in (0, 1)$. The ℓ_p norms in these cases are often used as soft measures of sparsity, as opposed to the hard

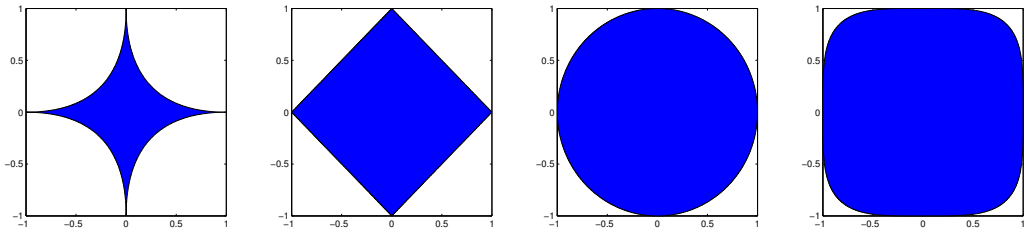


Figure 2.1: ℓ_p norm balls. Plots of the unit balls B_p^2 for $p = \frac{1}{2}, 1, 2, 4$, from left to right.

sparsity which is measured by the cardinality of $x \in \mathbb{R}^n$, defined as

$$\text{card}(x) := |\{i \ : \ x_i \neq 0\}|; \tag{2.3}$$

this is sometimes also called the ℓ_0 norm of x.

We will say that a vector $x \in \mathbb{R}^n$ is sparse, if $\text{card}(x) \ll n$. Similarly, we will say that $x \in \mathbb{R}^n$ is sparse in the "ℓ_p sense", for some $p \in (0, 1)$, if $\|x\|_p$ is relatively small. Here, relative means relative to the dimension n, assuming some fixed normalization, say $\|x\|_\infty = 1$.

For any $p \in [1, \infty]$, let p' denote its *Hölder conjugate*, defined via the relation $1/p+1/p' = 1$. Then, we have the Hölder inequality [42, Chap. 5]: for $x, y \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} |x_i y_i| \leq \|x\|_p \|y\|_{p'}. \tag{2.4}$$

In the special case $p = p' = 2$, (2.4) reduces to the Cauchy-Schwarz inequality. Another special case of interest for us is when $p = 1$ and $p' = \infty$. Applying (2.4) with $y_i = 1$, we obtain the upper bound in the useful relation

$$\|x\|_p \leq \|x\|_1 \leq n^{1/p'} \|x\|_p, \tag{2.5}$$

which holds for $p \in [1, \infty]$ and $x \in \mathbb{R}^n$. The lower bound is obtained by noting that for $p \in (1, \infty)$, the map $\alpha \mapsto \alpha^{p-1}$ is increasing on $(0, \infty)$. (Without loss of generality, assume $\sum_i |x_i| = 1$. Then, $|x_i|^{p-1} \leq 1$ for all i and $\sum_i |x_i|^p = \sum_i |x_i||x_i|^{p-1} \leq \sum_i |x_i| = 1$. The case $p = \infty$ is trivial.)

2.1.2 Classes of matrices and spectral theorems

We denote the class of real-valued m-by-n matrices as $\mathbb{R}^{m \times n}$, the class of n-by-n (real-valued) symmetric matrices as $\mathbb{S}^n$ and the class of n-by-n positive semidefinite (PSD) matrices as $\mathbb{S}^n_+$. Recall that a square matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^T$ where A^T is the transpose of A, and is PSD if in addition $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$. Thus, we have the inclusions $\mathbb{S}^n_+ \subset \mathbb{S}^n \subset \mathbb{R}^{n \times n}$.

The class $\mathbb{R}^{n \times n}$ is a vector space (with the usual matrix addition and scalar multiplication) and can be identified with $\mathbb{R}^{n^2}$ and hence has dimension n^2. The class $\mathbb{S}^n$ is subspace of $\mathbb{R}^{n \times n}$

of dimension $n(n+1)/2$ and $\mathbb{S}^n_+$ is a convex cone [53] in this subspace. The cone structure of $\mathbb{S}^n_+$ induces a natural partial order (sometimes called *Löwner oder*) which we denote by $\succeq$ and $\preceq$. More precisely, $A \succeq B$ and $B \preceq A$ if $A - B \in \mathbb{S}^n_+$.

We denote a generic eigenvalue of $A \in \mathbb{R}^{n\times n}$ as $\lambda(A)$. The minimum and maximum eigenvalues are denoted as $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. The eigenvalues in decreasing order are referred to as $\lambda_1^{\downarrow}(A) \geq \lambda_2^{\downarrow}(A) \geq \cdots \geq \lambda_n^{\downarrow}(A)$. Any member of the set of eigenvectors of A associated with an eigenvalue is denoted as $\vartheta(A)$. Similarly, $\vartheta_{\max}(A)$ represents any eigenvector associated with the maximal eigenvalue (occasionally referred to as a "maximal eigenvector") and $\vartheta_j^{\downarrow}(A)$ any eigenvector associated with $\lambda_j^{\downarrow}(A)$. We always assume that eigenvectors are normalized to unit ℓ_2-norm, and have a nonnegative first component. The sign convention guarantees uniqueness of the eigenvector associated with an eigenvalue with geometric multiplicity one.

By the spectral theorem [15, Chap. 1], any (real) symmetric matrix $A \in \mathbb{S}^n$ has a decomposition of the form

$$A = U\Lambda U^T \tag{2.6}$$

where Λ is diagonal with eigenvalues of A on its diagonal and $U \in \mathbb{R}^{n\times n}$ is an *orthogonal matrix* collecting the corresponding eigenvectors. Our notation for a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ with diagonal entries $\{\lambda_i\}_{i=1}^n$ is $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. By definition, an orthogonal matrix $U \in \mathbb{R}^{n\times n}$ is such that $U^TU = UU^T = I_n$ where I_n denotes the n-by-n identity matrix. The class of n-by-n orthogonal matrices will be denoted as $\mathbb{O}^n$. We refer to (2.6) as the *eigenvalue decomposition* (EVD), eigen-decomposition or spectral decomposition of A. Note the by

As a consequence of the spectral decomposition, for any matrix $A \in \mathbb{R}^{m\times n}$ one obtains a decomposition

$$A = U\Sigma V^T \tag{2.7}$$

where $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal, $U \in \mathbb{O}^m$ and $V \in \mathbb{O}^n$. The nonzero diagonal entries of Σ are called the *singular values* of A and denoted as $\{\sigma_i(A)\}$. They are positive by definition and their number is equal to the rank of A, denoted as $\mathrm{rank}(A)$. For example, if $m \leq n$ and A is full rank, we have $\Sigma = \begin{bmatrix}\Sigma_1 & 0\end{bmatrix}$ where $\Sigma_1 = \mathrm{diag}(\sigma_1(A), \ldots, \sigma_m(A))$. We refer to (2.7) as the singular value decomposition (SVD) of A.

Recalling the definition of a PSD matrix, we note that for $A \succeq 0$ and any U of compatible dimension, we have $U^TAU \succeq 0$. In particular, it follows from (2.6) that $\Lambda \succeq 0$, that is, all the eigenvalues of a PSD matrix are nonnegative.

2.1.3 Matrix operator norms

Now consider the class of real-valued m-by-n matrices, denoted as $\mathbb{R}^{m\times n}$. For a matrix $A \in \mathbb{R}^{m\times n}$, we use $|\!|\!| A |\!|\!|_{p,q}$ to denote its operator norm, when viewed as an operator $A : \ell_q^n \mapsto \ell_p^m$;

more precisely, we have

$$\|\!|A|\!\|_{p,q} := \max_{\|x\|_q=1} \|Ax\|_p. \tag{2.8}$$

When $p = q$, we also use $\|\!|A|\!\|_p = \|\!|A|\!\|_{p,p}$. A few cases of particular interest in this thesis are

- the *spectral norm*: $\|\!|A|\!\|_2 = \|\!|A|\!\|_{2,2} = \max_i\{\sigma_i(A)\}$,
- the ℓ_∞ *operator norm*: $\|\!|A|\!\|_\infty = \|\!|A|\!\|_{\infty,\infty} = \max_{i=1,\ldots,m} \sum_{j=1}^n |A_{ij}|$,
- the ℓ_1 *operator norm*: $\|\!|A|\!\|_1 = \|\!|A|\!\|_{1,1} = \max_{j=1,\ldots,n} \sum_{i=1}^m |A_{ij}|$,

where $\{\sigma_i(A)\}$ are the singular values of A (see §2.1.2). The above expression for the spectral norm follows from the definition (2.8), the SVD (2.7) and the invariance of the ℓ_2 norm under orthogonal transformations.

For a symmetric matrix $A \in \mathbb{S}^n$, one has another useful expression for the spectral norm,

$$\|\!|A|\!\|_2 = \sup\big\{|\lambda| \ : \ \lambda \text{ is an eigenvalue of } A\big\}, \tag{2.9}$$

which is a consequence of the EVD (2.6). The right-hand side of the above equation is called the *spectral radius* and is denoted as $\mathrm{spr}(A)$. In general, for a non-symmetric matrix A, one only has $\mathrm{spr}(A) \leq \|\!|A|\!\|_2$. Finally, for a PSD matrix $A \in \mathbb{S}^n_+$, we get form (2.9) that $\|\!|A|\!\|_2 = \lambda_{\max}(A)$.

As a consequence of the definition (2.8), for any vector $x \in \mathbb{R}^n$, we have

$$\|Ax\|_p \leq \|\!|A|\!\|_{p,q}\|x\|_q, \tag{2.10}$$

a property referred to as $\|\!|\cdot|\!\|_{p,q}$ being *consistent* with vector norms $\|\cdot\|_p$ and $\|\cdot\|_q$. By using (2.10) twice, it follows that operator norms are consistent with themselves,

$$\|\!|AB|\!\|_{p,q} \leq \|\!|A|\!\|_{p,r}\|\!|B|\!\|_{r,q}. \tag{2.11}$$

When all the norms in (2.11) are the same, this is called its *submultiplicative property*.

Recall that the *trace* of a square matrix $A \in \mathbb{R}^{n\times n}$ is given by $\mathrm{tr}(A) = \sum_i A_{ii}$. Given two square matrices $X, Y \in \mathbb{R}^{n\times n}$, we define the matrix inner product

$$\langle\!\langle X, Y \rangle\!\rangle \ := \ \mathrm{tr}(XY^T) = \sum_{i,j} X_{ij}Y_{ij}. \tag{2.12}$$

This inner product induces the *Hilbert-Schmidt norm* $\|\!|X|\!\|_{\mathrm{HS}} := \sqrt{\langle\!\langle X, X\rangle\!\rangle}$ (also called the Forbenius norm). It is not hard to see, using the SVD of X, that $\|\!|X|\!\|_{\mathrm{HS}} = \big(\sum_{i=1}^n \sigma_i^2(X)\big)^{1/2}$.

Note that both the spectral and the Hilbert-Schmidt norms depend only on the singular values of the matrix. Any such matrix norm is *unitarily invariant* in the sense that for any two orthogonal matrices U and V (of proper dimensions), one has

$$\|\!|UXV|\!\| = \|\!|X|\!\|. \tag{2.13}$$

Another useful example of a unitarily invariant norm is the nuclear norm $\|\!|X|\!\|_* := \sum_i \sigma_i(X)$ which is dual to the spectral norm with respect to the inner product (2.12).

Vector norms on matrices

We occasionally find it useful to apply vector ℓ_p norms to matrices, treating them as vectors, say, by stacking their columns on top of each other. [As the ℓ_p norms are symmetric, there is no ambiguity in arrangement of elements in passing from a matrix to a vector.] Our notation of using double bars $\|\cdot\|$ for vector norms and triple bars $|||\cdot|||$ for matrix norms should leave no confusion as to which type of norm is being applying to the matrix.

For example, for $B \in \mathbb{R}^{n\times k}$, we have

$$\|B\|_\infty := \max_{1\le i\le n,\ 1\le j\le k} |B_{ij}|.$$

The following "mixed-norm" inequality will be useful to us in §3.F,

$$\|AB\|_\infty \le |||A|||_{\infty,\infty} \|B\|_\infty, \tag{2.14}$$

where $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$. For the proof, let $b_1, \ldots, b_k$ denote the columns of B. Then,

$$\begin{aligned}\|AB\|_\infty = \|\,[Ab_1, \ldots, Ab_k]\,\|_\infty &= \max_{1\le i\le k} \|Ab_i\|_\infty \\ &\le |||A|||_{\infty,\infty} \max_{1\le i\le k} \|b_i\|_\infty = |||A|||_{\infty,\infty}\, \|B\|_\infty\end{aligned}$$

where we have used (2.10). For more details, see any of the standard books [54, 89].

2.2 Matrix perturbation theory

Occasionally, one has a matrix with known eigen-decomposition which is perturbed by a relatively unknown (noise) matrix and one wants to find out how close the eigenvalues and eigenvectors of the perturbed matrix are to the original one. In this section, we collect some inequalities bounding such deviations. There are many interesting approaches to deriving perturbation inequalities. We restrict ourselves to an illustrative sample. For more details, see [15, 89].

The first result is Weyl theorem on perturbation of eigenvalues [15, Cor. III.2.6]. Recall that $\{\lambda_j^\downarrow(A)\}$ denotes the eigenvalues of matrix A in nonincreasing order.

Theorem 1. *(Weyl) Let $A, B \in \mathbb{S}^n$. Then,*

$$\max_{1\le j\le n} |\lambda_j^\downarrow(A) - \lambda_j^\downarrow(B)| \le |||A - B|||_2. \tag{2.15}$$

Next, we consider perturbation of spectral (or invariant) subspaces. As an example, one has Davis-Kahan "$\sin\Theta$ theorem". We will use Bhatia's notation [15, p. 211] regarding projection operators for the remainder of this section: for $S \subset \mathbb{R}$ and a symmetric matrix A, we write $P_A(S)$ to denote the orthogonal projection onto the eigenspace of A corresponding to its eigenvalues that lie in S.

Let A and B be symmetric matrices, S_1 an interval $[a, b] \subset \mathbb{R}$ and $S_2 = (-\infty, a-\delta] \cup [b+\delta, \infty)$, for some $\delta > 0$. Let $E := P_A(S_1)$ and $F := P_B(S_2)$. Then, one has the following [15, Theorem VII.3.1, p. 211].

Theorem 2. *(Davis-Kahan)*

$$\|\!|EF|\!\|_2 \leq \frac{1}{\delta}\|\!|A - B|\!\|_2. \tag{2.16}$$

Let $F^\perp := I - F$ be the projection onto the orthogonal complement of $\operatorname{ran} F$. As a consequence of the CS-decomposition [15, Theorem VII.1.8], the nonzero singular values of $E - F^\perp$ are the nonzero singular values of EF repeated twice. In particular, $\|\!|E - F^\perp|\!\|_2 = \|\!|EF|\!\|_2$. The nonzero singular values of EF are themselves the sines of the *canonical angles* between the subspace $\operatorname{ran} E$ and $\operatorname{ran} F^\perp$, hence the name "$\sin\Theta$ theorem".

Using Theorems 1 and 2 and the discussion above, one can derive perturbation inequalities for invariant subspaces. Let us first agree on a distance between subspaces. Let $\mathcal{E}$ and $\mathcal{F}$ be two subspaces and $P_\mathcal{E}$ and $P_\mathcal{F}$ be orthogonal projections onto them. One defines the *projection-2 distance* between them as

$$\mathrm{d}_2(\mathcal{E}, \mathcal{F}) := \|\!|P_\mathcal{E} - P_\mathcal{F}|\!\|_2. \tag{2.17}$$

Consider $A, \Delta \in \mathbb{S}^n$. We think of $A + \Delta$ as a perturbation of A. Let $\lambda_j := \lambda_j^\downarrow(A)$ and note that $\{\lambda_1, \ldots, \lambda_j\} \subset [\lambda_j, \lambda_1]$. Let $P_j := P_A([\lambda_j, \lambda_1])$, i.e., projection onto the eigenspace of A corresponding to the j-th largest eigenvalues. Similarly, let $\widetilde{\lambda}_j := \lambda_j^\downarrow(A + \Delta)$ and $\widetilde{P}_j := P_{A+\Delta}([\widetilde{\lambda}_j, \widetilde{\lambda}_1])$. Now, let $\delta_j := \lambda_j - \lambda_{j+1} > 0$ be the gap to right of the j-th eigenvalue of A. If $\|\!|\Delta|\!\|_2 < \frac{1}{2}\delta_j$, Weyl inequality implies that $[\widetilde{\lambda}_j, \widetilde{\lambda}_1] \subset (\lambda_j - \frac{1}{2}\delta_j, \lambda_1 + \frac{1}{2}\delta_j)$ and $[\widetilde{\lambda}_n, \widetilde{\lambda}_{j+1}] \subset (-\infty, \lambda_{j+1} + \frac{1}{2}\delta_j) = (-\infty, \lambda_j - \frac{1}{2}\delta_j)$. Letting $S_2^c := (\lambda_j - \frac{1}{2}\delta_j, \lambda_1 + \frac{1}{2}\delta_j)$, it follows that

$$P_{A+\Delta}([\widetilde{\lambda}_j, \widetilde{\lambda}_1]) = P_{A+\Delta}(S_2^c) = P^\perp_{A+\Delta}(S_2). \tag{2.18}$$

Applying Theorem 2 with $S_1 = [\lambda_j, \lambda_1]$ and $B = A + \Delta$, we obtain the following for the perturbation of the j-th leading invariant subspace.

Corollary 1. *With the above notation,*

$$\|\!|P_j - \widetilde{P}_j|\!\|_2 \leq \min\Big\{\frac{2}{\delta_j}\|\!|\Delta|\!\|_2, 1\Big\}. \tag{2.19}$$

Taking $j = 1$ in Corollary 1 gives a perturbation inequality for the largest eigenvector, assuming that the largest eigenvalue is simple; that is, assuming λ_1 is separated from the rest of the spectrum. Similar bounds maybe obtained for other eigenvectors corresponding

to simple eigenvalues. As an example, consider the previous setup, fix some $j \in \{1, \dots, n\}$ and assume λ_j is simple. Let

$$d_j := \min\left\{|\lambda_j - \lambda_{j+1}|, |\lambda_j - \lambda_{j-1}|\right\} > 0, \tag{2.20}$$

the two-sided gap in spectrum at λ_j. We assume $\lambda_{n+1} := \lambda_n$ and $\lambda_0 := \lambda_1$ for the above to make sense in the corner cases. Let $z_j := \vartheta_j^{\downarrow}(A)$, i.e., the j-th eigenvector of A, with possible sign ambiguity. Note that $z_j z_j^T = P_A([\lambda_j, \lambda_j]) = P_A(\{\lambda_j\})$. Also let $\widetilde{z}_j := \vartheta_j^{\downarrow}(A + \Delta)$. Assuming $|||\Delta|||_2 < \frac{1}{2}d_j$, we have by Weyl inequality $\{\widetilde{\lambda}_j\} \subset (\lambda_j - \frac{1}{2}d_j, \lambda_j + \frac{1}{2}d_j)$ and

$$\{\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_n\} \setminus \{\widetilde{\lambda}_j\} \subset (-\infty, \lambda_{j+1} + \tfrac{1}{2}d_j] \cup [\lambda_{j-1} - \tfrac{1}{2}d_j, \infty) \subset (\lambda_j - \tfrac{1}{2}d_j, \lambda_j + \tfrac{1}{2}d_j)^c.$$

Letting $S_2^c := (\lambda_j - \frac{1}{2}d_j, \lambda_j + \frac{1}{2}d_j)$, we observe that

$$\widetilde{z}_j \widetilde{z}_j^T = P_{A+\Delta}(\{\widetilde{\lambda}_j\}) = P_{A+\Delta}(S_2^c) = P^{\perp}_{A+\Delta}(S_2).$$

Finally, assume z_j and $\widetilde{z}_j$ are *aligned* or *sign-matched*, that is $z_j^T \widetilde{z}_j \geq 0$. Letting $c_j := z_j^T \widetilde{z}_j \in [0, 1]$ and recalling that $z_j, \widetilde{z}_j \in S^{n-1}$, we have

$$\|z_j - \widetilde{z}_j\|_2^2 = 2(1 - c_j) \leq 2(1 - c_j^2) = 2|||z_j z_j - \widetilde{z}_j \widetilde{z}_j^T|||_2^2$$

where the last equality is a consequence of CS decomposition (cf. Lemma 34 of Chapter 5). Applying Theorem 2 with $S_1 = \{\lambda_j\}$ and S_2 as above, we get the following for perturbation of aligned eigenvectors.

Corollary 2. *With the above notation,*

$$\|z_j - \widetilde{z}_j\|_2 \leq \sqrt{2}|||z_j z_j - \widetilde{z}_j \widetilde{z}_j^T|||_2 \leq \min\left\{\frac{2\sqrt{2}}{d_j}|||\Delta|||_2, 1\right\}. \tag{2.21}$$

There is a more direct way of deriving the above corollary due to Bosq which will work for any norm derived from an inner product. We refer the reader to [22, Lemma 4.3].

The takeaway from this section is that given a perturbation matrix Δ, having a good control on $|||\Delta|||_2$ leads to a good control on the eigen-structure of the perturbed matrix $A + \Delta$. This, for example, is all one needs to establish the classical consistency theory for PCA. However, in some cases in our high-dimensional models, more refined bounds are required, as we will see in later chapters.

2.3 Concentration inequalities

Concentration inequalities are ubiquitous in recent approaches to analysis of consistency of M-estimators. Also called tail bounds or large deviation bounds, they provide powerful

tools in deriving *finite sample* probability bounds. (That is, non-asymptotic bounds which hold for sufficiently large number of samples.) They are also connected with the notion of *concentration of measure* which we will briefly touch upon towards the end of this section.

Consider a random variable X with mean $\mathbb{E}\,X$. Then, $|X - \mathbb{E}\,X|$ measures the (two-sided) deviation of X from its mean. A concentration inequality is an upper bound on the probability of the deviation being larger than say t, with the upper bound going to zero exponentially fast as $t \to 0$. More precisely, an inequality of the form

$$\mathbb{P}\big(|X - \mathbb{E}\,X| > t\big) \leq c_1 \exp(-c_2 t^2), \quad \forall t \in [0, \varepsilon), \tag{2.22}$$

for some constants $c_1, c_2, \varepsilon > 0$. In words, such inequalities capture sharp "concentration" of X around its mean, in a probabilistic sense. A one-sided version of the above, i.e., a bound on $\mathbb{P}(X - \mathbb{E}\,X > t)$ for example, is sometimes called a *deviation bound*; we however will not make that distinction. The exponent of t in (2.22) can be other than 2, although, and exponent of 2 is what we encounter in most cases of interest to us.

Sub-Gaussian variables

Perhaps the most well-known approach to deriving these inequalities is the *Chernoff bounding* technique (using an upper bound on the moment generating function); the easiest case perhaps that of *sub-Gaussian* random variables. Recall that a sub-Gaussian (zero-mean) random variable X is one whose moment generating function (m.g.f. for short) is bounded uniformly by that of a Gaussian random variable, that is,

$$\mathbb{E}\exp(\lambda X) \leq \exp\big(\frac{\sigma^2\lambda^2}{2}\big), \quad \forall \lambda \in \mathbb{R}, \tag{2.23}$$

for some constant $\sigma \in [0, \infty)$ which we call "a" sub-Gaussian standard of X. (One usually calls the smallest σ satisfying (2.23) "the" sub-Gaussian standard of X; it is a norm on the space of sub-Gaussian random variables which turns it into a Banach space. We do not however insist on working with the smallest σ.) We denote a sub-Gaussian random variable X with standard σ as

$$X \sim \mathrm{SubGauss}(\sigma). \tag{2.24}$$

If the random variable is not zero-mean the above means that $X - \mathbb{E}\,X$ satisfies (2.22).

Clearly a zero-mean Gaussian random variable is sub-Gaussian. A more interesting example is a zero-mean random variable which is bounded almost surely, that is, $|X| \leq C$, a.s., for some constant $C > 0$. It is not hard to see that the sum of independent sub-Gaussian random variables with standards σ_i is sub-Gaussian with standard $(\sum_i \sigma_i^2)^{1/2}$.

It follows form (2.23) and Markov inequality [46] that for $\lambda, t \in [0, \infty)$,

$$\mathbb{P}\big(X \geq t\big) = \mathbb{P}\big(e^{\lambda X} \geq e^{\lambda t}\big) \leq e^{-\lambda t}\,\mathbb{E}\,e^{\lambda X} \leq \exp\big(\frac{\sigma^2\lambda^2}{2} - \lambda t\big). \tag{2.25}$$

The right-hand side is minimized by taking $\lambda = \frac{t}{\sigma^2}$, leading to

$$\mathbb{P}\big(X \geq t\big) \leq \exp\big(-\frac{t^2}{2\sigma^2}\big), \quad t \geq 0. \tag{2.26}$$

Since the same inequality holds for X replaced by $-X$, we get an inequality of the form (2.22) with $c_1 = 2$, $c_2 = (2\sigma^2)^{-1}$ and $\varepsilon = \infty$. We also get the following, the special case of which for bounded random variables is called the *Hoeffding inequality*.

Lemma 1. *Let $X_i \sim \mathrm{SubGauss}(\sigma_i)$, $i = 1, \ldots, n$ be independent. Then,*

$$\mathbb{P}\big(\big|\sum_i (X_i - \mathbb{E}\, X_i)\big| > t\big) \leq 2\exp\big(-\frac{t^2}{2\sum_i \sigma_i^2}\big), \quad t \in [0, \infty). \tag{2.27}$$

Sub-Exponential variables

If a zero-mean random variable X satisfies (2.23) only on a neighborhood of zero, say for $\lambda \in (-\Lambda, \Lambda)$, we call it sub-exponential with parameters σ and Λ, denoted as

$$X \sim \mathrm{SubExp}(\sigma, \Lambda). \tag{2.28}$$

If X is not zero-mean, the above means that $X - \mathbb{E}\, X$ satisfies the necessary condition. There are alternative characterizations of sub-exponentiality. For example, the above definition is equivalent, under $\mathbb{E}\, X = 0$, to either of the following two:

(a) $\mathbb{E}\, e^{\lambda X} < \infty$, $\lambda \in (-\Gamma, \Gamma)$, for some $\Gamma > 0$,

(b) $\mathbb{E}\, e^{\lambda_0 |X|} < \infty$, for some $\lambda_0 > 0$.

This first one is called Cramér condition. As an example, let $X \sim N(0,1)$; then $\mathbb{E}\, e^{\lambda X^2} = (1-2\lambda)^{-1/2}$ for $\lambda \in (-\infty, \frac{1}{2})$, and ∞ otherwise. Thus, $X^2 - 1$ satisfies the Cramér condition on $(-\frac{1}{2}, \frac{1}{2})$, hence it is sub-exponential; or according to our convention one just says that X^2 is sub-exponential. It is easy to show that sums of independent sub-exponential random variables are sub-exponential. Recall that a *chi-square* random variable with n degrees of freedom, denoted as χ_n^2, is the sum of squares of n independent standard Gaussian random variables. It follows that χ_n^2 is sub-exponential.

For a (zero-mean) sub-exponential random variable, we still have (2.25) but only for $\lambda \in (-\Lambda, \Lambda)$. If $t/\sigma^2 < \Lambda$, the right-hand side (RHS) is again minimized at $\lambda = t/\sigma^2$ and we get the sub-Gaussian type bound (2.26). If $t/\sigma^2 \geq \Lambda$, the infimum of RHS is achieved as $\lambda \uparrow \Lambda$, leading to the sub-exponential type bound

$$\mathbb{P}(X \geq t) \leq \exp\big(\frac{\sigma^2\Lambda^2}{2} - \Lambda t\big) \leq \exp\big(-\frac{\Lambda t}{2}\big). \tag{2.29}$$

The two bounds can be summarized as

$$\mathbb{P}(X \geq t) \leq \exp\big(-\min\big\{\frac{t^2}{2\sigma^2}, \frac{\Lambda t}{2}\big\}\big), \quad t \in [0, \infty). \tag{2.30}$$

Note that in a neighborhood of zero we still have a bound of the form (2.22). We also have.

Lemma 2. *Let $X_i \sim \mathrm{SubExp}(\sigma_i, \Lambda)$, $i = 1, \ldots, n$. Then,*

$$\mathbb{P}\big(\big|\sum_i (X_i - \mathbb{E}\, X_i)\big| > t\big) \leq 2\exp\big(-\min\big\{\frac{t^2}{2\sum_i \sigma_i^2}, \frac{\Lambda t}{2}\big\}\big), \quad t \in [0, \infty). \tag{2.31}$$

The special case of the above for a χ^2_n random variable is frequently used in this thesis. In particular, there is a constant $t_0 > 0$ such that

$$\mathbb{P}\big(\big|\frac{\chi^2_n}{n} - 1\big| > t\big) \leq 2\exp\big(-\frac{nt^2}{2}\big), \quad t \leq t_0.$$

More refined versions of χ^2 concentration are discussed in §3.A. There are other variations on Lemma 1 and Lemma 2. For example, using variance information, one can obtain Bernstein type inequalities. These inequalities are also closely tied with another characterization of sub-Gaussian and sub-exponential random variables, namely, one in terms of the order of growth of their moments. For example, a random variable X is sub-Gaussian, if and only if $(\mathbb{E}\,|X|^p)^{1/p} \leq K\sqrt{p}$ for some constant $K > 0$ and all $p \geq 1$. For a more details, we refer to [25, 98].

Vector-valued variables

It is possible to extend results of the form (2.22) to random vectors, random matrices and with some success even to random elements of a general Banach space. In such cases, one is usually interested in bounding the deviation $\|X - \mathbb{E}\, X\|$ for some appropriate norm $\|\cdot\|$. For example, consider a random vector $X \in \mathbb{R}^p$ with sub-Gaussian entries, $X_i \sim \mathrm{SubGauss}(\sigma_i)$. Let $\sigma_\infty := \|(\sigma_i)\|_\infty$. Then, a simple application of union bound combined with the above exponential bounds yields

$$\mathbb{P}\big(\|X - \mathbb{E}\, X\|_\infty > t\big) \leq 2p\exp\big(-\frac{t^2}{2\sigma_\infty}\big), \quad t \in [0, \infty). \tag{2.32}$$

By the equivalence of norms on finite-dimensional spaces, this inequality can be translated to deviations in other norms.

A more interesting deviation bound along these lines is the Ahlswede-Winter matrix bound [1, 98, 91] for the deviations of the form $\|S_n - \mathbb{E}\, S_n\|_2$ where $S_n = \sum_{i=1}^n X_i$ with X_i are independent random matrices with controlled entries. It is interesting in that the proof

uses the matrix analogue of Chernoff bounding. For a version of the bound see Lemma 26 in §4.A.2.

Bounds similar to (2.32) and also the Ahlswede-Winter explicitly depend on the dimension p of the vectors (or matrices) involved. Using martingale techniques, one can obtain bounds on the deviations of the form $\big|\|S_n\| - \mathbb{E}\,\|S_n\|\big|$, where S_n is the sum of independent random elements, that do not explicitly depend on the dimension of the random elements. In particular, these bounds hold for (controlled) random elements in a Banach space. We refer to [66] for details.

Gaussian concentration

Another rather powerful approach in obtaining concentration inequalities which goes beyond sums of independent random elements is using the concentration of underlying measures. We briefly mention the idea here, focusing on the Gaussian measure; the interested reader is referred to [65] for thorough discussions. Consider a metric space (Ω, d) equipped with a Borel probability measure μ. For any Borel set A, let $A^\varepsilon = \{\omega \in \Omega : \ d(\omega, A) < \varepsilon\}$ be the ε-neighborhood of A. Also, let $A^{\varepsilon c}$ denote the complement of A^ε. The measure μ has Gaussian (or normal) concentration if there are constants $C, c > 0$ such that for all A, with $\mu(A) \geq \frac{1}{2}$,

$$\mu(A^{\varepsilon c}) \leq C \exp(-c\,\varepsilon^2), \quad \varepsilon > 0. \tag{2.33}$$

In particular, let γ^p denote canonical Gaussian measure on $\mathbb{R}^p$ with density $(2\pi)^{-p/2} e^{-\|x\|_2^2/2}$ with respect to Lebesgue measure. One can show that γ^p viewed as a (Borel) probability measure on $(\mathbb{R}^p, \|\cdot\|_2)$ satisfies (2.33) with $C = 1$ and $c = \frac{1}{2}$. (This is sometimes called *dimension free* concentration as the constants do not depend on p.)

If a measure μ satisfies (2.33), the Lipschitz functions on (Ω, d) satisfy a concentration inequality of the form (2.22). For functions $f : \Omega \to \mathbb{R}$, the Lipschitz seminorm is defined as

$$\|f\|_L := \sup_{\omega \neq \omega'} \frac{|f(\omega) - f(\omega')|}{d(\omega, \omega')}.$$

A function f is Lipschitz if $\|f\|_L < \infty$. By looking at $f/\|f\|_L$ we can (and will) restrict our attention to 1-Lipschitz functions, i.e., those with $\|f\|_L = 1$.

A median of f with respect to μ is a number m_f such that both $\mu(\{f \geq m_f\})$ and $\mu(\{f \leq m_f\})$ are greater that or equal to $\frac{1}{2}$. Take $A = \{f \leq m_f\}$, so that $\mu(A) \geq \frac{1}{2}$. Then, it is easily verified that $A^\varepsilon \subset \{f < m_f + \varepsilon\}$. (To see this, pick $x \in A^\varepsilon$. Then, by definition, there exists $y \in A$ such that $d(x, y) < \varepsilon$. It follows that $f(x) < f(y) + \varepsilon \leq m_f + \varepsilon$}.) Hence, (2.33) implies

$$\mu(\{f \geq m_f + \varepsilon\}) \leq C \exp(-c\varepsilon^2), \quad \varepsilon > 0.$$

By applying the same argument to $-f$ and combining the results, we get the two sided bound

$$\mu(\{|f - m_f| \geq \varepsilon\}) \leq 2C \exp(-c\,\varepsilon^2), \quad \varepsilon > 0 \tag{2.34}$$

which shows sharp concentration of f around its median. One can show that, in fact, having concentration inequalities for Lipschitz functions is equivalent to having (2.33) for the underlying measure. Letting $\mu f := \int f d\mu$ denote the mean of f and integrating (2.34) over $\varepsilon \in (0, \infty)$ one obtains that f is integrable and $|\mu f - m_f| \leq C\sqrt{\frac{\pi}{c}}$. Hence by modifying the constants appropriately, one also has concentration of f around its mean,

$$\mu(\{|f - \mu f| \geq \varepsilon\}) \leq C' \exp(-c'\,\varepsilon^2), \quad \varepsilon > 0.$$

This, in particular holds for 1-Lipschitz functions on a Gaussian space. In this case, using different techniques, it is in fact possible to obtain concentration around the mean without the need to modify the constants. We summarize this useful result in the following lemma (cf. [66, 65]).

Lemma 3. *Let X_i, $i = 1, \ldots, p$ be i.i.d. $N(0, 1)$. Let $f : \mathbb{R}^p \to \mathbb{R}$ be 1-Lipschitz with respect to $\|\cdot\|_2$ on $\mathbb{R}^p$ and let $f = f(X_1, \ldots, X_p)$. Then, $\mathbb{E}\,|f| < \infty$ and*

$$\mathbb{P}\big(|f - \mathbb{E}\,f| > t\big) \leq 2 \exp\big(-\frac{t^2}{2}\big), \quad t > 0. \tag{2.35}$$

2.4 PCA and its SDP formulation

As mentioned before, principal component analysis (PCA) is the central theme of this thesis. In this section, we briefly introduce it and discuss a lesser known semidefinite programming (SDP) formulation of it. We then review its consistency in the classical setting and some recent inconsistency results in the high-dimensional setting. Our focus will be on the first principal component.

PCA as subspace of maxmal variance

Consider a collection of data points $x_i, i = 1, \ldots, n$ in $\mathbb{R}^p$, drawn i.i.d. from a distribution $\mathbb{P}$. We denote the expectation with respect to this distribution by $\mathbb{E}$. For simplicity, a generic point from the distribution is denoted as x. Assume that the distribution is centered, i.e., $\mathbb{E}\,x = 0$, and that $\mathbb{E}\,\|x\|_2^2 < \infty$. We usually collect $\{x_i\}$ in a matrix $X \in \mathbb{R}^{n \times p}$. Thus, x_i represents the i-th row of X. Let Σ and $\widehat{\Sigma} = \widehat{\Sigma}_n$ denote the population covariance and the sample covariance, respectively; more specifically,

$$\Sigma := \mathbb{E}\,xx^T, \quad \widehat{\Sigma} := \frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T. \tag{2.36}$$

The first *principal component* (PC) of the distribution $\mathbb{P}$ is a vector $z^* \in \mathbb{R}^p$ satisfying

$$z^* \in \arg\max_{\|z\|_2=1} \mathbb{E}(z^T x)^2, \tag{2.37}$$

that is, z^* is a direction such that the projection of the distribution along which has maximal variance. We should warn the reader that some authors call the projection $((z^*)^T x)z^*$ or even $(z^*)^T x$ the first principal component of x. To avoid confusion, we call these the first *principal projection* and the first *principal coordinate* of x, respectively. Noting that $\mathbb{E}(z^T x)^2 = \mathbb{E}(z^T x)(x^T z) = z^T(\mathbb{E}\, xx^T)z$, we obtain

$$z^* \in \arg\max_{\|z\|_2=1} z^T \Sigma z. \tag{2.38}$$

By a well-known result in linear analysis, called Rayleigh-Ritz or Courant-Fischer theorem [15, 54, 42], (2.38) is the variational characterization of maximal eigenvectors of Σ. In the notation introduced in §2.1.2, $z^* = \vartheta_{\max}(\Sigma)$. If there are multiple maximal eigenvectors, any one of them satisfies (2.38) and can be taken as the first principal component.

The second PC is obtained by removing the contribution form the first PC and applying the same procedure; that is, obtaining the first PC of $x - ((z^*)^T x)z^*$. The subsequent PCs are obtained recursively until all the variance in x is explained, i.e., the remainder is zero. In case of ambiguity, one chooses a direction orthogonal to all the previous components. Thus, PCs form an orthonormal basis for the eigen-space of Σ corresponding to nonzero eigenvalues.

PCA as best linear approximation

A slightly different viewpoint on PCA is through a linear approximation framework. Let $\mathcal{P}_1$ be the collection of all rank-one projection operators on $\mathbb{R}^p$. That is,

$$\mathcal{P} = \left\{ zz^T : \ z \in S^{p-1} \right\}.$$

Pick some $P \in \mathcal{P}_1$. For any vector $x \in \mathbb{R}^p$, we think of Px as an approximation of x. When $x \sim \mathbb{P}$ is random, $\mathbb{E}\,\|x - Px\|_2^2$ measures the approximation error, in an average sense. We can now find the "best" P in the restricted class $\mathcal{P}_1$, as

$$P^* = \arg\min_{P \in \mathcal{P}_1} \mathbb{E}\,\|x - Px\|_2^2. \tag{2.39}$$

It is not hard to see that this formulation is equivalent to (2.37); indeed, it is equivalent to maximizing $\mathbb{E}\,\langle x, Px\rangle_2$. In particular, $P^* = z^*(z^*)^T$ where z^* is the principal component of the distribution of x. The above easily generalizes, by replacing $\mathcal{P}_1$ with $\mathcal{P}_d$, the class of rank d projection operators.

SDP formulation

Let us now derive a SDP equivalent to (2.38). By a property of the trace, $\mathrm{tr}(z^T\Sigma z) = \mathrm{tr}(\Sigma z z^T)$. For a matrix $Z \in \mathbb{R}^{p\times p}$, $Z \succeq 0$ and $\mathrm{rank}(Z) = 1$ is equivalent to $Z = zz^T$ for some $z \in \mathbb{R}^p$. Imposing the additional condition $\mathrm{tr}(Z) = 1$ is equivalent to the additional constraint $\|z\|_2 = 1$. Dropping the $\mathrm{rank}(Z) = 1$, we obtain a relaxation of (2.38) as follows

$$Z^* \in \arg \max_{Z\succeq 0,\ \mathrm{tr}(Z)=1} \mathrm{tr}(\Sigma Z). \tag{2.40}$$

It turns out that this relaxation is in fact exact. That is,

Lemma 4. *There is always a rank one solution $Z^* = z^*(z^*)^T$ of (2.40) where $z^* = \vartheta_{\max}(\Sigma)$.*

Proof. It is enough to show that for all Z feasible for (2.40), one has $\mathrm{tr}(\Sigma Z) \le \lambda_{\max}(\Sigma)$. Using the EVD of $Z = \sum_i \lambda_i u_i u_i^T$, as in (2.6), this is equivalent to $\sum_i \lambda_i u_i^T \Sigma u_i \le \lambda_{\max}(\Sigma)$. But this is true, by (2.38) and $\sum_i \lambda_i = 1$. □

As the optimization problem in (2.40) is over the cone of semidefinite matrices ($Z \succeq 0$) with an objective and extra constraints which are linear in Z, the problem (2.40) is a textbook example of a SDP. The SDPs belong to the class of conic programs for which fast methods of solution are currently available. For more information about semidefinite programming, see [23, 96].

Noisy samples

In practice, of course, one does not have access to the population covariance, but instead must rely on a "noisy" version of the form

$$\widehat{\Sigma} = \Sigma + \Delta, \tag{2.41}$$

where $\Delta = \Delta_n$ denotes a random noise matrix, typically arising from having only a finite number of samples. Unless otherwise stated, we assume the estimate $\widehat{\Sigma}$ to be the usual sample covariance, as in (2.36). One then applies the procedure described above to $\widehat{\Sigma}$, instead of Σ, and obtains the *sample principal components*. This is what is usually referred to as principal component analysis. A natural question in assessing the performance of PCA is under what conditions the sample PCs (i.e., based on $\widehat{\Sigma}$) are consistent estimators of their population analogues.

2.4.1 Classical consistency theory (fixed p, large n)

In the classical theory of PCA, the model dimension p is viewed as fixed, and asymptotic statements are established as the number of the observations n tends to infinity. With this scaling the influence of the noise matrix Δ dies off, so that the sample eigenvectors and eigenvalues are $\sqrt{n}$-consistent estimators of their population analogues.

To be more specific, by the results of §2.2, the errors in the sample eigenvalues and eigenvectors (or projection operators) relative to the population versions are $\mathcal{O}(\|\Delta_n\|_2) = \mathcal{O}(\|\widehat{\Sigma}_n - \Sigma\|_2)$. By the strong law of large number [e.g., 94, 60], $\Delta_n \to 0$ as $n \to \infty$, almost surely, element-wise. Since p is fixed, $\|\Delta_n\|_\infty = \max_{i,j} |[\Delta_n]_{ij}| \to 0$, almost surely. Recall that all norms on a finite-dimensional vector space are equivalent with constants depending on the dimension [21, Chap. 4]. That is, all norms on $\mathbb{S}^p$ are equivalent up to constants depending on p which itself is constant. In particular,

$$\|\Delta_n\|_2 \leq p^2 \|\Delta_n\|_\infty \xrightarrow{\text{a.s.}} 0, \tag{2.42}$$

proving strong consistency. Furthermore, since by the multivariate central limit theorem [e.g., 94, 60], we have $\|\Delta_n\|_\infty = \mathcal{O}_p(n^{-1/2})$, it follows that $\|\Delta_n\|_2 = \mathcal{O}_p(n^{-1/2})$. This provides the rate of convergence for the sample eigenvalues and eigenvectors and proves $\sqrt{n}$-consistency[1].

2.4.2 Random Matrix Theory

Before considering the high-dimensional performance of PCA, we review some relevant parts of random matrix theory [9, 6, 57]. Classical random matrix theory goes back to the observation of Wigner that the empirical distribution of the eigenvalues of an n-by-n (symmetric) Gaussian matrix (or more precisely Gaussian ensemble) converges almost surely, as $n \to \infty$, to a semicircle law [103]. This result was later refined and extended to include a large class of random symmetric matrices, with i.i.d. entries on and above diagonal, satisfying some mild moment conditions; see [9, 6] and the reference therein.

Marchenko-Pastur law

A similar behavior is observed for the sample covariance which is of interest to us. To state the result, consider a data matrix[2] $X = (x_{ij}) \in \mathbb{R}^{n \times p}$ and the corresponding sample covariance matrix $\widehat{\Sigma}_n = \frac{1}{n} X^T X \in \mathbb{S}^p_+$, as in §2.4. One can look at the empirical distribution of the eigenvalues of $\widehat{\Sigma}_n$, which we denote in this section as $\mu_{\widehat{\Sigma}_n}$; this is a (discrete) probability measure, putting equal mass at each eigenvalue, that is

$$\mu_{\widehat{\Sigma}_n} := \frac{1}{p} \sum_{j=1}^{p} \delta_{\lambda_j(\widehat{\Sigma}_n)}$$

where δ_x denotes a unit point mass at x (or Dirac delta measure at x). Note that $\mu_{\widehat{\Sigma}_n}$ is a random measure.

[1] Here and elsewhere $\mathcal{O}_p(\cdot)$ denotes a stochastic "O" notation as is common in treatments of asymptotics in classical statistics. In short, $X_n = \mathcal{O}_p(a_n)$ means that $\{a_n^{-1} X_n\}$ is a *tight* sequence, that is, it is a bounded sequence in a probabilistic sense. See [94] for more details.

[2] To be precise, we consider a doubly infinite array of elements $\{x_{ij}\}$ and for each n and $p = p(n)$, take $X = X_n$ to be the n-by-p section of it, located at the upper-left corner, for example. We then have a well-defined model which we can study as $n \to \infty$.

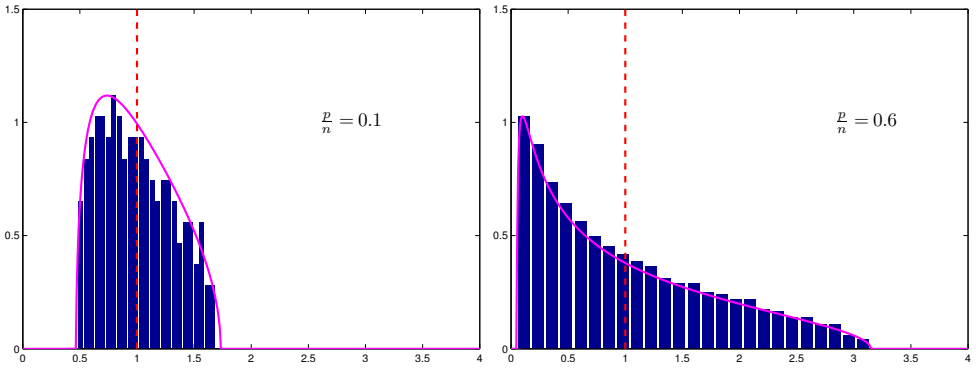


Figure 2.2: Marchenko-Pastur density, and normalized histogram of the empirical eigenvalues for $n = 2000$ and two values of $\alpha = p/n$.

Theorem 3. *Assume $\{x_{ij}\}$ to be i.i.d. mean zero, with variance σ^2. Let $(n, p) \to \infty$ such that $p/n \to \alpha \in (0, \infty)$. Then, with probability one, $\mu_{\widehat{\Sigma}_n}$ converges weakly to*

$$\mu_{MP} = \begin{cases} \nu, & \text{if } \alpha \in (0, 1] \\ (1 - \alpha^{-1})\delta_0 + \nu, & \text{if } \alpha \in (1, \infty) \end{cases} \tag{2.43}$$

where ν is absolutely continuous w.r.t. Lebesgue measure (dx) with density

$$\frac{d\nu}{dx}(x) = \frac{1}{2\pi\sigma^2} \frac{\sqrt{(b_+ - x)(x - b_-)}}{\alpha x} 1\{x \in [b_-, b_+]\}$$

where $b_\pm = \sigma^2(1 \pm \sqrt{\alpha})^2$.

The theorem was originally proved, under stronger conditions, by Marchenko and Pastur [67], in whose honor the limiting distribution is now called. The form stated above follows from a result due to Yin [106], see also [9]. Note that in the case where $\alpha > 1$, Marchenko-Pastur distribution has a point mass at 0, represented by δ_0 in (2.43).

Fig. 2.2 shows the plots of the Marchenko-Pastur density for $n = 2000$ and two values of α, namely 0.1 and 0.6. The important observation here is that while the expected value of $\widehat{\Sigma}_n$, which is equal to $\sigma^2 I_p$ under assumptions, has all its eignvalues concentrated at 1, the sample covariance itself, asymptotically, shows a spread of the eigenvalues around 1, with the smallest and largest eigenvalues tending towards the extreme points of the support of the density. Fig. 2.2 also shows the normalized histogram of the eigenvalues of a single Gaussian random matrix with $\sigma^2 = 1$. One observes that the simulation gets very close to the theoretical prediction even at samples of size $n = 2000$.

The statement above about the limiting behavior of the extreme eigenvalues does not immediately follow from Theorem 3, though separate results have established this intuitive observation. The first result in this direction was due to Gemen [43] which established convergence of largest eigenvalue of $\widehat{\Sigma}_n$ to b_+, under some growth conditions on moments of

$\{x_{ij}\}$. Subsequent results relaxed the conditions to the finiteness of only the forth moment (see [9, Chap. 5] for the intermediate results). The following is from [10] (cf. [9, Thm. 5.11]).

Theorem 4 (Bai-Yin). *Assume $\{x_{ij}\}$ to be i.i.d. with mean zero, variance σ^2 and finite fourth moment. Let $p/n \to \alpha$ as $n \to \infty$. Then, almost surely*

$$\lim_{n\to\infty} \lambda_{\max}(\widehat{\Sigma}_n) = \sigma^2(1+\sqrt{\alpha})^2, \quad \alpha \in (0,\infty), \tag{2.44}$$

$$\lim_{n\to\infty} \lambda_{\min}(\widehat{\Sigma}_n) = \sigma^2(1-\sqrt{\alpha})^2, \quad \alpha \in (0,1). \tag{2.45}$$

These theorems suggest a dramatic failure of the $\widehat{\Sigma}_n$ as an estimate of the population covaraince, in the so called "high-dimensional setting" where p (data dimension) and n (sample size) go to infinity simultanuously. It also hints at the break-down of multivariate methods, such as PCA, which base their inference on the sample covariance $\widehat{\Sigma}_n$. To be more specific, recall that $\Sigma = \mathbb{E}\,\widehat{\Sigma}_n = \sigma^2 I_p$ and note that

$$\begin{aligned}\|\widehat{\Sigma}_n - \Sigma\|_2 = &= \max_j |\lambda_j(\widehat{\Sigma}_n - \sigma^2 I_p)| \\ &= \max_j |\lambda_j(\widehat{\Sigma}_n) - \sigma^2| = \max\{|\lambda_{\min}(\widehat{\Sigma}_n) - \sigma^2|, |\lambda_{\max}(\widehat{\Sigma}_n) - \sigma^2|\}.\end{aligned}$$

Hence, by Theorem 4, for $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \|\widehat{\Sigma}_n - \Sigma\|_2 = \sigma^2(2\sqrt{\alpha} + \alpha) > 0 \tag{2.46}$$

which clearly is nonzero. That is, $\widehat{\Sigma}_n$ is not an operator-norm consistent estimate of Σ. In particular, the argument in §2.4.1 for consistency of PCA does not go through and in fact, we will see shortly that PCA is not consistent unless $p/n \to 0$ (cf. §2.4.3).

The inconsistency of $\widehat{\Sigma}$ has lead to investigation of covariance structures which lead to consistent estimators in high dimensions. Notable among these are sparse covariance estimation methods, for example, by thresholdings [18, 39] or by ℓ_1-regularization [31, 108].

Non-asymptotic results

Let us briefly look at some non-asymptotic results which closely match the asymptotic behavior discussed above. These types of results and their extensions, combined with concentration inequalities discussed in §2.3, are useful in establishing finite sample bounds on the performance of M-estimators which base their estimates on large random matrices.

We will focus on the classical case of a Gaussian matrix W with independent entries.

Theorem 5 (Gordon). *Consider $W \in \mathbb{R}^{n\times p}$ with i.i.d. standard Gaussian entries. Then,*

$$\sqrt{n} - \sqrt{p} \le \mathbb{E}\,\sigma_{\min}(W) \le \mathbb{E}\,\sigma_{\max}(W) \le \sqrt{n} + \sqrt{p}. \tag{2.47}$$

Recall that $\sigma_{\max}(W)$, for example, denotes the maximum singular value of W. The proof is based on Slepian's inequality and its generalization by Gordon to minimax of Gaussian processes [66, Section 3.3]. We refer the reader to [34] for details of the proof. However, we provide a sketch below as the technique will be useful to us in subsequent chapters.

We will only consider the proof of upper bound which is based on the Slepian's inequality, a form of "comparison inequality" for suprema of Gaussian process. We state it below for future reference. (See, [66, Section 3.3] for the proof.)

Lemma 5 (Slepian)**.** *Consider two Gaussian processes $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ whose increments satisfy $\mathbb{E}\,|X_s - X_t|^2 \le \mathbb{E}\,|Y_s - T_t|^2$ for all $s, t \in T$. Then,*

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t$$

Theorem 5 is proved by considering a Gaussian process $X_{u,v} = \langle Wu, v\rangle$ indexed by $(u, v) \in S^{p-1} \times S^{n-1}$ and comparing it with the process $Y_{u,v} = \langle g, u\rangle + \langle h, v\rangle$, where $g \in \mathbb{R}^p$ and $h \in \mathbb{R}^n$ are Gaussian vectors with i.i.d. standard entries. One then verifies that the increment inequality of Lemma 5 holds, hence the expected supremum of the Y-process dominates that of X-process. It is easy to verify using SVD of W, that supremum of X-process over its index set is $\sigma_{\max}(W)$. Using Jensen's inequality one can bound the supremum of Y-process as $\sqrt{n} + \sqrt{p}$. Modifications and extensions to this result will be presented in Chapters 3 and 5, where the argument is carried out in more details.

Combining the result of Theorem 5 with Lemma 3 on concentration of Lipschitz functions of Gaussian vectors, we can obtain full probabilistic bounds on singular values of W. In particular, treat W as a vector in the Euclidean space $\mathbb{R}^{np}$, the corresponding norm being $\|W\|_{\text{HS}} = (\sum_{i,j} W_{ij}^2)^{1/2}$. As a consequence of Weyl's Theorem (cf. (2.15)), singular values are 1-Lipschitz functions on this space, i.e., $|\sigma_i(W) - \sigma_i(W')| \le \|W - W'\|_{\text{HS}}$. Hence, Lemma 3 applies and we obtain the following.

Corollary 3. *For W of Theorem 5, we have, with probability at least $1 - 2\exp(-t^2/2)$,*

$$\sqrt{n} - \sqrt{p} - t \le \sigma_{\min}(W) \le \sigma_{\max}(W) \le \sqrt{n} + \sqrt{p} + t. \tag{2.48}$$

We note that the above result is of the right order as that predicted by asymptotics of Theorem 4. To see this, assume that $p/n = \alpha < 1$. Let $W = U[\begin{smallmatrix} S \\ 0 \end{smallmatrix}]V^T$ be the (full) SVD of W, where $S = \text{diag}(\sigma_1(W), \ldots, \sigma_p(W)) \in \mathbb{R}^{p\times p}$. We have,

$$\|\frac{1}{n}W^TW - I_p\|_2 \le \|\frac{1}{n}S^TS - I_p\|_2 = \max_i \Big|\frac{\sigma_i^2(W)}{n} - 1\Big|.$$

Taking $t = \varepsilon\sqrt{p}$ for some fixed $\varepsilon > 0$, in Corollary 3, we have, with probability at least $1 - 2\exp(-\varepsilon^2 p/2)$ that

$$\Big|\frac{\sigma_i(W)}{\sqrt{n}} - 1\Big| \le (1 + \varepsilon)\sqrt{\alpha}$$

For any $a \geq 0$, we have that $|a-1| \leq \delta$ implies[3] $|a^2-1| \leq 3\max\{\delta, \delta^2\}$. Letting $\widehat{\Sigma}_n = \frac{1}{n}W^TW$ be the sample covaraince based on W, we obtain

$$\|\widehat{\Sigma}_n - I_p\| \leq 3(1+\varepsilon)^2 \max\{\sqrt{\alpha}, \alpha\}$$

with probability at least $1 - 2\exp(-\varepsilon p/2)$ which is a non-asymptotic version of (2.46).

Part of the on-going research in non-asymptotic random matrix theory is to extend these type of results to more general ensembles, for example, matrices with just independent rows of sub-Gaussian vectors; see [98] and the references therein.

2.4.3 PCA inconsistency in high-dimensional setting

In this section, we briefly look at some inconsistency results for PCA, in the high-dimensional setting where $(n, p) \to \infty$. We will focus on the spiked covariance model proposed by [58], as this will be our model for the discussion of sparse PCA in Chapter 3. We further focus on a single-spiked model here.

Recall from §2.4 that in the context of PCA, we observe data points $\{x_1, \ldots, x_n\}$ i.i.d. from a distribution with population covaraince matrix $\Sigma = \mathbb{E}\, x_1 x_1^T$. The (single) spiked covariance model assumes the following structure on Σ,

$$\Sigma = \beta z^*(z^*)^T + I_p \tag{2.49}$$

where $\beta > 0$ is some positive constant, interpreted as a measure of signal-to-noise ratio (SNR). It is easily verified that the eigenvalues of Σ are all equal to 1 except for the largest one which is $1+\beta$. It follows that z^* is the leading PC for Σ. One then forms the sample covariance $\widehat{\Sigma}$ as in §2.4 and obtains its maximal eigenvector, $\widehat{z}$, hoping that $\widehat{z}$ is a consistent estimate[4] of z^*.

This unfortunately does not happen unless $p/n \to 0$ as shown by Paul and Johnston [75] among others (see also [57]). More specifically, under technical conditions, as $(p, n) \to \infty$ while $p/n \to \alpha > 0$, asymptotically, the following phase transition occurs:

$$\langle \widehat{z}, z^* \rangle_2 \to \begin{cases} 0, & \beta \leq \sqrt{\alpha} \\ \frac{1-\alpha/\beta^2}{1+\alpha/\beta^2}, & \beta > \sqrt{\alpha}. \end{cases} \tag{2.50}$$

Note that $\langle \widehat{z}, z^* \rangle_2$ measures cosine of the angle between $\widehat{z}$ and z^* and is related to the projection 2-distance between the corresponding 1-dimensional subspaces. (See §2.2 for definition of projection 2-distance.)

[3] To see this, note that $|a-1| \leq \delta$ implies $a+1 \leq 1+\delta+1 \leq 3\max\{\delta, 1\}$, hence $|a^2-1| = (a+1)|a-1| \leq 3\max\{\delta, 1\}\delta$.

[4] There is sign ambiguity in $\widehat{z}$; or rather we have a freedom in choosing the sign of z^*. We will assume that $\widehat{z}$ matches z^* in sign, it the sense that $\langle \widehat{z}, z^* \rangle \geq 0$.

In particular, neither case in (2.50) shows consistency, i.e., $\langle \widehat{z}, z^* \rangle_2 \to 1$. Interestingly, if the SNR β is below some threshold, $\widehat{z}$ is asymptotically orthogonal to z^*, a complete opposite of consistency.

This has led to research on additional structure/constraints that one may impose on z^* to allow for consistent estimation. As already mentioned, one such constraint is sparsity which will be the focus of our Chapter 3. Let us mention that there are other approaches to high-dimensional analysis of PCA, notable among them is the High-Dimension Low Sample Size (HDLSS) setting where one fixes the sample size n and lets the dimension p go to infinity. The HDLSS framework was introduced by Hall et. al. [50]; further analysis is provided by Ahn et. al. [2].

2.5 Hilbert spaces and reproducing kernels

We will need some facts on Hilbert spaces and functional analysis in general which we will review here. This material is mainly used from Chapter 4 onward. Details can be found in any standard text on functional analysis [21, 29, 64, 97].

We recall that a Hilbert space is a complete inner product space; it is usually used as an infinite-dimensional analogue of the Euclidean space. We occasionally represent the inner product of a Hilbert space $\mathcal{H}$ as $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the norm it induces as $\| \cdot \|_{\mathcal{H}}$; we might omit the subscript if $\mathcal{H}$ is understood from the context. Particular example of interest to us is when $\mathcal{H}$ is a space of functions $f : \mathcal{X} \to \mathbb{R}$ on some domain $\mathcal{X}$ (usually a subset of $\mathbb{R}^d$). As a more specific example, consider $\mathcal{X} \subset \mathbb{R}^d$ and let $\mathbb{P}$ be a (Borel) probability measure on $\mathcal{X}$. Then, we have the space $L^2 := L^2(\mathcal{X}, \mathbb{P})$ of (equivalence classes[5]) of square-integrable functions, defined as

$$L^2(\mathcal{X}, \mathbb{P}) := \Big\{ f : \mathcal{X} \to \mathbb{R} \;\Big|\; \int f^2 d\mathbb{P} < \infty \Big\}.$$

This space becomes a Hilbert space with the inner product $\langle f, g \rangle_{L^2} := \int f g d\mathbb{P}$, for $f, g \in L^2$.

We now summarize some key elements of the theory; $\mathcal{H}$ and $\mathcal{K}$ will be real Hilbert spaces; all Hilbert spaces are assumed over reals, unless otherwise stated.

- Linear operators and functionals: The basic objects of study are linear maps between Hilbert spaces, called (linear) operators, and linear maps from a Hilbert space to real numbers, called (linear) functionals. The collection of linear functionals on a Hilbert space is called its algebraic dual.

- Operator norm: For an operator $L : \mathcal{H} \to \mathcal{K}$, its operator norm is defined as

$$|\!|\!| L |\!|\!| := |\!|\!| L |\!|\!|_{\mathcal{H},\mathcal{K}} := \sup_{\|x\|_{\mathcal{H}} \le 1} \|Lx\|_{\mathcal{K}} = \sup_{\|x\|_{\mathcal{H}} = 1} \|Lx\|_{\mathcal{K}} = \sup_{x \neq 0} \frac{\|Lx\|_{\mathcal{K}}}{\|x\|_{\mathcal{H}}}. \tag{2.51}$$

[5] When dealing with L^p spaces, one usually identifies functions that are equal almost surely. This is so that we have $\|f\|_{L^2} = 0$ implies $f = 0$, making $\| \cdot \|_{L^2}$ a legitimate norm.

- Boundedness and continuity: When $\|\|\|L\|\|\| < \infty$, L is called a bounded operator. One then shows that for an operator, boundedness, continuity everywhere and continuity at the origin are all equivalent. We will denote the collection of bounded operators from $\mathcal{H}$ to $\mathcal{K}$ as $\mathcal{B}(\mathcal{H}, \mathcal{K})$. Taking $\mathcal{K} = \mathbb{R}$, one obtains parallel definitions and results for linear functionals on $\mathcal{H}$. The collection of all bounded linear functionals on $\mathcal{H}$ is called the (topological) dual of $\mathcal{H}$ and denoted as $\mathcal{H}^*$. Deviating slightly from our convention, we let $\|\phi\|_{\mathcal{H}*}$ denote the norm of a functional $\phi \in \mathcal{H}^*$.

- Riesz-Fréchet representation theorem: states that $\mathcal{H}^*$ is isometrically isomorphic to $\mathcal{H}$. More specifically, for any $\phi \in \mathcal{H}^*$, there is a unique $h_\phi \in \mathcal{H}$ such that $\phi(f) = \langle f, h_\phi \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, and we have $\|\phi\|_{\mathcal{H}^*} = \|h_\phi\|_{\mathcal{H}}$. The map $\phi \mapsto h_\phi$ is usually used to identify the dual space, $\mathcal{H}^*$, with the Hilbert space itself, $\mathcal{H}$.

- Orthonormal basis: A collection $\{\psi_k\} \subset \mathcal{H}$ such that $\|\psi_k\|_{\mathcal{H}} = 1$ and $\langle \psi_k, \psi_j \rangle_{\mathcal{H}} = \delta_{kj}$ for all k, j is called an orthonormal system. If furthermore, any $f \in \mathcal{H}$ can be (uniquely) represented as $f = \sum_k \langle f, \psi_k \rangle \psi_k$ (where the sum converges in $\mathcal{H}$-norm), the collection is called an orthonormal basis[6]. Any Hilbert space has an orthonormal basis which can be taken to be countable if the Hilbert space is separable; a condition which we always assume.

- Projections onto convex sets: For any closed convex set $\mathcal{C} \subset \mathcal{H}$ and any point $h \in \mathcal{H}$, there exists a unique point $\widehat{h} \in \mathcal{C}$ which is closest to h among all the points of $\mathcal{C}$, i.e., $\|h - \widehat{h}\|_2 = \inf_{c \in \mathcal{C}} \|h - c\|_2$. The map $h \mapsto \widehat{h}$ is called projection onto convex set $\mathcal{C}$, which we denote as $P_{\mathcal{C}}$.

- Projections onto linear subspaces: An important special case of the above is when $\mathcal{L} \subset \mathcal{H}$ is a closed linear subspace of $\mathcal{H}$. In this case, there is an alternative characterization of the projection map, in terms of the orthogonal complement of $\mathcal{L}$ which is defined as

$$\mathcal{L}^\perp := \big\{ f \in \mathcal{H} : \ \langle f, y \rangle_{\mathcal{H}} \text{ for all } y \in \mathcal{L} \big\}.$$

  One can verify that $\mathcal{L}^\perp$ is a closed linear subspace of $\mathcal{H}$[7].

  For any $h \in \mathcal{H}$, its projection onto $\mathcal{L}$ is the unique element $\widehat{h} \in \mathcal{L}$ such that $h - \widehat{h} \in \mathcal{L}^\perp$. This is sometimes called the "orthogonality principle ": the error $h - \widehat{h}$ is orthogonal to the subspace $\mathcal{L}$. From this it follows that $\mathcal{H}$ decomposes into direct sum of $\mathcal{L}$ and $\mathcal{L}^\perp$, represented as,

$$\mathcal{H} = \mathcal{L} \bigoplus \mathcal{L}^\perp, \tag{2.52}$$

[6]This type of basis is usually called Schauder basis, in contrast to the linear-algebraic notion of the basis, called the Hamel basis, of which any vector space has one. This later notion refers to a collection of elements of $\mathcal{H}$ such that any $f \in \mathcal{H}$ has a (unique) expansion in terms of a "finite" subcollection.

[7]This is true even if $\mathcal{L}$ itself is not closed or linear.

meaning that any $h \in \mathcal{H}$ has a unique decomposition as $h = y + z$ where $y \in \mathcal{L}$ and $z \in \mathcal{L}^\perp$. Another consequence is that $P_\mathcal{L}$ is a (bounded) linear operator of norm 1 (assuming $\mathcal{L}$ is not the trivial subspace $\{0\}$). For the above reasons, one calls $P_\mathcal{L}$ the "orthogonal" projection onto $\mathcal{L}$. We often omit the "orthogonal" label if it is implicitly understood.

- Kernel and image: Recall that for a linear operator $L : \mathcal{H} \to \mathcal{K}$, its kernel, $\operatorname{Ker} L$, and its image or range, $\operatorname{Ra} L$, are defined as

$$\begin{aligned} \operatorname{Ker} L &:= \big\{x \in \mathcal{H} : \ Lx = 0\big\}, \\ \operatorname{Ra} L &:= \big\{y \in \mathcal{K} : \ Lx = y \text{ for some } x \in \mathcal{H} \ \big\}. \end{aligned}$$

  Both are linear subspaces of their respective spaces. For a bounded linear operator, the kernel is always closed, while the image is not necessarily so. For example, for the projection operator, $P_\mathcal{L} : \mathcal{H} \to \mathcal{H}$, onto a closed linear subspace $\mathcal{L} \subset \mathcal{H}$, we have $\operatorname{Ker} P_\mathcal{L} = \mathcal{L}^\perp$ and $\operatorname{Ra} P_\mathcal{L} = \mathcal{L}$.

- Adjoints: For a linear operator $T : \mathcal{H} \to \mathcal{K}$, its adjoint $T^* : \mathcal{K} \to \mathcal{H}$ is the unique linear operator satisfying

$$\langle Tx, y\rangle_\mathcal{K} = \langle x, T^*y\rangle_\mathcal{H}\,, \quad \text{for all } x \in \mathcal{H},\ y \in \mathcal{K}.$$

  If T is bounded so is T^* and we have $\operatorname{Ker} T^* = (\operatorname{Ra} T)^\perp$. In particular, since $(\operatorname{Ra} T)^\perp = (\overline{\operatorname{Ra} T})^\perp$ we have, the direct sum decomposition[8]

$$\mathcal{K} = \overline{\operatorname{Ra} T} \bigoplus \operatorname{Ker} T^*. \tag{2.53}$$

  An operator $T \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ is self-adjoint if $T^* = T$. Self-adjoint operators are extensions of symmetric matrices. Projection operators $P_\mathcal{L}$ introduced above are examples of self-adjoint operators[9].

- Compact operators: A linear operator T is compact if it maps bounded sets to pre-compact sets (i.e., those whose closures are compact.) As a pre-compact set in a metric space is bounded, compact operators are a subclass of bounded operators. They are in fact a proper subclass as the identity map $I : \mathcal{H} \to \mathcal{H}$ on an infinite-dimensional Hilbert space is bounded but not compact[10]. Moreover, the class of compact operators is a closed[11] linear subspace of bounded operators. Finite-rank operators (i.e., operators T with finite-dimensional image, i.e., $\dim \operatorname{Ra} T < \infty$) are examples of compact operators,

[8] $\overline{\operatorname{Ra} T}$ is the closure of $\operatorname{Ra} T$ and we are relying on that (2.52) is valid for a closed subspace $\mathcal{L}$.

[9] This is seen by noting that orthogonality principle implies $\langle P_\mathcal{L} x, y\rangle = \langle P_\mathcal{L} x, P_\mathcal{L} y\rangle = \langle x, P_\mathcal{L} y\rangle$.

[10] This is a consequence of the well-known theorem that the unit ball of an infinite-dimensional normed space is never compact.

[11] Closure is in the topology induced by the operator norm defined in (2.51).

as are their limits. In fact, informally, compact operators behave in many ways like finite-rank operators (in much the same way as compact sets behave mostly like finite sets.) We will use $\mathcal{B}_0(\mathcal{H}, \mathcal{K})$ to denote compact operators from $\mathcal{H}$ to $\mathcal{K}$.

- Spectrum: of $T \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ is the collection of "complex" numbers λ such that $\lambda I - T$ does not have a bounded inverse, usually denoted as $\sigma(T)$. (Here, I is the identity map on $\mathcal{H}$.) One can show[12] that $\lambda \in \sigma(T)$ if and only if $\lambda I - T$ is not bijective. The points of the spectrum are further classified depending on how $\lambda I - T$ fails to be bijective.

  As we will be focusing on compact operators, we are mostly interested in the "point spectrum" of T, denoted as $\sigma_p(T)$, which consists of all the (complex) points λ such that $\mathrm{Ker}(\lambda I - T) \neq \{0\}$, that is, $\lambda I - T$ is not injective. In analogy with finite dimensional case, an element $\lambda \in \sigma_p(T)$ is called an eigenvalue of T, while $\mathrm{Ker}(\lambda I - T)$ is the corresponding eigenspace (i.e., each element is an eigenvector). In general, the point spectrum is a proper subset of the spectrum.

- Spectrum of a compact operator: One shows that for $T \in \mathcal{B}_0(\mathcal{H}, \mathcal{H})$,

$$\sigma(T) = \sigma_p(T) \cup \{0\}.$$

  Furthermore, $\sigma_p(T)$ is countable and can be ordered as a sequence $\{\lambda_1, \lambda_2, \dots\}$ so that $\lambda_k \to 0$ as $k \to 0$. Moreover, the eigenspace corresponding to each λ_k is finite-dimensional, i.e., $\dim \mathrm{Ker}(\lambda_k I - T) < \infty$. (One might call the aforementioned dimension the multiplicity of λ_k.)

  An important consequence of the above result is that a compact operator on $\mathcal{H}$, always has at least one eigenvector.

- Spectral theorem for compact self-adjoint operator: is an extension of EVD for symmetric matrices (cf. (2.6)). A simple statement is as follows:

  Theorem 6. *Let $\mathcal{H}$ be a (separable) Hilbert space and let $T \in \mathcal{B}_0(\mathcal{H}, \mathcal{H})$ be self-adjoint. Then, there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of T.*

  This theorem is stated in different formats in the literature. For example, let $\{\phi_k\}$ be the orthonormal basis of eigenvectors of T guaranteed above, corresponding to the sequence $\{\lambda_k\}$ of eigenvalues (i.e, $T\phi_k = \lambda_k \phi_k$. Then, any $f \in \mathcal{H}$ can be represented as $f = \sum_k \langle x, \phi_k \rangle \phi_k$ which then implies, by continuity and linearity of T,

$$Tf = \sum_k \lambda_k \langle f, \phi_k \rangle \phi_k, \quad f \in \mathcal{H}. \tag{2.54}$$

[12] This is a consequence of inverse mapping theorem, for example.

Another way of expressing the above is in terms of elementary tensors $\phi_k \otimes \phi_k$. We recall that for $h \in \mathcal{H}$, $h \otimes h \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ can be thought of as the rank-one projection onto the span of $\{h\}$. Hence[13],

$$T = \sum_k \lambda_k \, \phi_k \otimes \phi_k.$$

The spectral theorem above can be extended to compact operators between two spaces $\mathcal{H}$ and $\mathcal{K}$ which serves as an extension of SVD for general matrices (cf. (2.7)).

Reproducing Kernel Hilbert Spaces (RKHS)

We now turn to a brief review of RKHS theory. The classical reference on the subject is [8]. Here, we follow closely the treatment of [76]. See also [88, Chap. 4]. Again, our Hilbert spaces are mostly over real numbers although the results usually hold for the complex case with minor modifications.

Consider a set $\mathcal{X}$ and let $\mathcal{F}(\mathcal{X}, \mathbb{R})$ be the collection of real-valued functions on $\mathcal{X}$. In other words, $\mathcal{F}(\mathcal{X}, \mathbb{R}) := \mathbb{R}^{\mathcal{X}}$. This is clearly a vector space with the usual operations of addition and scalar multiplication.

- We say that $\mathcal{H}$ is a "reproducing kernel Hilbert space (RKHS)" on $\mathcal{X}$ over $\mathbb{R}$, if
  - $\mathcal{H}$ is a vector subspace of $\mathcal{F}(\mathcal{X}, \mathbb{R})$.
  - $\mathcal{H}$ is a Hilbert space with respect to some inner product.
  - for every $y \in \mathcal{X}$, the "evaluation functional" $\delta_y : \mathcal{H} \to \mathbb{R}$, defined as $\delta_y f := f(y)$ is bounded (equivalently continuous).
- By Riesz-Fréchet theorem, for any $y \in \mathcal{X}$, the bounded linear function δ_y can be represented by a function $k_y \in \mathcal{H}$. That is, $\delta_y(f) = \langle f, k_y \rangle$ for all $f \in \mathcal{H}$. We can then define the 2-variable "reproducing kernel" for $\mathcal{H}$ as

$$K(x, y) := k_y(x) = \langle k_y, k_x \rangle . \tag{2.55}$$

It is customary to write $K(\cdot, y)$ for k_y, so that the "reproducing property" of the kernel can be expressed as

$$\langle f, K(\cdot, y) \rangle = f(y), \quad f \in \mathcal{H}, \; y \in \mathcal{X} \tag{2.56}$$

[13] We are interpreting the series convergence in the sense of (2.54). It can also be shown that the series converges in the operator norm.

- As an example, consider the space Sobolev space

$$\mathcal{H} = \left\{ f : [0,1] \to \mathbb{R} \;\middle|\; f \text{ is absolutely continuous, } f(0) = 0 \text{ and } f' \in L^2[0,1]. \right\}$$

  with the inner product $\langle f, g \rangle = \int_0^1 f'(t)g'(t)dt$. One verifies that this is an RKHS with kernel $K(x, y) = \min(x, y)$.

- For an RKHS $\mathcal{H}$ on $\mathcal{X}$ with reproducing kernel $K(\cdot, \cdot)$, we have the following:
  - The linear span of $\{k_y : y \in \mathcal{X}\}$ is dense in $\mathcal{H}$ (in the norm topology).
  - Norm convergence in $\mathcal{H}$, implies point-wise convergence. That is, $\|f_n - f\|_{\mathcal{H}} \to 0$ implies $f_n(x) \to f(x)$ for all $x \in \mathcal{X}$. This is a consequence of (2.56) and Cauchy–Schwarz inequality.
  - If $\{\psi_k\}$ is an orthonormal basis for $\mathcal{H}$, then

$$K(x, y) := \sum_k \psi(x)\psi(y)$$

    where the series converges pointwise.

- We have the following characterization of reproducing kernels:
  - Two RKHS with equal kernel functions are equal, that is, they contain the same functions and their norms are the same. (By equality of kernels we mean pointwise equality as bivariate functions.)
  - One easily verifies that the reproducing kernel $K(\cdot, \cdot)$ of a RKHS is positive semidefinite. That is, for any finite collection $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the matrix $\big(K(x_i, x_j)\big)$ is positive semidefinite.

    The converse, which is rather deep, is also true: For any positive semidefinite function $K(\cdot, \cdot)$ on $\mathcal{X}$, there exists a RKHS on $\mathcal{X}$ whose reproducing kernel is $K(\cdot, \cdot)$. This result combined with the previous point shows that there is a one-to-one correspondence between positive semidefinite functions and RKHSs.

2.6 Minimax lower bounds via Fano inequality

One of the approachs in deriving minimax lower bounds is through the use of Fano inequality. Here, we give a brief introduction. For more details we refer to [107, 104, 48].

2.6.1 Decision theory and minimax criterion

Consider a collection of probability distributions $\{P_\theta, \theta \in \Theta\}$, parametrized with θ, defined on a common (measurable) space $\mathcal{X}$. The classical decision-theoretic approach to the problem of statistical inference is as follows: we observe a random variable X distributed according to P_θ for some unknown, underlying, true θ; the goal of us, as statisticians, is to construct an estimator of the (true) parameter, θ, based on the observation, X, under the assumption that there is a cost, depending on the mismatch between the estimator and (true) θ, which has to be minimized. We denote the estimator as $\widehat{\theta} = \widehat{\theta}(X)$.

Usually, P_θ stands for the distribution of n i.i.d. samples from a simpler distribution. That is, $P_\theta = Q_\theta^{\otimes n}$, for some Q_θ, where the notation $^{\otimes n}$ stands for n-fold product of a measure with itself. In other words, X consists of n i.i.d. copies from Q_θ. We ignore this minor detail in the following.

To formalize the notion of cost, one puts a distance (or metric) d on Θ to measure the discrepancy between θ and $\widehat{\theta}$ as $\mathrm{d}(\theta, \widehat{\theta})$. At a rather abstract level, there is usually another piece to the story, a loss function: an increasing function $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\ell(0) = 0$. The cost of a discrepancy $\mathrm{d}(\theta, \widehat{\theta})$ to the statistician is $\ell(\mathrm{d}(\theta, \widehat{\theta}))$. It is also customary to call $\ell(d(\cdot, \cdot))$ the loss function. Common loss functions are $\ell(x) = x^p$, for $p > 0$ and $\ell(x) = 1\{x > \delta\}$. The default for us is $\ell(x) = x^2$. Common distances are:

- The *discrete distance* $1\{\theta \neq \widehat{\theta}\}$, often used when θ takes on discrete values, say $\Theta \subset \{0, 1\}^d$. We call the corresponding loss $\ell(\mathrm{d}(\theta, \widehat{\theta})) = 1\{\theta \neq \widehat{\theta}\}$, the zero-one loss.
- The ℓ_2 distance $\|\theta - \widehat{\theta}\|_2$ when θ takes on real values, say $\Theta \subset \mathbb{R}^d$. We might call the corresponding loss $\ell(\mathrm{d}(\theta, \widehat{\theta})) = \|\theta - \widehat{\theta}\|_2^2$, the squared error loss.

The expected cost, also known as *risk*, that is, $R(\theta) := \mathbb{E}_\theta \, \ell(\mathrm{d}(\theta, \widehat{\theta}(X)))$ is what we are interested in minimizing. Here $\mathbb{E}_\theta$ denotes the expectation under P_θ, that is, assuming $X \sim P_\theta$. As for any particular estimator $\widehat{\theta}$, the risk depends in general on θ, one approach is to try minimizing the maximum risk,

$$R^*(\Theta) := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \, \ell(\mathrm{d}(\theta, \widehat{\theta}(X))).$$

An estimator the achieves the "inf" above is called *minimax optimal*, the corresponding risk, $R^*(\Theta)$, is called *minimax risk*. (For simplicity, we assume that the "inf" is taken over estimators $\widehat{\theta}$ taking values in Θ, so that there is no ambiguity in the definition of $R^*(\Theta)$. Also, a more detailed notation for the minimax risk is $R^*(\Theta; \mathrm{d}, \ell)$; we have omitted the dependence on the distance and loss assuming that those can be inferred from the context.)

2.6.2 Reduction by discretization and Bayesian averaging

In order to obtain lower bounds on the minimax risk, the usual approach is to reduce the problem to the case where Θ is a finite discrete set, and the loss is the zero-one loss, in which

case the inference problem is essentially a multiple hypothesis testing. More specifically, for some $\delta > 0$, let $F := F_\delta \subset \Theta$ be a finite subset of Θ with the property that

$$\mathrm{d}(\theta, \theta') \geq 2\delta, \quad \forall \theta, \theta' \in F_\delta.$$

Then it is easy to verify that

$$R^* \geq \ell(\delta)\, R^*(F_\delta) \tag{2.57}$$

where

$$R^*(F_\delta) := \inf_{\widehat{T}} \max_{\theta \in F_\delta} P_\theta\{\widehat{T} \neq \theta\}$$

where "inf" is now over all estimators $\widehat{T} = \widehat{T}(X)$ taking values in F_δ. One usually tries to choose δ appropriately such that $R^*(F_\delta)$ is bounded below by a numerical constant, in which case, by (2.57), $\ell(\delta)$ determines a lower bound on minimax risk up to constants. Of course reduction (2.57) is unnecessary, if we start with a finite discrete Θ, as for example in the problem of model (or subset) selection where $\Theta \subset \{0,1\}^d$.

The next common step is noting that the minimax risk $R^*(F_\delta)$ is bounded below by optimal Bayes risk for any prior on F_δ. In particular, let w be a probability measure on F_δ and put $w_\theta := w(\{\theta\})$, for $\theta \in F_\delta$. Then since $\sum_{\theta \in F_\delta} w_\theta = 1$ and $w_\theta \geq 0$, for $\theta \in F_\delta$, it is clear that

$$R^*(F_\delta) \geq r^*(w) \tag{2.58}$$

where

$$r^*(w) := \inf_{\widehat{T}} \sum_{\theta \in F_\delta} w_\theta P_\theta\{\widehat{T} \neq \theta\} \tag{2.59}$$

is the optimal Bayes risk for (prior) w. To see the Bayesian interpretation, let T be a random variable with distribution w and given $T = \theta$, let X be distributed as P_θ. Then, (2.59) can be restated as

$$r^*(w) = \inf_{\widehat{T}} \mathbb{P}_w(\widehat{T} \neq T) \tag{2.60}$$

where the subscript w on $\mathbb{P}_w$ signifies the underlying prior on $w(\cdot)$ the parameter θ.

2.6.3 Fano inequality

We can now apply Fano inequality to (2.60). Assume that the prior w on θ is uniform, i.e. $w_\theta = \frac{1}{|F_\delta|}$ for all $\theta \in F_\delta$. Then, one form of the inequality states

$$\inf_{\widehat{T}} \mathbb{P}(\widehat{T}(X) \neq T) \;\geq\; 1 - \frac{I(X;T) + \log 2}{\log |F_\delta|} \tag{2.61}$$

where $I(X;T)$ is the *mutual information* between X and T [30]. Hence, the problem of finding a minimax lower bound is reduced to finding an upper-bound on the mutual information $I(X;T)$. Before mentioning some such bounds, let us discuss some equivalent definitions of mutual information.

The usual definition is via *entropies*. Assume that each $P_\theta, \theta \in \Theta$ is absolutely continuous with respect to a (common) underlying measure μ, with density $x \mapsto p_\theta(x)$. Let c be the counting measure on F_δ. Then, (X,T) has a distribution with density $(x,\theta) \mapsto w_\theta p_\theta(x)$ with respect to $\mu \otimes c$. Let us denote this (joint) density as $p_{X,T}$. Let $p_X(x) := \sum_{\theta \in F_\delta} w_\theta p_\theta(x)$ be the marginal density of X with respect to μ. Also let $p_{X|T}(x|\theta) := p_\theta(x)$ denote the conditional density of X given $T = \theta$.

The (differential) entropy of X and the conditional entropy of X given T are

$$H(X) := -\int p_X \log p_X d\mu, = -\mathbb{E} \log p_X(X) \tag{2.62}$$

$$H(X|T) := -\int p_{X,T} \log p_{X|T} \; d\mu \otimes c = -\mathbb{E} \log p_{X|T}(X|T) \tag{2.63}$$

respectively. The mutual information is defined as

$$I(X;T) := H(X) - H(X|T) \tag{2.64}$$

This form is suitable for upper bounding the mutual information using the maximum entropy principal. See §3.5 for an example.

Another useful expression for the mutual information is in terms of the Kullback-Leibler (KL) divergence between two distributions. For two distributions P and Q with densities p and q with respect to μ, it is given by

$$D(P \,\|\, Q) := D(p \,\|\, q) := \int p \log \frac{p}{q} d\mu. \tag{2.65}$$

Expanding (2.64), we have

$$I(X;T) = \sum_\theta w_\theta \int p_\theta \log \frac{p_\theta}{p_X} d\mu = \sum_\theta w_\theta D(p_\theta \,\|\, p_X), \tag{2.66}$$

hence the mutual information has an interpretation as an averaged KL divergence.

Recall that we are assuming w_θ to be uniform on F_δ. For simplicity, let $N := |F_\delta|$. In this case, it is natural to denote the distribution corresponding to $p_X = \frac{1}{N} \sum_\theta p_\theta$ as $\overline{P}$; it is the center of the collection $\{P_\theta\}$. Then, we obtain an alternative form of Fano inequality,

$$\inf_{\widehat{T}} \mathbb{P}(\widehat{T} \neq T) \;\geq\; 1 - \frac{\frac{1}{N} \sum_{\theta \in F_\delta} D(P_\theta \,\|\, \overline{P}) + \log 2}{\log N} \tag{2.67}$$

2.6.4 Bounds on mutual information

A rather simple series of bounds based on representation (2.66) is

$$I(X;T) = \frac{1}{N}\sum_{\theta\in F_\delta} D(P_\theta \,\|\, \overline{P}) \overset{(a)}{\le} \frac{1}{N^2}\sum_{\theta,\theta'\in F_\delta} D(P_\theta \,\|\, P_{\theta'}) \overset{(b)}{\le} \max_{\theta,\theta'\in F_\delta} D(P_\theta \,\|\, P_{\theta'}) \tag{2.68}$$

where (a) follows from convexity of the KL divergence and (b) is trivial. Note that (b) can be interpreted as follows: the mutual information is bounded by the "diameter" of the set F_δ in KL divergence. This rather crude bound—which is still useful in some cases—can be refined to give a more sophisticated bound due to Yang and Barron [104].

To derive the Yang-Barron inequality, we first note that $I(X;T) = \inf_Q \frac{1}{N}\sum_{\theta\in F_\delta} D(P_\theta \,\|\, Q)$ where the infimum is taken over all distributions Q of X. Using this variational characterization, any choice of Q provides an upper bound. In particular, let $E \subset \Theta$ some finite subset of Θ, possibly other than F_δ. Let Q be the distribution of X obtained by setting a uniform prior on E, i.e., Q has the density $q = \frac{1}{|E|}\sum_{t\in E} p_t$. For each $\theta \in F_\delta$, let $t_\theta \in E$ be such that $D(P_\theta \,\|\, P_{t_\theta}) = \min_{t\in E} D(P_\theta \,\|\, P_t)$—that is, we are projecting P_θ onto $\{P_t,\ t \in E\}$. Then,

$$\begin{aligned} I(X;T) \le \frac{1}{N}\sum_{\theta\in F_\delta} D(P_\theta \,\|\, Q) &= \frac{1}{N}\sum_{\theta\in F_\delta}\int p_\theta \log \frac{p_\theta}{|E|^{-1}\sum_{t\in E} p_t} d\mu \\ &\overset{(a)}{\le} \frac{1}{N}\sum_{\theta\in F_\delta}\int p_\theta \log \frac{p_\theta}{|E|^{-1} p_{t_\theta}} d\mu \\ &= \log|E| + \frac{1}{N}\sum_{\theta\in F_\delta} D(P_\theta \,\|\, P_{t_\theta}) \\ &\le \log|E| + \max_{\theta\in F_\delta} D(P_\theta \,\|\, P_{t_\theta}) \end{aligned} \tag{2.69}$$

where (a) follows since dropping nonnegative terms of a sum does not make it bigger. (We are dropping every term except p_{t_θ},) Using the definition of t_θ, we can restate this as the Yang-Barron inequality

$$I(X;T) \le \log|E| + \max_{\theta\in F_\delta}\min_{t\in E} D(P_\theta \,\|\, P_t). \tag{2.70}$$

To see the usefulness of (2.70), let $N_{KL}(\varepsilon)$ be the cardinality of the smallest set $E \subset \Theta$ such that any $\theta \in \Theta$ is within a ball of radius ε around some element of E, in square root KL (pseudo-)metric, that is, for any $\theta \in \Theta$, there is $t \in E$ such that $\sqrt{D(P_\theta \,\|\, P_t)} \le \varepsilon$. In other words, $N_{KL}(\varepsilon)$ is the *covering number* of set Θ in square root KL (pseudo-)metric. It follows that the RHS of (2.70) is upper-bounded by $\log N_{KL}(\varepsilon) + \varepsilon^2$. As this holds for any $\varepsilon > 0$, we obtain

$$I(X;T) \le \inf_{\varepsilon>0}\left\{\varepsilon^2 + \log N_{KL}(\varepsilon)\right\}. \tag{2.71}$$

2.6.5 Symmetry and risk flatness

The lower bound on the minimax risk using Fano inequality is based on the uniform prior on the parameter space. The question one might ask is whether there are other priors which lead to tighter bounds. It is known that under some general regularity conditions [16], there are the so-called *least favorable* priors for which the optimal Bayes risk is equal to the minimax risk, i.e., for which inequality (2.58) is tight. In this subsection, we consider settings in which symmetry considerations lead to the rather opposite conclusion, namely that all priors are effectively the same. To introduce the setup, we need some notations.

Consider the case where $\Theta \subset \mathbb{R}^d$. Let π be a permutation of $[d] := \{1, \ldots, d\}$. The set of all such permutations is the symmetric group of $[d]$ denoted as $\mathfrak{S}_d$. For any vector $x \in \mathbb{R}^d$, let $x \circ \pi$ be the vector in $\mathbb{R}^d$, with entries $(x \circ \pi)_i = x_{\pi(i)}$. That is,

$$x \circ \pi = (x_{\pi(1)}, \ldots, x_{\pi(d)}).$$

We are interested in collections $\{P_\theta : \theta \in F_\delta\}$ that have the following property: for any $\theta \in F_\delta$ and $\pi \in \mathfrak{S}_d$,

$$X \sim P_\theta \implies X \circ \pi \sim P_{\theta \circ \pi}. \tag{2.72}$$

Let us call such collections of distributions *symmetric*. Then, one can argue that in determining the minimax risk, we only need to focus on estimators $\widetilde{T}(X)$ that have the corresponding symmetry property, i.e., for $\pi \in \mathfrak{S}_d$

$$\widetilde{T}(X \circ \pi) \sim \widetilde{T}(X) \circ \pi.$$

We will actually argue sufficiency of an even more restricted class of estimators, namely those with the property that, for $\pi \in \mathfrak{S}_d$,

$$\widetilde{T}(X \circ \pi) = \widetilde{T}(X) \circ \pi. \tag{2.73}$$

Let us call an estimator $\widetilde{T}$ satisfying the above *strongly symmetric*. In particular, we have the following which we state without proof.

Proposition 1. *Consider a symmetric collection $\{P_\theta : \theta \in F_\delta\}$ of distributions, in the sense of (2.72). Then, for any estimator $\widehat{T}(X)$, there is a strongly symmetric estimator $\widetilde{T}(X)$, in the sense of (2.73), such that*

$$\max_{\theta \in F_\delta} P_\theta(\widetilde{T} \neq \theta) \leq \max_{\theta \in F_\delta} P_\theta(\widehat{T} \neq \theta).$$

Chapter 3

High-dimensional sparse PCA

3.1 Sparse spiked covariance model

The primary focus of this chapter is the (sparse) *spiked covariance model*, in which some base covariance matrix is perturbed by the addition of a sparse eigenvector $z^* \in \mathbb{R}^p$. In particular, in the notations of §2.4, we study sequences of covariance matrices of the form

$$\Sigma_p \;=\; \beta\, z^*(z^*)^T + \begin{bmatrix} I_k & 0 \\ 0 & \Gamma_{p-k} \end{bmatrix} \;=\; \beta\, z^*(z^*)^T + \Gamma \tag{3.1}$$

where $\Gamma_{p-k} \in \mathbb{S}_+^{p-k}$ is a symmetric PSD matrix with $\lambda_{\max}(\Gamma_{p-k}) \leq 1$. The vector z^* is assumed k-sparse, that is, having exactly k nonzero entries,

$$\operatorname{card}(z^*) = k.$$

Without loss of generality, by re-ordering the indices as necessary, we assume that the non-zero entries of z^* are indexed by $\{1, \ldots, k\}$, so that equation (3.1) is the form of the covariance after any re-ordering. We also assume that the non-zero part of z^* has entries $z_i^* \in \frac{1}{\sqrt{k}}\{-1, +1\}$, so that $\|z^*\|_2 = 1$.

The spiked covariance model (3.1) was first proposed by Johnson [55], who focused on the spiked identity covariance matrix (i.e., model (3.1) with $\Gamma_{p-k} = I_{p-k}$). As mentioned earlier (cf. ?), Johnstone and Lu [58] established that, for the spiked identity model and a Gaussian ensemble, the sample maximal eigenvector, based on a sample of size n, is inconsistent as an estimator of z^* whenever $p/n \to c > 0$. These asymptotic results were refined by later work [74, 11].

In this chapter, we study a slightly more general family of spiked covariance models, in which the matrix Γ_{p-k} is required to satisfy the following conditions:

$$\textbf{A1}. \qquad \|\!|\sqrt{\Gamma_{p-k}}|\!\|_{\infty,\infty} = \mathcal{O}(1), \quad \text{and} \tag{3.2a}$$

$$\textbf{A2}. \qquad \lambda_{\max}(\Gamma_{p-k}) \leq \min\left\{1,\; \lambda_{\min}(\Gamma_{p-k}) + \frac{\beta}{8}\right\}. \tag{3.2b}$$

Here $\sqrt{\Gamma_{p-k}}$ denotes the symmetric square root. These conditions are trivially satisfied by the identity matrix I_{p-k}, but also can hold for more general non-diagonal matrices. Thus, under the model (3.1), the population covariance matrix Σ itself need not be sparse, since (at least generically) it has $k^2 + (p-k)^2 = \Theta(p^2)$ non-zero entries. Assumption (A2) on the eigenspectrum of the matrix Γ_{p-k} ensures that as long as $\beta > 0$, then the vector z^* is the unique maximal eigenvector of Σ, with associated eigenvalue $(1+\beta)$. Since the remaining eigenvalues are bounded above by 1, the parameter $\beta > 0$ represents a signal-to-noise ratio, characterizing the separation between the maximal eigenvalue and the remainder of the eigenspectrum. Assumption (A1) is related to the fact that recovering the correct signed support means that the estimate $\widehat{z}$ must satisfy $\|\widehat{z} - z^*\|_\infty \leq 1/\sqrt{k}$. As will be clarified by our analysis (see §3.4.4), controlling this ℓ_∞ norm requires bounds on terms of the form $\|\sqrt{\Gamma_{p-k}}\, u\|_\infty$, which requires control of the ℓ_∞-operator norm $|||\sqrt{\Gamma_{p-k}}|||_{\infty,\infty}$.

3.1.1 Model selection problem

In this chapter, we study the *model selection problem* for eigenvectors: i.e., we assume that the maximal eigenvector z^* is k-sparse, meaning that it has exactly k non-zero entries, and our goal is to recover this support, along with the sign of z^* on its support. We let

$$S := \text{supp}(z^*) := \{i \mid z_i^* \neq 0\}$$

denote the support set of the maximal eigenvector; recall that $\text{supp}(z^*) = \{1, \ldots, k\}$ by our assumed ordering of the indices. We also define the function $\text{supp}_\pm : \mathbb{R}^p \to \{-1, 0, +1\}^p$ by

$$[\text{supp}_\pm(u)]_i \;:=\; \begin{cases} \text{sign}(u_i) & \text{if } u_i \neq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

so that

$$S^*_\pm := \text{supp}_\pm(z^*)$$

encodes the *signed support* of the maximal eigenvector.

Given some estimate $\widehat{S}_\pm$ of the true signed support $S^*_\pm$, we assess it based on the 0–1 loss $1\{\widehat{S}_\pm \neq S^*_\pm\}$, so that the associated risk is simply the probability of incorrect decision $\mathbb{P}(\widehat{S}_\pm \neq S^*_\pm)$. Our goal is to specify conditions on the scaling of the triplet (n, p, k) such that this error probability vanishes, or conversely fails to vanish asymptotically. We consider methods that operate based on a set of n samples $x_1, \ldots, x_n$, drawn i.i.d. with Gaussian distribution $N(0, \Sigma_p)$. Under the spiked covariance model (3.1), each sample can be written as

$$x_i \;=\; \sqrt{\beta}\, v_i z^* + \sqrt{\Gamma}\, g_i, \tag{3.4}$$

where $v_i \sim N(0,1)$ is standard Gaussian, and $g_i \sim N(0, I_p)$ is a standard Gaussian p-vector, independent of v_i, so that $\sqrt{\Gamma}\, g_i \sim N(0, \Gamma)$.

3.2 Two methods of support recovery and results

The data $\{x_i\}_{i=1}^n$ defines the sample covariance matrix

$$\widehat{\Sigma} \;:=\; \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T, \tag{3.5}$$

which follows a p-variate Wishart distribution [7]. In the following, we analyze the high-dimensional scaling of two methods for recovering the (signed) support of the maximal eigenvector which operate on $\widehat{\Sigma}$. It will be assumed throughout that the size k of the support of z^* is available to the methods a priori, i.e., we do not make any attempt at estimating k.

3.2.1 Diagonal cut-off

Under the spiked covariance model (3.1), note that the diagonal elements of the population covariance satisfy $[\Sigma]_{\ell\ell} = 1 + \beta/k$ for all $\ell \in S$, and $[\Sigma]_{\ell\ell} \leq 1$ for all $\ell \notin S$. This latter bound follows since for all $\ell \notin S$, we have $[\Sigma]_{\ell\ell} \leq |||\Gamma_{p-k}|||_{2,2} \leq 1$. This observation motivates a natural approach to recovering information about the support set S, previously used as a pre-processing step by Johnstone and Lu [58].

Let $D_\ell, \ell = 1, \ldots, p$ be the diagonal elements of the sample covariance matrix—viz.

$$D_\ell \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_{i\ell})^2 \;=\; [\widehat{\Sigma}]_{\ell\ell}.$$

Form the associated order statistics

$$D_{(1)} \leq D_{(2)} \leq \cdots \leq D_{(p-1)} \leq D_{(p)},$$

and output the random subset $\widehat{S}_d$ of cardinality k specified by the indices of the largest k elements $\{D_{(p-k+1)}, \ldots, D_{(p)}\}$. That is, if π_d is the permutation of $[p]$ such that $D_{(j)} = D_{\pi_d(j)}$, then

$$\widehat{S}_d := \big\{\pi_d(p-k+1), \ldots, \pi_d(p)\big\}.$$

The chief appeal of this method is its low computational complexity: apart from the order $\mathcal{O}(np)$ of computing the diagonal elements of $\widehat{\Sigma}$, it requires only performing a sorting operation, with complexity $\mathcal{O}(p \log p)$.

Note that this method provides only an estimate of the support S, as opposed to the signed support $S_\pm^*$. One could imagine extending the method to extract sign information as well, but our main interest in studying this method is to provide a simple benchmark by which to calibrate our later results on the performance of the more complex SDP relaxation. In particular, the following result provides a precise characterization of the statistical behavior of the diagonal cut-off method:

Proposition 2 (Performance of diagonal cut-off). *For $k = \mathcal{O}(p^{1-\delta})$ for any $\delta \in (0,1)$, the probability of successful recovery using diagonal cut-off undergoes a phase transition as a function of the rescaled sample size*

$$\theta_{\mathrm{dia}}(n,p,k) \;=\; \frac{n}{k^2 \log(p-k)}. \tag{3.6}$$

More precisely, there exists a constant θ_u such that if $n > \theta_u k^2 \log(p-k)$, then

$$\mathbb{P}\big(\widehat{S}_d = S\big) \;\geq\; 1 - \exp\big(-\Theta(k^2 \log(p-k))\big) \;\to\; 1, \tag{3.7}$$

so that the method succeeds w.a.p. one, and a constant $\theta_\ell > 0$ such that if $n \leq \theta_\ell k^2 \log(p-k)$, then

$$\mathbb{P}\big(\widehat{S}_d = S\big) \;\leq\; \exp\big(-\Theta(\log(p-k))\big) \;\to\; 0, \tag{3.8}$$

so that the method fails w.a.p. one.

Remarks. The proof of Proposition 2, provided in §3.3, is based on large deviations bounds on χ^2-variates. The achievability assertion (3.7) uses known upper bounds on the tails of χ^2-variates [e.g., 19, 58]. The converse result (3.8) requires an exponentially tight lower bound on the tails of χ^2-variates, which we derive in Appendix 3.B.

To illustrate the prediction of Proposition 2, we provide some results on the diagonal cut-off method. For all experiments reported here, we generated n samples $\{x_1, \ldots, x_n\}$ in an i.i.d. manner from the spiked covariance ensemble (3.1), with $\Gamma = I$ and $\beta = 3$. Figure 3.1 illustrates the behavior predicted by Proposition 2. Each panel plots the success probability $\mathbb{P}\big(\widehat{S}_d = S\big)$ versus the rescaled sample size $\theta_{\mathrm{dia}}(n,p,n) = n/[k^2 \log(p-k)]$. Each panel shows five model dimensions ($p \in \{100, 200, 300, 600, 1200\}$), with panel (a) showing the logarithmic sparsity index $k = \mathcal{O}(\log p)$, and panel (b) showing the case $k = \mathcal{O}(\sqrt{p})$. Each point on each curve corresponds to the average of 100 independent trials. As predicted by Proposition 2, the curves all coincide, even though they correspond to very different regimes of (p,k).

3.2.2 Semidefinite-programming relaxation

We now describe the approach to sparse PCA developed by d'Aspremont et al. [32]. Recall from §? that $\mathbb{S}^p_+$ represents the cone of positive semidefinite p-by-p matrices. Given n i.i.d. observations from the model $N(0, \Sigma_p)$, let $\widehat{\Sigma}$ be the sample covariance matrix (3.5), and let $\lambda_n > 0$ be a user-defined regularization parameter. d'Aspremont et al. [32] propose estimating z^* by solving the optimization problem

$$\widehat{Z} \;:=\; \arg \max_{Z \in \mathbb{S}^p_+,\, \mathrm{tr}(Z)=1} \Big[\mathrm{tr}(\widehat{\Sigma}\, Z) - \lambda_n \sum_{i,j} |Z_{ij}| \Big] \tag{3.9}$$

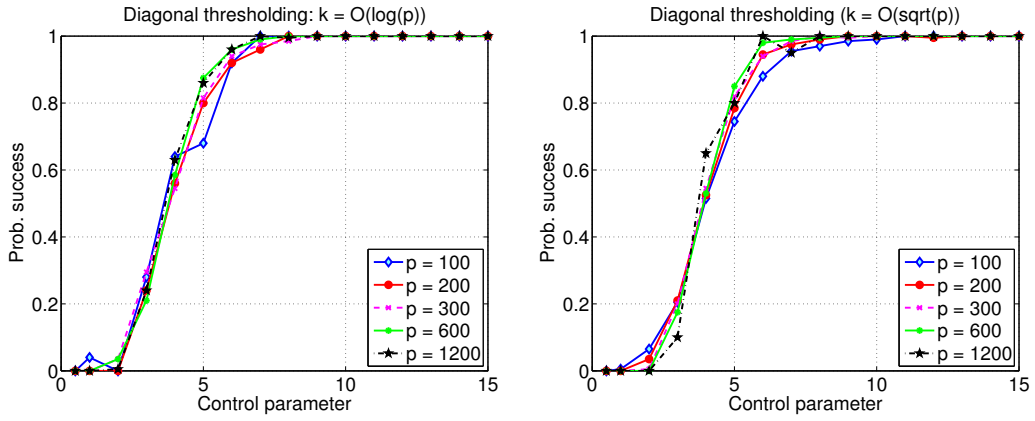


Figure 3.1: Plot of the success probability $\mathbb{P}(\widehat{S}_d = S)$ versus the rescaled sample size $\theta_{\mathrm{dia}}(n, p, k) = n/[k^2 \log(p-k)]$. The five curves in each panel correspond to model dimensions $p \in \{100, 200, 300, 600, 1200\}$, SNR parameter $\beta = 3$, and sparsity indices $k = \mathcal{O}(\log p)$ in panel (a), and $k = \mathcal{O}(\sqrt{p})$ in panel (b). As predicted by Proposition 2, the success probability undergoes a phase transition, with the curves for different model sizes and different sparsity indices all lying on top of one another.

and computing the maximal eigenvector $\widehat{z} = \vartheta_{\max}(\widehat{Z})$. The optimization problem (3.9) is a semidefinite program (SDP), a class of convex conic programs that can be solved exactly in polynomial time. Indeed, d'Asprémont et al. [32] describe an $\mathcal{O}(p^4 \log p)$ algorithm, with an implementation posted online, that we use for all simulations reported in this paper.

To gain some intuition for (3.9), recall the exact SDP formulation of the ordinary PCA, that is (2.40) of §§ 2.4. In particular, replacing Σ with $\widehat{\Sigma}$ in (2.40), we get the sample version

$$\max_{Z \in \mathbb{S}_+^p,\, \mathrm{tr}(Z)=1} \mathrm{tr}(\widehat{\Sigma} Z). \tag{3.10}$$

As mentioned before, this problem always has (at least) a rank-one solution. In particular, if $z = \vartheta_{\max}(\widehat{\Sigma})$ is any maximal eigenvector of $\widehat{\Sigma}$, then $Z = zz^T$ is a solution of (3.10). (If the maximal eigenvalue is not simple, there are also higher rank solutions.) If we were given *a priori* information that the maximal eigenvector were sparse, then it might be natural to solve the same semidefinite program with the addition of an ℓ_0 constraint. Given the intractability of such an ℓ_0-optimization problem, the SDP program (3.9) is a natural relaxation.

In particular, the following result provides sufficient conditions for the SDP relaxation (3.9) to succeed in recovering the correct signed support of the maximal eigenvector:

Theorem 7 (SDP performance guarantees)**.** *Impose conditions* (3.2a) *and* (3.2b) *on the sequence of population covariance matrices* $\{\Sigma_p\}$*, and suppose moreover that* $\lambda_n = \beta/(2k)$ *and* $k = \mathcal{O}(\log p)$*. Then,*

(a) Rank guarantee: there exists a constant $\theta_{wr} = \theta_{wr}(\Gamma, \beta)$ *such that for all sequences* (n, p, k) *satisfying* $\theta_{\mathrm{dia}}(n, p, k) > \theta_{wr}$*, the semidefinite program* (3.9) *has a rank one solution with high probability, and*

(b) *Critical scaling: there exists a constant* $\theta_{\text{crit}} = \theta_{\text{crit}}(\Gamma, \beta)$ *such that if the sequence* (n, p, k) *satisfies*

$$\theta_{\text{sdp}}(n, p, k) \;:=\; \frac{n}{k \log(p-k)} \;>\; \theta_{\text{crit}}, \tag{3.11}$$

and if there exists a rank one solution, then it specifies the correct signed support with probability converging to one.

Remarks. Part (a) of the theorem shows that rank one solutions of the SDP (3.9) are not uncommon; in particular, they are guaranteed to exist with high probability at least under the weaker scaling of the diagonal cut-off method. The main contribution of Theorem 7 is its part (b), which provides sufficient conditions for signed support recovery using the SDP, when a rank one solution exists. The bulk of our technical effort is devoted to part (b); indeed, the proof of part (a) is straightforward once all the pieces of the proof of part (b) have been introduced, and so will be deferred to the last appendix. For technical reasons, our current proof(s) require the condition $k = \mathcal{O}(\log p)$; however, it should be possible to remove this restriction, and indeed, the empirical results do not appear to require it.

3.2.3 Minimax lower bound

Proposition 2 and Theorem 7 apply to the performance of specific (polynomial-time) methods. It is natural then to ask whether there exists any algorithm, possibly with super-polynomial complexity, that has greater statistical efficiency. The following result is information-theoretic in nature, and characterizes the fundamental limitations of any algorithm regardless of its computational complexity:

Theorem 8 (Information-theoretic limitations)**.** *Consider the problem of recovering the eigenvector support in the spiked covariance model* (3.1) *with* $\Gamma = I_p$*. For any sequence* $(n, p, k) \to +\infty$ *such that*

$$\theta_{\text{sdp}}(n, p, k) \;:=\; \frac{n}{k \log(p-k)} \;<\; \frac{1+\beta}{\beta^2}, \tag{3.12}$$

the probability of error of any method is at least $1/2$*.*

Remarks. Together with Theorem 7, this result establishes the sharpness of the threshold (3.11) in characterizing the behavior of SDP relaxation, and moreover, it guarantees optimality of the SDP scaling (3.11), up to constant factors, for the spiked identity ensemble.

To illustrate the predictions of Theorem 7 and 8, we applied the SDP relaxation to the spiked identity covariance ensemble, again generating n i.i.d. samples. We solved the SDP relaxation using publically available code provided by d'Asprémont et al. [32]. Figure 3.2 shows

the corresponding plots for the SDP relaxation. Here we plot the probability $\mathbb{P}\big(\operatorname{supp}_{\pm}(\widehat{z}) = S^*_{\pm}\big)$ that the SDP relaxation correctly recovers the signed support of the unknown eigenvector z^*, where the signs are chosen uniformly in $\{-1, +1\}$ at random. Following Theorem 7, the horizontal axis plots the rescaled sample size $\theta_{\mathrm{sdp}}(n, p, k) = n/[k \log(p-k)]$. Each panel shows plots for three different problem sizes, $p \in \{100, 200, 300\}$, with panel (a) corresponding to logarithmic sparsity ($k = \mathcal{O}(\log p)$), and panel (b) to linear sparsity ($k = 0.1p$).

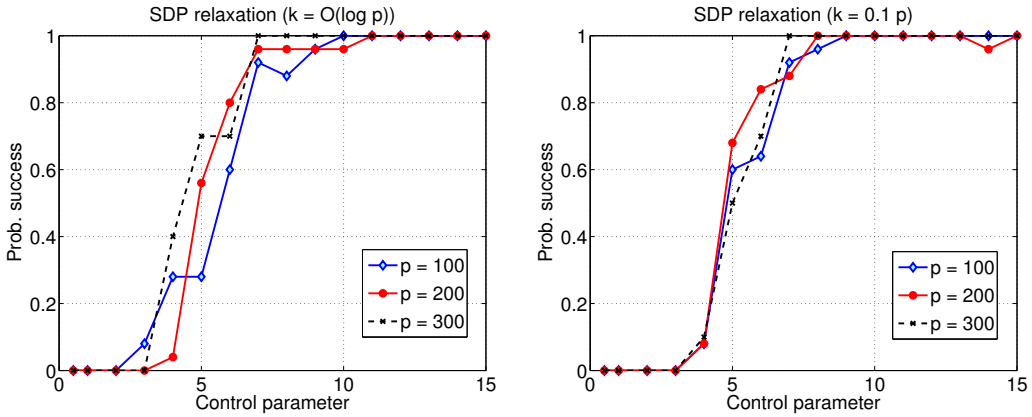


Figure 3.2: Performance of the SDP relaxation for the spiked identity ensemble, plotting the success probability $\mathbb{P}\big(\operatorname{supp}_{\pm}(\widehat{z}) = S^*_{\pm}\big)$ versus the rescaled sample size $\theta_{\mathrm{sdp}}(n, p, k) = n/[k \log(p-k)]$. The three curves in each panel correspond to model dimensions $p \in \{100, 200, 300\}$, SNR parameter $\beta = 3$, and sparsity indices $k = \mathcal{O}(\log p)$ in panel (a), and $k = 0.1p$ in panel (b). As predicted by Theorem 7, the curves in panel (a) all lie on top of one another, and transition to success once the order parameter θ_{sdp} is sufficiently large.

Consistent with the prediction of Theorem 7, the success probability rapidly approaches one once the rescaled sample size exceeds some critical threshold. (Strictly speaking, Theorem 7 only covers the case of logarithmic sparsity shown in panel (a), but the linear sparsity curves in panel (b) show the same qualitative behavior.) Note that this empirical behavior is consistent with our conclusion that the order parameter $\theta_{\mathrm{sdp}}(n, p, k) = n/[k \log(p-k)]$ is a sharp description of the SDP threshold.

3.3 Proof of Proposition 2 – diagonal cut-off

This section contains the proof of Proposition 2. We begin by proving the achievability result (3.7). We provide a detailed proof for the case $\Gamma_{p-k} = I_{p-k}$, and discuss necessary modifications for the general case at the end. For $\ell = 1, \ldots, p$, we have

$$D_\ell \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_{i\ell})^2 = \frac{1}{n}\sum_{i=1}^{n}\big(\sqrt{\beta} v_i z^*_\ell + g_{i\ell}\big)^2. \tag{3.13}$$

Since $(\sqrt{\beta} z_\ell^* v_i + g_{i\ell}) \sim N(0, \beta(z_\ell^*)^2 + 1)$ for each i, the rescaled variate $\frac{n}{\beta(z_\ell^*)^2+1} D_\ell$ is central χ^2_n with n degrees of freedom. Consequently, we have

$$\mathbb{E}[D_\ell] \;=\; \begin{cases} 1 & \text{for all } \ell \in S^c \\ 1 + \frac{\beta}{k} & \text{for all } \ell \in S, \end{cases}$$

where we have used the fact that $(z_\ell^*)^2 = 1/k$ by assumption.

A sufficient condition for success of the diagonal cut-off decoder is a threshold τ_k such that $D_\ell \geq (1 + \tau_k)$ for all $\ell \in S$, and $D_\ell < (1 + \tau_k)$ for all $\ell \in S^c$. Using the union bound and the tail bound (3.64) on central χ^2, we have

$$\mathbb{P}\Big\{ \max_{\ell \in S^c} D_\ell \geq (1 + \tau_k) \Big\} \;\leq\; (p-k)\mathbb{P}\Big\{ \frac{\chi^2_n}{n} \geq 1 + \tau_k \Big\} \;\leq\; (p-k)\, \exp\left(-\frac{3n}{16} \tau_k^2 \right),$$

so that the probability of false inclusion vanishes as long as $n > \frac{16}{3} (\tau_k)^{-2} \log(p-k)$.

On the other hand, using the union bound and the tail bound (3.63b), we have

$$\begin{aligned} \mathbb{P}\Big\{ \min_{\ell \in S} D_\ell < (1 + \tau_k) \Big\} \;&\leq\; k\, \mathbb{P}\Big\{ \frac{\chi^2_n}{n} - 1 < \frac{1 + \tau_k}{1 + \frac{\beta}{k}} - 1 \Big\} \\ &=\; k\, \mathbb{P}\Big\{ \frac{\chi^2_n}{n} - 1 < \frac{\tau_k - \frac{\beta}{k}}{1 + \frac{\beta}{k}} \Big\} \\ &\leq\; k\, \mathbb{P}\Big\{ \frac{\chi^2_n}{n} - 1 < \tau_k - \frac{\beta}{k} \Big\} \end{aligned}$$

As long as $\tau_k < \beta/k$, we may choose $x = \frac{n}{4}(\frac{\beta}{k} - \tau_k)^2$ in equation (3.63b), thereby obtaining the upper bound

$$\mathbb{P}\Big\{ \min_{\ell \in S} D_\ell < n(1 + \tau_k) \Big\} \;\leq\; k\, \exp\Big(-\frac{n}{4} \big(\frac{\beta}{k} - \tau_k\big)^2 \Big),$$

so that the probability of false exclusion vanishes as long as $n > 4(\frac{\beta}{k} - \tau_k)^{-2} \log k$. Choosing $\tau_k = \frac{\beta}{2k}$ ensures that the probability of both types of error vanish asymptotically as long as

$$n \;>\; \max\left\{ \frac{64}{3\beta^2} k^2 \log(p-k), \frac{16}{\beta^2} k^2 \log k \right\}.$$

Since $k = o(p)$, the $\log(p-k)$ term is the dominant requirement. The modifications required for the case of general Γ_{p-k} are straightforward. Since $\mathrm{var}([\sqrt{\Gamma} g_i]_\ell) = [\Gamma_{p-k}]_{\ell\ell} \leq 1$ for all $\ell \in S^c$ and samples $i = 1, \ldots, n$, we need to adjust the scaling of the χ^2_n variates. For general Γ_{p-k}, the variates $\{D_\ell, \ell \in S^c\}$ need no longer be independent, but our proof used only union bound, and so is valid regardless of the dependence structure.

We now prove the converse claim (3.8) for the spiked identity ensemble. At a high-level, this portion of the proof consists of the following steps. For a positive real t, define the events

$$\mathbb{A}_1(t) := \Big\{ \max_{\ell \in S^c} D_\ell > 1 + t \Big\}, \quad \text{and} \quad \mathbb{A}_2(t) := \Big\{ \min_{\ell \in S} D_\ell < 1 + t \Big\}.$$

Noting that the event $\mathbb{A}_1(t) \cap \mathbb{A}_2(t)$ implies failure of the diagonal cutoff decoder, it suffices to show the existence of some $t > 0$ such that $\mathbb{P}[\mathbb{A}_1(t)] \to 1$ and $\mathbb{P}[\mathbb{A}_2(t)] \to 1$.

Analysis of event $\mathbb{A}_1$. Central to the analysis of event $\mathbb{A}_1$ is the following large-deviations *lower bound* on χ^2-variates:

Lemma 6. *For a central χ^2_n variable, there exists a constant $C > 0$ such that*

$$\mathbb{P}\Big\{ \frac{\chi^2_n}{n} > 1 + t \Big\} \;\geq\; \frac{C}{\sqrt{n}} \exp\Big(- \frac{nt^2}{2} \Big), \quad t \in (0, 1).$$

See Appendix 3.B for the proof.

We exploit this lemma as follows. First define the integer-valued random variable

$$Z(t) := \sum_{\ell \in S^c} 1\{D_\ell > 1 + t\}$$

corresponding to the number of indices $\ell \in S^c$ for which the diagonal entry D_ℓ exceeds $1 + t$, and note that $\mathbb{P}[\mathbb{A}_1(t)] = \mathbb{P}[Z(t) > 0]$. By a one-sided Chebyshev inequality [46], we have

$$\mathbb{P}[\mathbb{A}_1(t)] \;=\; \mathbb{P}\Big\{ Z(t) > 0 \Big\} \;\geq\; \frac{(\mathbb{E}[Z(t)])^2}{(\mathbb{E}[Z(t)])^2 + \mathrm{var}(Z(t))}. \tag{3.14}$$

Note that $Z(t)$ is a sum of $(p - k)$ independent Bernoulli indicators, each with the same parameter $q(t) := \mathbb{P}[D_\ell > 1 + t]$. Computing the mean $\mathbb{E}[Z(t)] = (p - k)q(t)$ and variance $\mathrm{var}(Z(t)) = (p - k)q(t)\,(1 - q(t))$, and then substituting into the Chebyshev bound (3.14), we obtain

$$\mathbb{P}\Big\{ \mathbb{A}_1(t) \Big\} \;\geq\; \frac{(p-k)^2 q^2(t)}{(p-k)^2 q^2(t) + (p-k)q(t)\,(1 - q(t))} \;\geq\; \frac{(p-k)\,q(t)}{(p-k)\,q(t) + 1} \;\geq\; 1 - \frac{1}{(p-k)q(t)}.$$

Consequently, the condition $(p - k)q(t) \to \infty$ implies that $\mathbb{P}[\mathbb{A}_1(t)] \to 1$.

Let us set $t = \sqrt{\frac{\delta \log(p-k)}{n}}$ where $\delta \in (0, 1)$ is the parameter from the assumption $k = \mathcal{O}(p^{1-\delta})$. From Lemma 6, we have $q(t) \geq \frac{C}{\sqrt{n}} \exp(-nt^2/2)$, so that

$$\begin{aligned}(p-k)q\Big(\sqrt{\frac{\delta \log(p-k)}{n}}\Big) \;&\geq\; \frac{C(p-k)}{\sqrt{n}} \exp\Big(- \frac{\delta}{2} \log(p-k) \Big) \\ &=\; \frac{C(p-k)^{1-\delta/2}}{\sqrt{n}}.\end{aligned}$$

Since $n \leq Lk^2 \log(p-k)$ for some $L < +\infty$ by assumption, we have

$$(p-k)q\Big(\sqrt{\frac{\delta \log(p-k)}{n}}\Big) \;\geq\; \frac{C}{\sqrt{L}} \frac{(p-k)^{1-\delta}}{k} \frac{(p-k)^{\delta/2}}{\sqrt{\log(p-k)}},$$

which diverges to infinity, since $k = \mathcal{O}(p^{1-\delta})$.

Analysis of event $\mathbb{A}_2$. In order to analyze this event, we first need to condition on the random vector $v := (v_1, \ldots, v_n)$, so as to decouple the random variables $\{D_\ell, \ell \in S\}$. After conditioning on v, each variate $nD_\ell, \ell \in S$ is a non-central χ^2_{n,ν^*}, with n degrees of freedom and non-centrality parameter $\nu^* = \frac{\beta}{k}\|v\|_2^2$, so that each D_ℓ has mean $(\nu^* + n)$.

Since v is a standard Gaussian n-vector, we have $\|v\|_2^2 \sim \chi^2_n$. Therefore, if we define the event $\mathbb{B}(v) := \Big\{\frac{\|v\|_2^2}{n} > \frac{3}{2}\Big\}$, the large deviations bound (3.63a) implies that $\mathbb{P}[\mathbb{B}] \leq \exp(-n/16)$. Therefore, by conditioning on $\mathbb{B}$ and its complement, we obtain

$$\begin{aligned} \mathbb{P}[\mathbb{A}_2^c] &\leq \mathbb{P}\Big\{ \min_{\ell \in S} D_\ell > 1 + t \,\Big|\, \mathbb{B}^c \Big\} + \mathbb{P}[\mathbb{B}] \\ &\leq \Big(\mathbb{P}\Big\{\chi^2_{n,\nu^*} > n(1+t) \,\Big|\, \mathbb{B}^c\Big\}\Big)^k + \exp(-n/16), \end{aligned} \tag{3.15}$$

where we have used the conditional independence of $\{D_\ell, \ell \in S\}$. Finally, since $\frac{\|v\|_2^2}{n} \leq \frac{3}{2}$ on the event $\mathbb{B}^c$, we have $\nu^* \leq \frac{3\beta}{2k} n$, and thus

$$\mathbb{P}\Big\{\chi^2_{n,\nu^*} > n(1+t) \mid \mathbb{B}^c\Big\} \;\leq\; \mathbb{P}\Big\{\chi^2_{n,\nu^*} > \{n + \nu^*\} + n\{t - \frac{3\beta}{2k}\} \mid \mathbb{B}^c\Big\}.$$

Since $t = \sqrt{\delta \log(p-k)/n}$ and $n < Lk^2 \log(p-k)$, we have $t \geq \sqrt{\frac{\delta}{L}} \frac{1}{k}$, so that the quantity $\epsilon := \min\{\frac{1}{2}, t - \frac{3\beta}{2k}\}$ is positive for the pre-factor $L > 0$ chosen sufficiently small. Thus, we have

$$\begin{aligned} \mathbb{P}\Big\{\chi^2_{n,\nu^*} > n(1+t) \mid \mathbb{B}^c\Big\} &\leq \mathbb{P}\Big\{\chi^2_{n,\nu^*} > \{n + \nu^*\} + n\epsilon\Big\} \\ &\leq \exp\left(-\frac{n\,\epsilon^2}{16\,(1 + 2\frac{3}{2})}\right) \;=\; \exp\left(-\frac{n\,\epsilon^2}{64}\right) \end{aligned}$$

using the χ^2 tail bound (3.66). Substituting this upper bound into equation (3.15), we obtain

$$\mathbb{P}[\mathbb{A}_2^c] \;\leq\; \exp\left(-\frac{kn\,\epsilon^2}{64}\right) + \exp\left(-n/16\right),$$

which certainly vanishes if $\epsilon = \frac{1}{2}$. Otherwise, we have $\epsilon = t - \frac{3\beta}{2k}$ with $t = \sqrt{\frac{\delta \log(p-k)}{n}}$, and we need the quantity

$$\sqrt{kn}\left(t - \frac{3\beta}{2k}\right) \;=\; \sqrt{\delta k \log(p-k)} - \frac{3\beta}{2}\sqrt{\frac{n}{k}}$$

to diverge to $+\infty$. This divergence is guaranteed by choosing $n < Lk^2 \log(p-k)$, for L sufficiently small.

3.4 Proof of Theorem 7(b) – SDP relaxation

Theorem 7(b) is the main result of this chapter. Its proof is constructive in nature, based on the notion of a *primal-dual certificate*: that is, a primal feasible solution and a dual feasible solution that together satisfy the optimality conditions associated with the SDP (3.9).

3.4.1 High-level proof outline

We first provide a high-level outline of the main steps in our proof. Under the stated assumptions of Theorem 7, it suffices to construct a rank-one optimal solution $\widehat{Z} = \widehat{z}\widehat{z}^T$, constructed from a vector with $\|\widehat{z}\|_2 = 1$, as well as the following properties:

$$\textbf{Correct sign:} \quad \operatorname{sign}(\widehat{z}_i) = \operatorname{sign}(z_i^*) \quad \text{for all } i \in S, \quad \text{and} \tag{3.16a}$$

$$\textbf{Correct exclusion:} \quad \widehat{z}_j = 0 \quad \text{for all } j \in S^c. \tag{3.16b}$$

Note that our objective function $f(Z) = \operatorname{tr}(\widehat{\Sigma}\, Z) - \lambda_n \sum_{i,j} |Z_{ij}|$ is concave but not differentiable. However, it still possesses a subdifferential (see the books [84, 53] for more details), so that it may be shown that the following conditions are sufficient to verify the optimality of $\widehat{Z} = \widehat{z}\widehat{z}^T$. Let

$$S^{p-1} := \big\{ x \in \mathbb{R}^p : \ \|x\|_2 = 1 \big\}$$

Lemma 7. *Suppose that for each $x \in S^{p-1}$, there exists a sign matrix $\widehat{U} = \widehat{U}(x)$ such that*

(a) the matrix $\widehat{U}$ satisfies

$$\widehat{U}_{ij} = \begin{cases} \operatorname{sign}(\widehat{z}_i)\operatorname{sign}(\widehat{z}_j) & \text{if } \widehat{z}_i\widehat{z}_j \neq 0 \\ \in [-1,+1] & \text{otherwise.} \end{cases} \tag{3.17}$$

(b) The vector $\widehat{z}$ satisfies of $x^T\big(\widehat{\Sigma} - \lambda_n \widehat{U}(x)\big)x \leq \widehat{z}^T\big(\widehat{\Sigma} - \lambda_n \widehat{U}(x)\big)\widehat{z}$.

Then $\widehat{Z} = \widehat{z}\widehat{z}^T$ is an optimal rank-one solution.

Proof. The subdifferential $\partial f(\widehat{Z})$ of our objective function at $Z = \widehat{Z}$ consists of matrices of the form $\widehat{\Sigma} - \lambda_n U$, where U satisfies the condition (3.17). By the concavity of f, for any such U and for all $x \in S^{p-1}$, we have

$$f(xx^T) \leq f(\widehat{Z}) + \operatorname{tr}\big((\widehat{\Sigma} - \lambda_n U)\,(xx^T - \widehat{Z})\big).$$

Therefore, it suffices to demonstrate, for each $x \in S^{p-1}$, a valid sign matrix $\widehat{U}(x)$ such that $\mathrm{tr}((\widehat{\Sigma} - \lambda_n \widehat{U}(x))\,(xx^T - \widehat{Z})) \leq 0$. Since we have

$$\mathrm{tr}((\widehat{\Sigma} - \lambda_n \widehat{U}(x))\, xx^T) \;\; \leq \;\; \mathrm{tr}((\widehat{\Sigma} - \lambda_n \widehat{U}(x))\, \widehat{Z})$$

by assumption (b), the stated conditions are sufficient. □

Remarks. Note that if there is a $\widehat{U}$ independent of x such that $\widehat{z}$ satisfies condition (b) of Lemma 7, i.e. if $\widehat{z}$ is a maximal eigenvector of $\widehat{\Sigma} - \lambda_n \widehat{U}$, then the above argument shows that $\widehat{z}\widehat{z}^T$ is in fact "the" optimal solution (i.e., among all matrices in the constraint space, not necessarily rank-one).

The condition (3.17), when combined with the condition (3.16a), implies that we must have

$$\widehat{U}_{SS} \;\; = \;\; \mathrm{sign}(z^*_S)\,\mathrm{sign}(z^*_S)^T. \tag{3.18}$$

The remainder of the proof consists in choosing appropriately the remaining dual blocks $\widehat{U}_{SS^c}$ and $\widehat{U}_{S^cS^c}$, and verifying that the primal-dual optimality conditions are satisfied. To describe the remaining steps, it is convenient to define the matrix

$$\Phi \;\; := \;\; \widehat{\Sigma} - \lambda_n \widehat{U} - \Gamma \;\; = \;\; \beta z^* z^{*T} - \lambda_n \widehat{U} + \Delta, \tag{3.19}$$

where $\Delta := \widehat{\Sigma} - \Sigma$ is the effective noise in the sample covariance matrix. We divide our proof into three main steps, based on the block structure

$$\Phi \;\; = \;\; \begin{bmatrix} \Phi_{SS} & \Phi_{SS^c} \\ \Phi_{S^cS} & \Phi_{SS} \end{bmatrix} \;\; = \;\; \begin{bmatrix} \beta z^*_S z^{*T}_S - \lambda_n \widehat{U}_{SS} + \Delta_{SS} & -\lambda_n \widehat{U}_{SS^c} + \Delta_{SS^c} \\ -\lambda_n \widehat{U}_{S^cS} + \Delta_{S^cS} & -\lambda_n \widehat{U}_{S^cS^c} + \Delta_{S^cS^c} \end{bmatrix}. \tag{3.20}$$

(A) In Step A, we analyze the upper left block Φ_{SS}, using the fixed choice $\widehat{U}_{SS} = \mathrm{sign}(z^*_S)\,\mathrm{sign}(z^*_S)^T$. We establish conditions on the regularization parameter λ_n and the noise matrix Δ_{SS} under which the maximal eigenvector of Φ_{SS} has the same sign pattern as z^*_S. This maximal eigenvector specifies the k-dimensional subvector $\widehat{z}_S$ of our optimal primal solution.

(B) In Step B, we analyze the off-diagonal block Φ_{S^cS}, in particular establishing conditions on the noise matrix Δ_{S^cS} under which a valid sign matrix $\widehat{U}_{S^cS}$ can be chosen such that the p-vector $\widehat{z} := (\widehat{z}_S, \vec{0}_{S^c})$ is an eigenvector of the full matrix Φ.

(C) In Step C, we focus on the lower right block $\Phi_{S^cS^c}$, in particular analyzing conditions on $\Delta_{S^cS^c}$ such that a valid sign matrix $\widehat{U}_{S^cS^c}$ can be chosen such that $\widehat{z}$ defined in Step B satisfies condition (b) of Lemma 7.

Our primary interest is the effective noise matrix $\Delta = \widehat{\Sigma} - \Sigma$ induced by the usual i.i.d. sampling model. However, our results are actually somewhat more general, in that we can provide conditions on arbitrary noise matrices (which need not be of the Wishart type) under which it is possible to construct $(\widehat{z}, \widehat{U})$ as in Steps A through C. Accordingly, in order to make the proof as clear as possible, we divide our analysis into two parts: in §3.4.2, we specify sufficient properties on arbitrary noise matrices Δ, and in §3.4.3, we analyze the Wishart ensemble induced by the i.i.d. sampling model, and establish sufficient conditions on the sample size n. In §3.4.3, we focus exclusively on the special case of the spiked identity covariance, whereas §3.4.4 describes how our results extend to the more general spiked covariance ensembles covered by Theorem 7.

3.4.2 Sufficient conditions for general noise matrices

We now state a series of sufficient conditions, applicable to general noise matrices. So as to clarify the flow of the main proof, we defer the proofs of these technical lemmas to the appendix.

Sufficient conditions for step A

We begin with sufficient condition for the block (S, S). In particular, with the choice (3.18) of $\widehat{U}_{SS}$ and noting that $\operatorname{sign}(z^*_S) = \sqrt{k}\, z^*_S$ by assumption, we have

$$\Phi_{SS} = (\beta - \lambda_n k) z^*_S z^{*T}_S + \Delta_{SS} := \alpha z^*_S z^{*T}_S + \Delta_{SS},$$

where the quantity $\alpha := \beta - \lambda_n k < \beta$ represents a "post-regularization" signal-to-noise ratio. Throughout the remainder of the development, we enforce the constraint

$$\lambda_n \;=\; \frac{\beta}{2k}, \tag{3.21}$$

so that $\alpha = \beta/2$. The following lemma guarantees correct sign recovery (see equation (3.16a)), assuming that Δ_{SS} is "small" in a suitable sense:

Lemma 8. *(Correct sign recovery) Suppose that the upper left noise matrix Δ_{SS} satisfies*

$$\|\!|\Delta_{SS}|\!\|_{\infty,\infty} \leq \frac{\alpha}{10}, \qquad \text{and} \qquad \|\!|\Delta_{SS}|\!\|_{2,2} \to 0 \tag{3.22}$$

with probability 1 as $p \to +\infty$. Then w.a.p. one,

(a) The maximal eigenvalue $\gamma_1 := \lambda_{\max}(\Phi_{SS})$ converges to α, and its second largest eigenvalue γ_2 converges to zero.

(b) The upper left block Φ_{SS} has a unique maximal eigenvector $\widehat{z}_S$ with the correct sign property (i.e. $\operatorname{sign}(\widehat{z}_S) = \operatorname{sign}(z^*_S)$). More specifically, we have

$$\|\widehat{z}_S - z^*_S\|_\infty \quad \leq \quad \frac{1}{2\sqrt{k}}. \tag{3.23}$$

Sufficient conditions for step B

With the subvector $\widehat{z}_S$ specified, we can now specify the $(p-k) \times k$ submatrix $\widehat{U}_{S^cS}$ so that the vector

$$\widehat{z} \quad := \quad (\widehat{z}_S, \vec{0}_{S^c}) \in \mathbb{R}^p \tag{3.24}$$

is an eigenvector of the full matrix Φ. In particular, if we define the renormalized quantity $\widetilde{z}_S = \widehat{z}_S / \|\widehat{z}_S\|_1$, and choose

$$\widehat{U}_{S^cS} = \frac{1}{\lambda_n} \left(\Delta_{S^cS} \widetilde{z}_S\right) \operatorname{sign}(\widehat{z}_S)^T, \tag{3.25}$$

then some straightforward algebra shows that $(\Delta_{S^cS} - \lambda_n \widehat{U}_{S^cS})\widehat{z}_S = 0$, so that $\widehat{z}$ is an eigenvector of the matrix $\Phi = \beta z^*(z^*)^T - \lambda_n \widehat{U} + \Delta$. It remains to verify that the choice (3.25) is a valid sign matrix (meaning that its entries are bounded in absolute value by one).

Lemma 9. *Suppose that w.a.p. one, the matrix Δ satisfies conditions (3.22), and in addition, for sufficiently small $\delta > 0$, we have*

$$|||\Delta_{S^cS}|||_{\infty,2} \leq \frac{\delta}{\sqrt{k}}. \tag{3.26}$$

Then the specified $\widehat{U}_{S^cS}$ is a valid sign matrix w.a.p. one.

Sufficient conditions in Step C

Up to this point, we have established that $\widehat{z} := (\widehat{z}_S, \vec{0}_{S^c})$ is an eigenvector of $\widehat{\Sigma} - \lambda_n \widehat{U}$ and we have specified the sub-blocks $\widehat{U}_{SS}$ and $\widehat{U}_{SS^c}$ of the sign matrix. To complete the proof, it suffices to show that condition (b) in Lemma 7 can be satisfied—namely, that for each $x \in S^{p-1}$, there exists an extension $\widehat{U}_{S^cS^c}(x)$ to our sign matrix such that

$$\widehat{z}^T \left(\widehat{\Sigma} - \lambda_n \widehat{U}(x)\right) \widehat{z} \geq x^T \left(\widehat{\Sigma} - \lambda_n \widehat{U}(x)\right) x.$$

Note that it is sufficient to establish the above inequality with $\Phi(x)$ in place of $\widehat{\Sigma} - \lambda_n \widehat{U}(x)$[1]. Given any vector $x \in S^{p-1}$, recall the definition (3.19) of the matrix $\Phi = \Phi(x)$, and observe

[1]In particular, we have $x^T \Gamma x \leq |||\Gamma|||_{2,2} \, \|x\|_2^2 = \max\{1, |||\Gamma_{p-k}|||_{2,2}\} \, \|x\|_2^2 = 1$, while $\widehat{z}^T \Gamma \widehat{z} = \|\widehat{z}_S\|_2^2 = 1$; i.e., we have $x^T \Gamma x \leq \widehat{z}^T \Gamma \widehat{z}$.

that $(\widehat{z})^T\Phi(x)\widehat{z} = \gamma_1$ for any choice of $\widehat{U}_{S^cS^c}(x)$. Consider the partition $x = (u, v) \in S^{p-1}$, with $u \in \mathbb{R}^k$ and $v \in \mathbb{R}^m$, where $m = p - k$. We have

$$x^T\Phi x = u^T\Phi_{SS}u + 2v^T\Phi_{S^cS}u + v^T\Phi_{S^cS^c}v. \tag{3.27}$$

Let us decompose $u = \mu\widehat{z}_S + \widehat{z}_S^{\perp}$, where $|\mu| \leq 1$ and $\widehat{z}_S^{\perp}$ is an element of the orthogonal complement of the span of $\widehat{z}_S$. With this decomposition, we have

$$\begin{aligned} u^T\Phi_{SS}u &= \mu^2\widehat{z}_S^T\Phi_{SS}\widehat{z}_S + 2\mu\widehat{z}_S^T\Phi_{SS}\widehat{z}_S^{\perp} + (\widehat{z}_S^{\perp})^T\Phi_{SS}\widehat{z}_S^{\perp} \\ &= \mu^2\gamma_1 + (\widehat{z}_S^{\perp})^T\Phi_{SS}\widehat{z}_S^{\perp}, \end{aligned}$$

using the fact that $\widehat{z}_S$ is an eigenvector of Φ_{SS} with eigenvalue γ_1 by definition. Note that $\|\widehat{z}_S^{\perp}\|_2^2 \leq 1-\mu^2$, so that $(\widehat{z}_S^{\perp})^T\Phi_{SS}\widehat{z}_S^{\perp}$ is bounded by $(1-\mu^2)\gamma_2$, where γ_2 is the second largest eigenvalue of Φ_{SS}, which tends to zero according to Lemma 8. We thus conclude that

$$u^T\Phi_{SS}u \;\leq\; \mu^2\gamma_1 + (1-\mu^2)\gamma_2. \tag{3.28}$$

The following lemma addresses the remaining two terms in the decomposition (3.27):

Lemma 10. *Let $m = p-k$ and let $\mathbb{S} = \{(\eta_i, \ell_i)\}_i$ be a set of cardinality $|\mathbb{S}| = \mathcal{O}(m)$. Suppose that in addition to conditions (3.22) and (3.26), the noise matrix Δ satisfies w.p. 1*

$$\max_{\substack{\|v\|_2 \leq \eta, \\ \|v\|_1 \leq \ell}} \sqrt{v^T(\Delta_{S^cS^c} + \Gamma_m)v} \leq \eta + \frac{\delta}{\sqrt{k}}\ell + \varepsilon, \quad \forall(\eta, \ell) \in \mathbb{S}, \tag{3.29}$$

for sufficiently small $\delta, \epsilon > 0$ as $m \to +\infty$. Then w.p. 1, for all $x \in S^{p-1}$, there exists a valid sign matrix $\widehat{U}_{S^cS^c}(x)$ such that the matrix $\Phi(x) := \beta z^* z^{*T} - \lambda_n\widehat{U}(x) + \Delta$ satisfies

$$x^T\big(\Phi(x)\big)x \leq \mu^2\alpha + (1-\mu^2)\frac{\alpha}{2} \;\leq\; \alpha. \tag{3.30}$$

where $|\mu| = |x^T\widehat{z}| \leq 1$.

3.4.3 Noise in sample covariance – identity case

Having established general sufficient conditions on the effective noise matrix, we now turn to the case of i.i.d. samples $x_1, \ldots, x_n$ from the population covariance, and let the effective noise matrix correspond to the difference between the sample and population covariances. Our interest is in providing specific scalings of the triplet (n, p, k) that ensure that the constructions in Steps A through C can be carried out. So as to clarify the steps involved, we begin with the proof for the spiked identity ensemble ($\Gamma = I$). In §3.4.4, we provide the extension to non-identity spiked ensembles.

Recalling our sampling model $x_i = \sqrt{\beta}\, v_i z^* + g_i$, define the vector $h = \frac{1}{n}\sum_{i=1}^n v_i g_i$. The effective noise matrix $\Delta = \widehat{\Sigma} - \Sigma$ can be decomposed as follows:

$$\Delta \;=\; \underbrace{\beta\left(\frac{1}{n}\sum_{i=1}^n v_i^2 - 1\right) z^* z^{*T}}_{P} + \underbrace{\sqrt{\beta}\left(z^* h^T + h z^{*T}\right)}_{R} + \underbrace{\left(n^{-1}\sum_{i=1}^n g_i g_i^T - I_p\right)}_{W}. \quad (3.31)$$

We have named each of the three terms that appear in equation (3.31), so that we can deal with each one separately in our analysis. The decomposition can be summarized as

$$\Delta = \beta P + \sqrt{\beta} R + W.$$

The last term W is a *centered Wishart random matrix*, whereas the other two are cross-terms from the sampling model, involving both random vectors and the unknown eigenvector z^*. Defining the standard Gaussian random matrix $G = (g_{ij}) \in \mathbb{R}^{n\times p}$, we can express W concisely as

$$W = \frac{1}{n} G^T G - I_p. \quad (3.32)$$

Our strategy is to examine each of the terms βP, $\sqrt{\beta} R$ and W separately. For sub-block Δ_{SS}, the corresponding sub-blocks of all the three terms are present, while for sub-block Δ_{S^cS}, only $\sqrt{\beta} R_{S^cS}$ and W_{S^cS} have contributions. Since the conditions to be satisfied by these two sub-blocks are expressed in terms of their (operator) norms, the triangle inequality immediately yields the results for the whole sub-block, once we have established them separately for each of the contributing terms. On the other hand, although the conditions on $\Delta_{S^cS^c}$ (given in Lemma 10) do not have this (sub)additive property, only the Wishart term contributes to this sub-block, and it has a natural decomposition of the form required.

Regarding the Wishart term, the spectral norm of such a random matrix ($|||W|||_{2,2}$) is well-characterized [34, 43]; for instance, see claim (3.33a) in Lemma 12 for one precise statement. See also the discussion of § ?. The following lemma, concerning the mixed $(\infty, 2)$ norms of submatrices of centered Wishart matrices, is perhaps of independent interest, and plays a key role in our analysis:

Lemma 11. *Let $W \in \mathbb{R}^{p\times p}$ be a centered Wishart matrix as defined in (3.32). Let $I, J \subset \{1, \ldots, p\}$ be sets of indices, with cardinalities $|I|, |J| \to \infty$ as $n, p \to \infty$, and let $W_{I,J}$ denote the corresponding submatrix. Then as long as $\max\{|J|, \log|I|\}/n = o(1)$, we have*

$$|||W_{I,J}|||_{\infty,2} = \mathcal{O}\left(\frac{\sqrt{|J|} + \sqrt{\log|I|}}{\sqrt{n}}\right),$$

as $n, p \to +\infty$ with probability 1.

See Appendix 3.D for the proof of this claim.

Verifying steps A and B

First, let us look at the Wishart random matrix. The conditions on the upper-left sub-block W_{SS} and lower-left sub-block W_{S^cS} are addressed in the following:

Lemma 12. *As $(n, p, k) \to +\infty$, we have w.a.p. one*

$$|||W_{SS}|||_{2,2} = \mathcal{O}\left(\sqrt{\frac{k}{n}}\right), \tag{3.33a}$$

$$|||W_{SS}|||_{\infty,\infty} = \mathcal{O}\left(\sqrt{\frac{k^2}{n}}\right), \tag{3.33b}$$

$$|||W_{S^cS}|||_{\infty,2} = \mathcal{O}\left(\frac{\sqrt{k} + \sqrt{\log(p-k)}}{\sqrt{n}}\right). \tag{3.33c}$$

In particular, under the scaling $n > Lk\log(p-k)$ and $k = \mathcal{O}(\log p)$, the conditions of Lemma 8 and Lemma 9 are satisfied for W_{SS} and W_{S^cS} for sufficiently large L.

Proof. Assertion (3.33a) about the spectral norm of W_{SS} follows directly from known results on singular values of Gaussian random matrices (e.g., see [34, 43]). To bound the mixed norm $|||W_{S^cS}|||_{\infty,2}$, we apply Lemma 11 with the choices $I = S^c$ and $J = S$, noting that $|I| = p - k$ and $|J| = k$. Finally, to obtain a bound on $|||W_{SS}|||_{\infty,\infty}$, we first bound $|||W_{SS}|||_{\infty,2}$. Again using Lemma 11, this time with the choices $I = J = S$, we obtain

$$|||W_{SS}|||_{\infty,2} = \mathcal{O}\Big(\frac{\sqrt{k} + \sqrt{\log k}}{\sqrt{n}}\Big) = \mathcal{O}\Big(\sqrt{\frac{k}{n}}\Big), \tag{3.34}$$

as $n, k \to \infty$. Now, using the fact that for any $x \in \mathbb{R}^k$, $\|x\|_2 \leq \sqrt{k}\|x\|_\infty$, we obtain

$$|||W_{SS}|||_{\infty,\infty} = \max_{\|x\|_\infty \leq 1} \|W_{SS}x\|_\infty \leq \max_{\|x\|_2 \leq \sqrt{k}} \|W_{SS}x\|_\infty = \sqrt{k}|||W_{SS}|||_{\infty,2}.$$

Combined with the inequality (3.34), we obtain the stated claim (3.33b). □

We now turn to the cross-term R, and establish the following result:

Lemma 13. *The matrix $R = z^*h^T + hz^{*T}$, as defined in equation (3.31), satisfies the conditions of Lemmas 8 and 9.*

Proof. First observe that h may be viewed as a vector consisting of the off-diagonal elements of the first column of a $(p+1) \times (p+1)$ Wishart matrix, say W'. This representation follows since $h_j = \frac{1}{n}\sum_{i=1}^n v_i g_{ij}$, where the Gaussian variable v_i is independent of g_{ij} for all $1 \leq j \leq p$. For ease of reference, let us index rows/columns of W' by $1', 1, \ldots, p$, let $S' = \{1'\} \cup S$, and let $h = W'_{1',S\cup S^c}$. (Recall that $S \cup S^c$ is simply $\{1, \ldots, p\}$.)

Since the spectral norm of a matrix is an upper bound on the ℓ_2-norm of any column, we have

$$\|h_S\|_2 \leq |||W'_{S'S'}|||_{2,2} = \mathcal{O}\left(\frac{k+1}{n}\right), \tag{3.35}$$

where we used known bounds [34] on singular values of Gaussian random matrices. Under the scaling $n > Lk\log(p-k)$, we thus have $\|h_S\|_2 \xrightarrow{P} 0$. By Lemma 21, we have $\mathbb{P}[|W'_{ij}| > t] \leq C\exp(-cnt^2)$ for $t > 0$ sufficiently small, which implies (via union bound) that

$$\|h\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right) \;=\; \mathcal{O}\left(\frac{1}{\sqrt{k}}\right), \tag{3.36}$$

under our assumed scaling. Note also that $\|h\|_\infty = \max\{\|h_S\|_\infty, \|h_{S^c}\|_\infty\}$, i.e. the ∞-norm of each of these subvectors are also $\mathcal{O}(k^{-1/2})$. Assume for the following that L is chosen large enough so that $\|h\|_\infty \leq \delta/\sqrt{k}$.

Now, to complete the proof, let us first examine the spectral norm of $R_{SS} = z_S^* h_S^T + h_S z_S^{*T}$. The two (possibly) non-zero eigenvalues of this matrix are $z_S^{*T} h_S \pm \|z_S^*\|_2\|h_S\|_2$, whence we have

$$|||R_{SS}|||_{2,2} \;\leq\; \left|z_S^{*T} h_S\right| + \|z_S^*\|_2\|h_S\|_2 \;\leq\; 2\|h_S\|_2 \;\xrightarrow{P}\; 0.$$

As for the (matrix) ∞-norm of R_{SS}, let us exploit the "maximum row sum" interpretation, i.e. $|||R_{SS}|||_{\infty,\infty} = \max_{i\in S}\sum_{j\in S}|R_{ij}|$ (cf. §2.1.3) to deduce

$$\begin{aligned}
|||R_{SS}|||_{\infty,\infty} &\leq |||z_S^* h_S^T|||_{\infty,\infty} + |||h_S z_S^{*T}|||_{\infty,\infty} \\
&\leq \left(\max_{i\in S}|z_i^*|\right)\|h_S^T\|_1 + \left(\max_{i\in S}|h_i|\right)\|z_S^{*T}\|_1 \\
&\leq \frac{1}{\sqrt{k}}|||W'_{S'S'}|||_{\infty,\infty} + \|h_S\|_\infty\sqrt{k}.
\end{aligned}$$

From the argument of Lemma 12, we have $|||W'_{S'S'}|||_{\infty,\infty} = \mathcal{O}\left(\sqrt{\frac{k^2}{n}}\right)$, so that

$$\frac{1}{\sqrt{k}}|||W'_{S'S'}|||_{\infty,\infty} = \mathcal{O}\left(\sqrt{\frac{k}{n}}\right) \;\xrightarrow{P}\; 0,$$

and moreover, the norm $|||R_{SS}|||_{\infty,\infty}$ can be made smaller than 2δ, by choosing L sufficiently large in the relation $n > Lk\log(p-k)$.

Finally, to establish the additional condition required by Lemma 9—namely equation (3.26)—notice that

$$\begin{aligned}\|\!|R_{S^cS}|\!\|_{\infty,2} &= \max_{\|y\|_2=1} \|R_{S^cS}y\|_\infty \\ &= \max_{\|y\|_2=1} \|h_{S^c} z_S^{*T} y\|_\infty \\ &= \Big(\max_{\|y\|_2=1} |z_S^{*T} y|\Big)\|h_{S^c}\|_\infty \le \frac{\delta}{\sqrt{k}}\end{aligned}$$

where the last line uses $\max_{\|y\|_2=1} |z_S^{*T} y| = \|z_S^*\|_2 = 1$, thereby completing the proof. □

Finally, we examine the first term in equation (3.31), i.e. P. As this term only contributes to the upper-left block, we only need to establish that it satisfies Lemma 8:

Lemma 14. *The matrix P_{SS} satisfies condition (3.22) of Lemma 8.*

Proof. Note that for any matrix norm, we have $\|\!|P_{SS}|\!\| = \big|n^{-1}\sum_{i=1}^n (v_i)^2 - 1\big| \, \|\!|z_S^* z_S^{*T}|\!\|$. Now, notice that $\|\!|z_S^* z_S^{*T}|\!\|_{2,2} = |z_S^{*T} z_S^*| = 1$. Also, using the "maximum row sum" characterization of matrix ∞-norm, we have $\|\!|z_S^* z_S^{*T}|\!\|_{\infty,\infty} = \sum_{j=1}^k \big|\big(\pm\tfrac{1}{\sqrt{k}}\big)\big(\pm\tfrac{1}{\sqrt{k}}\big)\big| = 1$. Now by the strong law of large numbers, $\big|n^{-1}\sum_{i=1}^n (v_i)^2 - 1\big| \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. It follows that with probability 1

$$\|\!|P_{SS}|\!\|_{2,2} = \|\!|P_{SS}|\!\|_{\infty,\infty} \to 0,$$

which clearly implies condition (3.22). □

Verifying step C

For this step, we only need to consider the lower-right block of W; i.e., we only need to verify condition (3.29) of Lemma 10 for $\Delta_{S^cS^c} = W_{S^cS^c}$. Recall that $W = n^{-1}G^TG - I_p$ where G is a $n \times p$ (canonical) Gaussian matrix (see equation (3.32)). With a slight abuse of notation, let $G_{S^c} = (G_{ij})$ for $1 \le i \le n$ and $j \in S^c$. Note that $G_{S^c} \in \mathbb{R}^{n\times m}$ where $m = p - k$ and

$$\Delta_{S^cS^c} + I_m = W_{S^cS^c} + I_m = n^{-1}G_{S^c}^T G_{S^c}.$$

Now, we can simplify the quadratic form in (3.29) as

$$\sqrt{v^T(\Delta_{S^cS^c} + I_m)v} = \sqrt{\|n^{-1/2}G_{S^c}v\|_2^2} = \|n^{-1/2}G_{S^c}v\|_2.$$

for which we have the following lemma.

Lemma 15. *For any $M > 0$ and $\varepsilon > 0$, there exists a constant $B > 0$ such that for any set $\mathbb{S} = \{(\eta_i, \ell_i)\}_i$ with elements in $(0, M) \times \mathbb{R}^+$ and cardinality $|\mathbb{S}| = \mathcal{O}(m)$, we have*

$$\max_{\substack{\|v\|_2\le\eta,\\ \|v\|_1\le\ell}} \|n^{-1/2}G_{S^c}v\|_2 \le \eta + B\sqrt{\frac{\log m}{n}}\ell + \varepsilon, \quad \forall(\eta,\ell) \in \mathbb{S} \tag{3.37}$$

as $p \to \infty$, with probability 1. In particular, under the scaling $n > Lk \log m$, condition (3.29) of Lemma 10 is satisfied for L large enough.

Proof. Without loss of generality, assume $M = 1$. We begin by controlling the expectation of the left-hand side, using an argument based on the Gordon–Slepian theorem [66], similar to that used for establishing bounds on spectral norms of random Gaussian matrices [e.g. 34]. First, we require some notation: for a zero-mean random variable Z, define its standard deviation $\sigma(Z) = \big(\mathbb{E}\,|Z|^2\big)^{1/2}$. For vectors x, y of the same dimension, define the Euclidean inner product $\langle x, y\rangle = x^T y$. For matrices X, Y of the same dimension (although not necessarily symmetric), recall the Hilbert–Schmidt norm

$$\|\!|X|\!\|_{\mathrm{HS}} := \langle\!\langle X, X\rangle\!\rangle^{1/2} = \big(\sum_{i,j} X_{ij}^2\big)^{1/2}.$$

Given some (possibly uncountable) index set $\{t \in T\}$, let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be a pair of centered Gaussian processes. One version of the Gordon–Slepian theorem (see [66]) asserts that if $\sigma(X_s - X_t) \le \sigma(Y_s - Y_t)$ for all $s, t \in T$, then we have

$$\mathbb{E}[\sup_{t\in T} X_t] \;\le\; \mathbb{E}[\sup_{t\in T} Y_t]. \tag{3.38}$$

For simplicity in notation, define $\widetilde{H} := G_{S^c} \in \mathbb{R}^{n\times m}$, $H := n^{-1/2} G_{S^c}$, and fix some $\eta,\, \ell > 0$. We wish to bound

$$f(\widetilde{H}; \eta, \ell) := \max_{\substack{\|v\|_2 \le \eta,\\ \|v\|_1 \le \ell}} \|\widetilde{H} v\|_2 = \max_{\substack{\|v\|_2 \le \eta,\\ \|v\|_1 \le \ell,\\ \|u\|_2 = 1}} \langle \widetilde{H} v, u\rangle$$

where $v \in \mathbb{R}^m$, $u \in \mathbb{R}^n$. Note that $\langle \widetilde{H} v, u\rangle = u^T \widetilde{H} v = \mathrm{tr}(\widetilde{H} v u^T) = \langle\!\langle \widetilde{H}, uv^T\rangle\!\rangle$. Consider $\widetilde{H}$ to be a (canonical) Gaussian vector in $\mathbb{R}^{mn}$, take

$$T := \{t = (u, v) \in \mathbb{R}^n \times \mathbb{R}^m \;\mid\; \|v\|_2 \le \eta,\, \|v\|_1 \le \ell,\, \|u\|_2 = 1\}, \tag{3.39}$$

and define $X_t = \langle\!\langle \widetilde{H}, uv^T\rangle\!\rangle$ for $t \in T$. Observe that $(X_t)_{t\in T}$ is a (centered) canonical Gaussian process generated by $\widetilde{H}$, and $f(\widetilde{H}; \eta, \ell) = \max_{t\in T} X_t$. We compare this to the maximum of another Gaussian process $(Y_t)_{t\in T}$, defined as $Y_t = \langle (g, h), (u, v)\rangle$ where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^m$ are Gaussian vectors with $\mathbb{E}[gg^T] = \eta^2 I_n$ and $\mathbb{E}[hh^T] = I_m$. Note that, for example,

$$\sigma(\langle g, u\rangle) = \big(\mathbb{E}\langle g, u\rangle^2\big)^{1/2} = \big(u^T\, \mathbb{E}[gg^T] u\big)^{1/2} = \eta \|u\|_2,$$

in which the left-hand size is the norm of a process $\big(\langle g, u\rangle\big)_u$ expressed in terms of the norm of a vector (i.e., its index).

Let $t = (u, v) \in T$ and $t' = (u', v') \in T$. Assume, without loss of generality, that $\|v'\|_2 \le \|v\|_2$. Then, we have

$$\begin{aligned}
\sigma^2(X_t - X_{t'}) &= \|\!| uv^T - u'v'^T |\!\|_{\mathrm{HS}}^2 \\
&= \|\!| uv^T - u'v^T + u'v^T - u'v'^T |\!\|_{\mathrm{HS}}^2 \\
&= \|v\|_2^2 \|u - u'\|_2^2 + \|u'\|_2^2 \|v - v'\|_2^2 + 2\big(u^T u' - \|u'\|_2^2\big)\big(\|v\|_2^2 - v^T v'\big) \\
&\le \eta^2 \|u - u'\|_2^2 + \|v - v'\|_2^2 = \sigma^2(Y_t - Y_{t'}).
\end{aligned}$$

where we have used Cauchy–Schwarz inequality to deduce $|u^T u'| \le 1 = \|u'\|_2^2$ and $|v^T v'| \le \|v\|_2 \|v'\|_2 \le \|v\|_2^2$. Thus, the Gordon–Slepian lemma is applicable and we obtain

$$\begin{aligned}
\mathbb{E}\, f(\widetilde{H}; \eta, \ell) &\le \mathbb{E} \max_{t \in T} Y_t \\
&= \mathbb{E} \max_{\|u\|_2 = 1} \langle g, u \rangle + \mathbb{E} \max_{\|v\|_2 \le \eta,\, \|v\|_1 \le \ell} \langle h, v \rangle \\
&\le \mathbb{E}\, \|g\|_2 + \big(\mathbb{E}\, \|h\|_\infty\big)\, \ell \\
&< \sqrt{n}\, \eta + \Big(\sqrt{3 \log m}\Big)\, \ell.
\end{aligned}$$

where we have used $\big(\mathbb{E}\, \|g\|_2\big)^2 < \mathbb{E}\, \big(\|g\|_2^2\big) = \mathbb{E}\, \mathrm{tr}(gg^T) = \mathrm{tr}\, \mathbb{E}(gg^T) = n\, \eta^2$; the bound used for $\mathbb{E}\, \|h\|_\infty$ follows from standard Gaussian tail bounds [66]. Noting that $H = n^{-1/2} \widetilde{H}$, we obtain $\mathbb{E}\, f(H; \eta, \ell) \le \eta + \sqrt{\frac{3 \log m}{n}}\, \ell$.

The final step is to argue that $f(H; \eta, \ell)$ is sufficiently close to its mean. For this, we will use concentration of Gaussian measure [66, 65] for Lipschitz functions in $\mathbb{R}^{mn}$. To see that $A \to f(A; \eta, \ell)$ is in fact 1-Lipschitz, note that it satisfies the triangle inequality and it is bounded above by the spectral norm. Thus,

$$\big| f(\widetilde{H}; \eta, \ell) - f(\widetilde{F}; \eta, \ell) \big| \le f(\widetilde{H} - \widetilde{F}; \eta, \ell) \le \|\!| \widetilde{H} - \widetilde{F} |\!\|_{2,2} \le \|\!| \widetilde{H} - \widetilde{F} |\!\|_{\mathrm{HS}}$$

where we have used the assumption $\eta \le 1$. Noting that $H = n^{-1/2} \widetilde{H}$ and $f(H; \eta, \ell) = n^{-1/2} f(\widetilde{H}; \eta, \ell)$, Gaussian concentration of measure for 1-Lipschitz functions (cf. Lemma 3 of §2.3) implies that

$$\mathbb{P}\big[f(H; \eta, \ell) - \mathbb{E}[f(H; \eta, \ell)] > t \big] \le \exp(-nt^2/2).$$

Finally, we use union bound to establish the result uniformly over $\mathbb{S}$. By assumption, there exists some $K > 0$ such that $|\mathbb{S}| \le Km$. Thus

$$\mathbb{P}\Big[\max_{(\eta, \ell) \in \mathbb{S}} \big(f(H; \eta, \ell) - (\eta + \sqrt{(3 \log m)/n} \cdot \ell) \big) > t \Big] \le K \exp(-nt^2/2 + \log m).$$

Now, fix some $\varepsilon > 0$, take $t = \sqrt{\frac{6 \log m}{n}}$ and apply the Borell–Cantelli lemma to conclude that

$$\max_{(\eta,\ell)\in\mathbb{S}} \left[f(H;\eta,\ell) - \left(\eta + \sqrt{\frac{3\log m}{n}} \cdot \ell \right) \right] \;\leq\; \sqrt{\frac{6\log m}{n}} \;\leq\; \varepsilon$$

eventually (w.p. 1). □

3.4.4 Nonidentity noise covariance

In this section, we specify how the proof is extended to (population) covariance matrices having a more general base covariance term Γ_{p-k} in equation (3.1). For convenience, in this section, we write $v^i := v_i$ and $g^i := g_i$. Then, for example, $g^i_S := (g^i_j, j \in S) := (g_{ij}, j \in S)$. Let $\Gamma^{1/2}_{p-k}$ denote the (symmetric) square root of Γ_{p-k}. We can write samples from the general model as

$$\tilde{x}^i = \sqrt{\beta} v_i z^* + \tilde{g}^i, \quad i = 1, \ldots, n \tag{3.40}$$

where

$$\tilde{g}^i = \begin{pmatrix} g^i_S \\ \Gamma^{1/2}_{p-k} g^i_{S^c} \end{pmatrix} \tag{3.41}$$

with $g^i \sim N(0, I_p)$ and $v^i \sim N(0,1)$ standard independent Gaussian random variables.

Denoting the resulting sample covariance as $\widehat{\Sigma}$, we can obtain an expression for the noise matrix $\Delta = \widehat{\Sigma} - \Sigma$. The result will be similar to expansion (3.31) with h and W appropriately modified; more specifically, we have

$$\tilde{h}_S = h_S, \qquad \tilde{h}_{S^c} = \Gamma^{1/2}_{p-k} h_{S^c} \tag{3.42}$$

$$\widetilde{W}_{SS} = W_{SS}, \quad \widetilde{W}_{S^cS} = \Gamma^{1/2}_{p-k} W_{S^cS}, \quad \widetilde{W}_{S^cS^c} = \Gamma^{1/2}_{p-k} W_{S^cS^c} \Gamma^{1/2}_{p-k}. \tag{3.43}$$

Note that the P-term is unaffected.

Re-examining the proof presented for the case $\Gamma_{p-k} = I_{p-k}$, we can identify conditions imposed on h and W to guarantee optimality. By imposing sufficient constraints on Γ_{p-k}, we can make $\tilde{h}$ and $\widetilde{W}$ satisfy the same conditions. The rest of the proof will then be exactly the same as the case $\Gamma_{p-k} = I_{p-k}$. As before, we proceed by verifying Steps A through C in sequence.

Verifying steps A and B

Examining the proof of Lemma 13, we observe that we need bounds on $\|\tilde{h}_S\|_2$, $\|\tilde{h}_S\|_1$ and $\|\tilde{h}\|_\infty = \max\{\|\tilde{h}_S\|_\infty, \|\tilde{h}_{S^c}\|_\infty\}$. Since $\tilde{h}_S = h_S$, we should only be concerned with $\|\tilde{h}_{S^c}\|_\infty$, for which we simply have

$$\|\tilde{h}_{S^c}\|_\infty \leq |||\Gamma^{1/2}_{p-k}|||_{\infty,\infty} \|h_{S^c}\|_\infty.$$

Thus, assumption (3.2a)—i.e., $\|\Gamma^{1/2}\|_{\infty,\infty} = \mathcal{O}(1)$—guarantees that Lemma 13 also holds for (nonidentity) Γ.

Similarly, for Lemma 12 to hold, we need to investigate $\|\widetilde{W}_{S^cS}\|_{\infty,2}$, since this is the only norm (among those considered in the lemma) affected by a nonidentity Γ. Using sub-multiplicative property of operator norms (cf. (2.11) in §2.1.3), we have

$$\|\widetilde{W}_{S^cS}\|_{\infty,2} \leq \|\Gamma^{1/2}_{p-k}\|_{\infty,\infty}\|W_{S^cS}\|_{\infty,2},$$

so that the same boundedness assumption (3.2a) is sufficient.

Verifying step C

For the lower-right block $\widetilde{W}_{S^cS^c}$, we first have to verify Lemma 15. We also need to examine the proof of Lemma 10 where the result of Lemma 15—namely relation (3.37)—was used. Let $\widetilde{G} = (\tilde{g}^i_j)_{i,j=1,1}^{n,p}$ and let $\widetilde{G}_{S^c} = (\widetilde{G}_{ij})$ for $1 \leq i \leq n$ and $j \in S^c$. Note that $\widetilde{G}^T_{S^c} \in \mathbb{R}^{(p-k)\times n}$ and we have

$$\widetilde{G}^T_{S^c} = \left(\tilde{g}^1_{S^c}, \ldots, \tilde{g}^n_{S^c}\right) = \Gamma^{1/2}_{p-k}\left(g^1_{S^c}, \ldots, g^n_{S^c}\right) = \Gamma^{1/2}_{p-k}G^T_{S^c}.$$

Using this notation, we can write $\widetilde{W}_{S^cS^c} = n^{-1}\widetilde{G}^T_{S^c}\widetilde{G}_{S^c} - \Gamma_{p-k} = \Gamma^{1/2}_{p-k}\left(n^{-1}G^T_{S^c}G_{S^c} - I_{p-k}\right)\Gamma^{1/2}_{p-k}$, consistent with equation (3.43).

Now to establish a version of (3.37), we have to consider the maximum of

$$\|n^{-1/2}\widetilde{G}_{S^c}v\|_2 = \|n^{-1/2}G_{S^c}\Gamma^{1/2}_{p-k}v\|_2$$

over the set where $\|v\|_2 \leq \eta$ and $\|v\|_1 \leq \ell$. Let $\tilde{v} = \Gamma^{1/2}_{p-k}v$ and note that for any consistent pair of vector/matrix norms we have $\|\tilde{v}\| \leq \|\Gamma^{1/2}_{p-k}\|\,\|v\|$. Thus, for example, $\|v\|_2 \leq \eta$ implies $\|\tilde{v}\|_2 \leq \|\Gamma^{1/2}_{p-k}\|_{2,2}\eta$, and similarly for the ℓ_1-norm. Now, if we assume that Lemma 15 holds for G_{S^c}, we obtain, for all $(\eta, \ell) \in \mathbb{S}$, the inequality

$$\begin{aligned}\max_{\substack{\|v\|_2\leq\eta,\\ \|v\|_1\leq\ell}} \|n^{-1/2}\widetilde{G}_{S^c}v\|_2 &\leq \max_{\substack{\|\tilde{v}\|_2\leq\|\Gamma^{1/2}_{p-k}\|_{2,2}\eta,\\ \|\tilde{v}\|_1\leq\|\Gamma^{1/2}_{p-k}\|_{1,1}\ell}} \|n^{-1/2}G_{S^c}\tilde{v}\|_2 \\ &\leq \|\Gamma^{1/2}_{p-k}\|_{2,2}\eta + B\|\Gamma^{1/2}_{p-k}\|_{1,1}\sqrt{\frac{\log m}{n}}\ell + \varepsilon. \end{aligned} \tag{3.44}$$

Thus, one observes that the boundedness condition (3.2a) guarantees that

$$\|\Gamma^{1/2}_{p-k}\|_{1,1} = \|\Gamma^{1/2}_{p-k}\|_{\infty,\infty} \leq A_1,$$

thereby taking care of the second term in equation (3.44). More specifically, the constant A_1 is simply absorbed into some $B' = BA_1$. In addition, we also require a bound on

$\|\Gamma_{p-k}^{1/2}\|_{2,2}$, which follows from our assumption $\|\Gamma_{p-k}\|_{2,2} \leq 1$. However, the fact that the factor multiplying η in (3.44) is no longer unity has to be addressed more carefully.

Recall that inequality (3.37) was used in the proof of Lemma 10 to establish a bound on

$$v^{*T}\Delta_{S^cS^c}v^* = v^{*T}W_{S^cS^c}v^* = v^{*T}(H^TH - I_{p-k})v^* = \|Hv^*\|_2^2 - \|v^*\|_2^2$$

where $H = n^{-1/2}G_{S^c}$. The bound obtained on this term is given by (3.79). We focus on the core idea, omitting some technical details such as the discretization argument[2]. Replacing $W_{S^cS^c}$ with $\widetilde{W}_{S^cS^c}$, we need to establish a similar bound on

$$v^{*T}\widetilde{W}_{S^cS^c}v^* = v^{*T}(n^{-1}\widetilde{G}_{S^c}^T\widetilde{G}_{S^c} - \Gamma_{p-k})v^* = \|n^{-1/2}\widetilde{G}_{S^c}v^*\|_2^2 - \|\Gamma_{p-k}^{1/2}v^*\|_2^2.$$

Note that $\|v^*\|_2 \leq \|\Gamma_{p-k}^{-1/2}\|_{2,2}\|\Gamma_{p-k}^{1/2}v^*\|_2$ or, equivalently, $\|\Gamma_{p-k}^{-1/2}\|_{2,2}^{-1}\|v^*\|_2 \leq \|\Gamma_{p-k}^{1/2}v^*\|_2$. Thus, using (3.44), one obtains

$$\begin{aligned}\|n^{-1/2}\widetilde{G}_{S^c}v^*\|_2^2 - \|\Gamma_{p-k}^{1/2}v^*\|_2^2 \leq \Big(\|\Gamma_{p-k}^{1/2}\|_{2,2}^2 - \|\Gamma_{p-k}^{-1/2}\|_{2,2}^{-2}\Big)\|v^*\|_2^2 \\ + \big(\text{terms of lower order in } \|v^*\|_2\big).\end{aligned}$$

Note that unlike the case $\Gamma_{p-k} = I_{p-k}$, the term quadratic in $\|v^*\|_2$ does not vanish in general. Thus, we have to assume that its coefficient is eventually small compared to β. More specifically, we assume

$$\|\Gamma_{p-k}^{1/2}\|_{2,2}^2 - \|\Gamma_{p-k}^{-1/2}\|_{2,2}^{-2} \;\leq\; \frac{\alpha}{4} = \frac{\beta}{8}, \quad \text{eventually.} \tag{3.45}$$

The boundedness assumptions on $\|\Gamma_{p-k}^{1/2}\|_{1,1}$ and $\|\Gamma_{p-k}^{1/2}\|_{2,2}$ now allows for the rest of the terms to be made less than $\alpha/4$, using arguments similar to the proof of Lemma 10, so that the overall objective is less than $\alpha/2$, eventually. This concludes the proof.

Noting that $\|\Gamma_{p-k}^{1/2}\|_{2,2}^2 = \lambda_{\max}(\Gamma_{p-k})$ and $\|\Gamma_{p-k}^{-1/2}\|_{2,2}^{-2} = \lambda_{\min}(\Gamma_{p-k})$, we can summarize the conditions sufficient for Lemma 10 to extend to general covariance structure as follows

$$\|\Gamma_{p-k}^{1/2}\|_{1,1} \;=\; \|\Gamma_{p-k}^{1/2}\|_{\infty,\infty} \;=\; \mathcal{O}(1) \tag{3.46a}$$

$$\lambda_{\max}(\Gamma_{p-k}) \;\leq\; 1 \tag{3.46b}$$

$$\lambda_{\max}(\Gamma_{p-k}) - \lambda_{\min}(\Gamma_{p-k}) \;\leq\; \frac{\beta}{8}, \tag{3.46c}$$

as stated previously.

[2]In particular, we will assume that v^* saturates (3.44), so that $\|v^*\|_2 = \eta$. For a more careful argument see the proof of Lemma 10.

3.5 Proof of Theorem 8 – minimax lower bound

Our proof is based on the standard approach of applying Fano's inequality (e.g., [30, 52, 104, 101] and § 2.6). Let $\mathbb{S}$ denote the collection of all possible support sets, i.e. the collection of k-subsets of $\{1, \ldots, p\}$ with cardinality $|\mathbb{S}| = \binom{p}{k}$; we view S as a random variable distributed uniformly over $\mathbb{S}$. Let $\mathbb{P}_S$ denote the distribution of a sample $X \sim N(0, \Sigma_p(S))$ from a spiked covariance model, conditioned on the maximal eigenvector having support set S, and let $X^n = (x_1, \ldots, x_n)$ be a set of n i.i.d. samples. In information-theoretic terms, we view any method of support recovery as a decoder that operates on the data X^n and outputs an estimate of the support $\widehat{S} = \phi(X^n)$—in short, a (possibly random) map $\phi : (\mathbb{R}^p)^n \to \mathbb{S}$. Using the 0-1 loss to compare an estimate $\widehat{S}$ and the true support set S, the associated risk is simply the probability of error $\mathbb{P}[\text{error}] = \sum_{S \in \mathbb{S}} \binom{p}{k}^{-1} \mathbb{P}_S[\widehat{S} \neq S]$. Due to symmetry of the ensemble, in fact we only need to restrict our attention to symmetric estimators for which we have $\mathbb{P}[\text{error}] = \mathbb{P}_S[\widehat{S} \neq S]$, where S is some fixed but arbitrary support set, a property that we refer to as *risk flatness* (cf. §2.6.5).

In order to generate suitably tight lower bounds, we restrict attention to the following *sub-collection* $\widetilde{\mathbb{S}}$ of support sets:

$$\widetilde{\mathbb{S}} \;\; := \;\; \{S \in \mathbb{S} \mid \{1, \ldots, k-1\} \subset S\},$$

consisting of those k-element subsets that contain $\{1, \ldots, k-1\}$ and one element from $\{k, \ldots, p\}$. By risk flatness, the probability of error with S chosen uniformly at random from the original ensemble $\mathbb{S}$ is the same as the probability of error with S chosen uniformly from $\widetilde{\mathbb{S}}$. Letting U denote a subset chosen uniformly at random from $\widetilde{\mathbb{S}}$, using Fano's inequality, we have the lower bound

$$\mathbb{P}[\text{error}] \;\; \geq \;\; 1 - \frac{I(U; X^n) + \log 2}{\log |\widetilde{\mathbb{S}}|},$$

where $I(U; X^n)$ is the mutual information between the data X^n and the randomly chosen support set U, and $|\widetilde{\mathbb{S}}| = p - k + 1$ is the cardinality of $\widetilde{\mathbb{S}}$.

It remains to obtain an upper bound on $I(U; X^n) = H(X^n) - H(X^n|U)$. By chain rule for entropy, we have $H(X^n) \leq nH(x)$. Next, using the maximum entropy property of the Gaussian distribution [30], we have

$$H(X^n) \;\leq\; nH(x) \;\leq\; n\Big\{\frac{p}{2}[1 + \log(2\pi)] + \frac{1}{2}\log\det \mathbb{E}\,[xx^T]\Big\}, \tag{3.47}$$

where $\mathbb{E}[xx^T]$ is the covariance matrix of x. On the other hand, given $U = \overline{U}$, the vector X^n is a collection of n Gaussian p-vectors with covariance matrix $\Sigma_p(\overline{U})$. The determinant of this matrix is $1 + \beta$, independent of $\overline{U}$, so that we have

$$H(X^n|U) \;\; = \;\; \frac{np}{2}\Big[1 + \log(2\pi)\Big] + \frac{n}{2}\log(1 + \beta). \tag{3.48}$$

Combining equations (3.47) and (3.48), we obtain

$$I(U; X^n) \leq \frac{n}{2} \left\{ \log \det \mathbb{E}\left[xx^T\right] - \log(1+\beta) \right\}. \tag{3.49}$$

The following lemma, proved in Appendix 3.E, specifies the form of the log determinant of the covariance matrix $\Sigma_M := \mathbb{E}[xx^T]$.

Lemma 16. *The log determinant has the exact expression*

$$\log \det \Sigma_M \;=\; \log(1+\beta) + \log\left(1 - \frac{\beta}{1+\beta}\frac{p-k}{k(p-k+1)}\right) + (p-k)\log\left(1 + \frac{\beta}{k(p-k+1)}\right). \tag{3.50}$$

Substituting equation (3.50) into equation (3.49) and using the inequality $\log(1+\alpha) \leq \alpha$, we obtain

$$\begin{aligned} I(U; X^n) &\leq \frac{n}{2}\left\{ \log\left(1 - \frac{\beta}{1+\beta}\frac{p-k}{k(p-k+1)}\right) + (p-k)\log\left(1 + \frac{\beta}{k(p-k+1)}\right)\right\} \\ &\leq \frac{n}{2}\left\{ -\frac{\beta}{1+\beta}\frac{p-k}{k(p-k+1)} + \frac{\beta(p-k)}{k(p-k+1)}\right\} \\ &= \frac{n}{2}\left\{ \frac{\beta^2}{1+\beta}\frac{p-k}{k(p-k+1)}\right\} \\ &\leq \frac{\beta^2}{2(1+\beta)}\frac{n}{k}. \end{aligned}$$

From the Fano bound (3.47), the error probability is greater than $\frac{1}{2}$ if $\frac{\beta^2}{1+\beta}\frac{n}{k} < \log(p-k) < \log|\widetilde{\mathbb{S}}|$, which completes the proof.

3.6 Some results on ℓ_q sparsity

In this section, we provide some partial analysis of the SDP for the spiked covariance model with ℓ_q sparsity model. In particular, we assume a sampling model ? with the following constraint on z^*,

$$\sum_{j=1}^{p} |z_j^*|^q \leq \kappa \tag{3.51}$$

for some $q \in (0,2)$ and $\kappa = \kappa(p) > 0$. Recall that $\widehat{Z}$ is the solution and λ_n is the regularization parameter of SDP (3.9). Let us define

$$Z^* := z^*(z^*)^T.$$

We have the following theorem.

Theorem 9. *Let* $\lambda_n = c_1\kappa\sqrt{\frac{\beta \log p}{n}}$, *assume (3.51) holds and* $\|z^*\|_2 = 1$. *Then,*

$$\|\widehat{Z} - Z^*\|_{HS}^2 \leq c_2(\beta)\,\kappa^2\big(\frac{\log p}{n}\big)^{\frac{1}{1+q}} \tag{3.52}$$

with high probability at least $1 - c_3 p^{-c_4} - c_5 e^{-c_6 n}$.

The $c_2(\beta)$ constant in (3.52) is decreasing in β and can be taken to be $c_2(\beta) \sim \beta^{\frac{-q}{1+q}}$ as $\beta \to \infty$. The proof relies on the following lemma.

Lemma 17. *Let* f *be a convex function defined on a convex subset* $\mathcal{C}$ *of a normed space* $(\mathcal{X}, \|\cdot\|)$. *Let* $\widehat{x}$ *be a minimizer of* f *on* $\mathcal{C}$, *i.e.,*

$$f(\widehat{x}) = \inf_{x \in \mathcal{C}} f(x) \tag{3.53}$$

Furthermore assume that for some $x_0 \in \mathcal{C}$ *and* $R > 0$, *we have for all* $x \in \mathcal{C}$

$$\|x - x_0\| = R \implies f(x) > f(x_0). \tag{3.54}$$

Then, we have $\|\widehat{x} - x_0\| \leq R$.

Proof. Without loss of generality, we can assume $x_0 = 0$ and $f(x_0) = 0$, since $\widehat{x}$ is a minimizer of f on $\mathcal{C}$ if and only if $\widehat{x} - x_0$ is a minimizer of $g(x) = f(x + x_0) - f(x_0)$ over $x \in \mathcal{C} - x_0$.

Now, assume that the conclusion does not hold, $\|\widehat{x}\| > R$. Then, there exists $\lambda \in (0, 1)$ such that $z = (1 - \lambda)x_0 + \lambda\widehat{x}$ has norm R. (This holds for example with $\lambda = \frac{R}{\|\widehat{x}\|}$). Then,

$$f(z) \leq (1 - \lambda)f(x_0) + \lambda f(\widehat{x}) = \lambda f(\widehat{x}) \leq \lambda f(x_0) = 0. \tag{3.55}$$

But since $\|z\|_2 = R$, by assumption (3.54) we should have $f(z) > f(x_0) = 0$ which is a contradiction. □

3.6.1 Proof of Theorem 9

We will apply the Lemma (3.54) to the function $f(Z) = -\langle\langle\widehat{\Sigma} - I_p, Z\rangle\rangle + \lambda_n\|Z\|_1$ over the set $\mathcal{C} = \mathbb{S}_+^p \cap \{Z : \mathrm{tr}(Z) = 1\}$, with $x_0 = Z^*$ and $\widehat{x} = \widehat{Z}$. To simplify notation, let $\widehat{E} := \widehat{Z} - Z^*$ and $E := Z - Z^*$. Then,

$$f(Z) - f(Z^*) = -\langle\langle\widehat{\Sigma} - I_p, E\rangle\rangle + \lambda_n(\|Z^* + E\|_1 - \|Z^*\|_1) \tag{3.56}$$

Our strategy is to find a value of R such that for all $\|E\|_2 = \|E\|_{HS} = R$, we have $f(Z) - f(Z^*) > 0$ with high probability. It then follows from the lemma that $\|E\|_{HS} \leq R$.

Lemma 18. *We have* $\langle\langle -Z^*, E\rangle\rangle \geq \frac{1}{2}\|E\|_{HS}^2$.

Proof. In general, we have $\|\!|Z|\!\|_{HS} \le \|\!|Z|\!\|_*$ where $\|\!|\cdot|\!\|_*$ is the nuclear norm of the matrix—cf. §?. For a matrix $Z \in \mathbb{S}^p_+$, $\|\!|Z|\!\|_* = \mathrm{tr}(Z)$. It follows that, for $Z \in \mathcal{C}$,

$$\|\!|Z|\!\|_{HS} \le \mathrm{tr}(Z) = 1. \tag{3.57}$$

Now,

$$1 \ge \|\!|Z|\!\|^2_{HS} = \|\!|Z^* + E|\!\|^2_{HS} = \|\!|E|\!\|^2_{HS} + \|\!|Z^*|\!\|^2_{HS} + 2\langle\!\langle Z^*, E\rangle\!\rangle.$$

Using $\|\!|Z^*|\!\|_{HS} = \sqrt{\langle z^*, z^*\rangle} = 1$ and rearranging gives the inequality. □

As was observed previously (cf. ?), we can write $\widehat{\Sigma} = \beta(\frac{1}{n}\sum_i v_i^2)Z^* + \sqrt{\beta}(z^*h^T + h(z^*)^T) + \frac{1}{n}G^TG$, where $h = \frac{1}{n}\sum_{i=1}^n v_i g_i$ and $G = (g_{ij}) \in \mathbb{R}^{n\times p}$ is a standard Gaussian matrix. Since $E \in \mathbb{S}^p$, we have

$$-\langle\!\langle \widehat{\Sigma} - I_p, E\rangle\!\rangle = \underbrace{-\beta\big(\frac{1}{n}\sum_{i=1}^n v_i^2\big)\langle\!\langle Z^*, E\rangle\!\rangle}_{T_0} \underbrace{-\langle\!\langle 2\sqrt{\beta}h(z^*)^T + \frac{1}{n}G^TG - I_p, E\rangle\!\rangle}_{\widetilde{\Delta}} \tag{3.58}$$

By χ^2 concentration, $\frac{1}{n}\sum_i v_i^2 \ge \frac{1}{2}$ with probability at least $1 - \exp(-n/64)$. Hence, with the same probability, the first in (3.58) bounded below as $T_0 \ge -\frac{1}{2}\beta\langle\!\langle Z^*, E\rangle\!\rangle \ge \beta\frac{1}{4}\|\!|E|\!\|^2_{HS}$. As for the term involving $\widetilde{\Delta}$, we first note that the following bound on its vector ℓ_∞ norm which follows from our previous discussions.

Lemma 19. *With probability at least* $1 - c_2p^{-c_3}$

$$\|\widetilde{\Delta}\|_\infty \le c_1\sqrt{\frac{\beta\log p}{n}}.$$

Consider the set $S := \{i : |z^*_i| \ge \tau\} \subset [p]$, where $\tau > 0$ is some threshold to be determined later, and let $k = \mathrm{card}(S)$. We denote by $Z^*_S \subset \mathbb{R}^{k^2}$, a vector consisting of all elements Z^*_{ij}, $(i,j) \in S^2$. Similarly, let $Z^*_{S^c} \subset \mathbb{R}^{p^2-k^2}$ denote the vector consisting of Z^*_{ij}, $(i,j) \notin S^2$. Similar notations will used for $\widetilde{\Delta}$ and E.

Lemma 20. *As long as* $\|\widetilde{\Delta}\|_\infty \le \lambda_n$, *we have*

$$T_1 := -\langle\widetilde{\Delta}_S, E_S\rangle + \lambda_n(\|Z^*_S + E_S\|_1 - \|Z^*_S\|_1) \ge -2\lambda_n k\|E_S\|_2 \tag{3.59}$$

$$T_2 := -\langle\widetilde{\Delta}_{S^c}, E_{S^c}\rangle + \lambda_n(\|Z^*_{S^c} + E_{S^c}\|_1 - \|Z^*_{S^c}\|_1) \ge -4\lambda_n\tau^{1-q}\kappa^2. \tag{3.60}$$

We take $\lambda_n = c_1\sqrt{\frac{\beta\log p}{n}}$ so that the consequences of Lemma 20 hold. Note that $\|E_S\|_2 \le \|\!|E|\!\|_{HS} = R$. From Lemma 23 in Appendix 3.G, $k \le \tau^{-q}\kappa$. Then, putting the pieces together we have

$$f(Z) - f(Z^*) = T_0 + T_1 + T_2 \ge \frac{\beta}{4}R^2 - 2\lambda_n\tau^{-q}\kappa R - 4\lambda_n\tau^{1-q}\kappa^2. \tag{3.61}$$

Choose τ so that $\lambda_n \tau^{1-q}\kappa^2 = (\lambda_n \tau^{-q}\kappa)^2$ or equivalently, $\tau = \lambda_n^{\frac{1}{1+q}}$. Then, taking $R = c_3\lambda_n\tau^{-q}\kappa$ for some sufficiently large $c_3 = c_3(\beta) > 0$, we can make the right-hand side of (3.61) positive[3], implying that $f(Z) - f(Z^*) > 0$. Some algebra shows that

$$R = c_3\kappa\lambda_n^{\frac{1}{1+q}} = c_4\kappa\big(\frac{\log p}{n}\big)^{\frac{1}{2(1+q)}} \tag{3.62}$$

where $c_4 > 0$ could depend on β. The proof is complete. The proofs of auxiliary lemmas can be found in Appendix 3.G.

Appendix 3.A Large deviations for χ^2 variates

The following large-deviations bounds for centralized χ^2 are taken from Laurent and Massart [63]. Given a centralized χ^2-variate X with d degrees of freedom, then for all $x \geq 0$,

$$\mathbb{P}\Big\{X - d \geq 2\sqrt{d\,x} + 2x\Big\} \leq \exp(-x), \qquad \text{and} \tag{3.63a}$$

$$\mathbb{P}\Big\{X - d \leq -2\sqrt{d\,x}\Big\} \leq \exp(-x). \tag{3.63b}$$

We also use the following slightly different version of the bound (3.63a),

$$\mathbb{P}\Big\{X - d \geq d\,x\Big\} \leq \exp\left(-\frac{3}{16}\,d\,x^2\right), \qquad 0 \leq x < \frac{1}{2} \tag{3.64}$$

due to Johnstone [56]. More generally, the analogous tail bounds for *non-central* χ^2, taken from Birgé [19], can be established via the Chernoff bound. Let X be a noncentral χ^2 variable with d degrees of freedom and noncentrality parameter $\nu \geq 0$. Then for all $x > 0$,

$$\mathbb{P}\Big\{X \geq (d+\nu) + 2\sqrt{(d+2\nu)x} + 2x\Big\} \leq \exp(-x), \qquad \text{and} \tag{3.65a}$$

$$\mathbb{P}\Big\{X \leq (d+\nu) - 2\sqrt{(d+2\nu)x}\Big\} \leq \exp(-x). \tag{3.65b}$$

We derive here a slightly weakened but useful form of the bound (3.65a), valid when ν satisfies $\nu \leq Cd$ for a positive constant C. Under this assumption, then for any $\delta \in (0,1)$, we have

$$\mathbb{P}\Big\{X \geq (d+\nu) + 4d\sqrt{\delta}\Big\} \leq \exp\left(-\frac{\delta}{1+2C}d\right). \tag{3.66}$$

[3] With this choice of R, the RHS becomes $(\frac{\beta}{4}c_3^2 - 2c_3 - 4)\lambda_n\tau^{-q}\kappa$, which is positive if $c_3 > \frac{4}{\beta}(1+\sqrt{1+\beta})$.

To establish this bound, let $x = \frac{d^2\delta}{d+2\nu}$ for some $\delta \in (0,1)$. From equation (3.65a), we have

$$p^* \;:=\; \mathbb{P}\Big\{X \geq (d+\nu) + 2d\sqrt{\delta} + 2\frac{d^2}{d+2\nu}\delta\Big\} \;\leq\; \exp\left(-\frac{d^2\delta}{d+2\nu}\right) \;\leq\; \exp\left(-\frac{\delta}{1+2C}d\right).$$

Moreover, we have

$$p^* \;\geq\; \mathbb{P}\Big\{X \geq (d+\nu) + 2d\sqrt{\delta} + 2d\delta\Big\} \;\geq\; \mathbb{P}\Big\{X \geq (d+\nu) + 4d\sqrt{\delta}\Big\},$$

since $\sqrt{\delta} \geq \delta$ for $\delta \in (0,1)$.

Appendix 3.B Proof of Lemma 6

Using the form of the χ^2_n PDF, we have, for even n and any $t > 0$,

$$\begin{aligned}
\mathbb{P}\Big\{\frac{\chi^2_n}{n} > 1+t\Big\} &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_{(1+t)n}^{\infty} x^{\frac{n}{2}-1}\exp(-x/2)dx \\
&= \frac{1}{2^{n/2}\Gamma(n/2)} \left\{ \frac{(n/2-1)!}{\left(\frac{1}{2}\right)^{(n/2-1)+1}} \exp\left(-\frac{n(1+t)}{2}\right) \sum_{i=0}^{n/2-1} \frac{1}{i!}\left(\frac{n(1+t)}{2}\right)^i \right\} \\
&\geq \exp(-nt/2)\left\{\frac{\exp(-n/2)\,(n/2)^{n/2-1}}{(n/2-1)!}\right\}(1+t)^{n/2-1}
\end{aligned}$$

where the second line uses standard integral formula (cf. §3.35 in the reference book [45]). Using Stirling's approximation for $(n/2-1)!$, the term within square brackets is lower bounded by $2C/\sqrt{n}$. Also, over $t \in (0,1)$, we have $(1+t)^{-1} > 1/2$, so we conclude that

$$\mathbb{P}\Big\{\frac{\chi^2_n}{n} > 1+t\Big\} \;\geq\; \frac{C}{\sqrt{n}}\exp\left(-\frac{n}{2}\Big\{t - \log(1+t)\Big\}\right). \tag{3.67}$$

Defining the function $f(t) = \log(1+t)$, we calculate $f(0) = 0$, $f'(0) = 1$ and $f''(t) = -1/(1+t)^2$. Note that $f''(t) \geq -1$, for all $t \in \mathbb{R}$. Consequently, via a second-order Taylor series expansion, we have $f(t) - t \geq -t^2/2$. Substituting this bound into equation (3.67) yields

$$\mathbb{P}\Big\{\frac{\chi^2_n}{n} > 1+t\Big\} \;\geq\; \frac{C}{\sqrt{n}}\exp\left(-\frac{nt^2}{2}\right)$$

as claimed.

Appendix 3.C Proofs for §3.4.2

3.C.1 Proof of Lemma 8

The argument we present here has a deterministic nature. In other words, we will show that if the conditions of the lemma hold for a nonrandom sequence of matrices Δ_{SS}, the conclusions will follow. Thus, for example, all the references to limits may be regarded as deterministic. Then, since the conditions of the lemma are assumed to hold for a random Δ_{SS} a.a.s., it immediately follows that the conclusions hold a.a.s.. To simplify the argument let us assume that $\alpha^{-1}|\!|\!|\Delta_{SS}|\!|\!|_{\infty,\infty} \leq \varepsilon$ for sufficiently small $\varepsilon > 0$; it turns out that $\varepsilon = \frac{1}{10}$ is enough.

We prove the lemma in steps. First, by Weyl's theorem (cf. Theorem 1 in § 2.2 and [54, 89]) , eigenvalues of the perturbed matrix $\alpha z_S^* z_S^{*T} + \Delta_{SS}$ are contained in intervals of length $2|\!|\!|\Delta_{SS}|\!|\!|_{2,2}$ centered at eigenvalues of $\alpha z_S^* z_S^{*T}$. Since the matrix $z_S^* z_S^{*T}$ is rank-one, one eigenvalue of the perturbed matrix is in the interval $[\alpha \pm |\!|\!|\Delta_{SS}|\!|\!|_{2,2}]$, and the remaining $k-1$ eigenvalues are in the interval $[0 \pm |\!|\!|\Delta_{SS}|\!|\!|_{2,2}]$. Since by assumption $2|\!|\!|\Delta_{SS}|\!|\!|_{2,2} \leq \alpha$ eventually, the two intervals are disjoint, and the first one contains the maximal eigenvalue γ_1 while the second contains the second largest eigenvalue γ_2. In other words, $|\gamma_1 - \alpha| \leq |\!|\!|\Delta_{SS}|\!|\!|_{2,2}$ and $|\gamma_2| \leq |\!|\!|\Delta_{SS}|\!|\!|_{2,2}$. Since $|\!|\!|\Delta_{SS}|\!|\!|_{2,2} \to 0$ by assumption, we conclude that $\gamma_1 \to \alpha$ and $\gamma_2 \to 0$. For the rest of the proof, take n large enough so

$$|\gamma_1 \alpha^{-1} - 1| \;\leq\; \varepsilon, \tag{3.68}$$

where $\varepsilon > 0$ is a small number to be determined.

Now, let $\widehat{z}_S \in \mathbb{R}^k$ with $\|\widehat{z}_S\|_2 = 1$ be the eigenvector associated with γ_1, i.e.

$$(\alpha z_S^* z_S^{*T} + \Delta_{SS})\widehat{z}_S = \gamma_1 \widehat{z}_S. \tag{3.69}$$

Taking inner products with $\widehat{z}_S$, one obtains $\alpha(z_S^{*T}\widehat{z}_S)^2 + \widehat{z}_S^T \Delta_{SS} \widehat{z}_S = \gamma_1$. Noting that $|\widehat{z}_S^T \Delta_{SS} \widehat{z}_S|$ is upper-bounded by $|\!|\!|\Delta_{SS}|\!|\!|_{2,2}$, we have by triangle inequality

$$\begin{aligned} |\alpha - \alpha(z_S^{*T}\widehat{z}_S)^2| &= |\alpha - \gamma_1 + \gamma_1 - \alpha(z_S^{*T}\widehat{z}_S)^2| \\ &\leq |\alpha - \gamma_1| + |\gamma_1 - \alpha(z_S^{*T}\widehat{z}_S)^2| \leq 2|\!|\!|\Delta_{SS}|\!|\!|_{2,2} \end{aligned}$$

which implies $z_S^{*T}\widehat{z}_S \to 1$ (taking into account our sign convention). Take n large enough so that

$$|z_S^{*T}\widehat{z}_S - 1| \leq \varepsilon, \tag{3.70}$$

and let u be the solution of

$$\alpha z_S^* + \Delta_{SS} u = \alpha u \tag{3.71}$$

which is an approximation of equation (3.69) satisfied by $\widehat{z}_S$. Using triangle inequality, one has $\|u\|_\infty \leq \|z_S^*\|_\infty + \alpha^{-1}|\!|\!|\Delta_{SS}|\!|\!|_{\infty,\infty}\|u\|_\infty$, which implies that

$$\|u\|_\infty \leq (1 - \alpha^{-1}|\!|\!|\Delta_{SS}|\!|\!|_{\infty,\infty})^{-1}\|z_S^*\|_\infty \leq (1-\varepsilon)^{-1}\|z_S^*\|_\infty. \tag{3.72}$$

We also have

$$\|u - z_S^*\|_\infty \leq \alpha^{-1} |||\Delta_{SS}|||_{\infty,\infty} \|u\|_\infty \leq \varepsilon(1-\varepsilon)^{-1} \|z_S^*\|_\infty. \tag{3.73}$$

Subtracting equation (3.71) from equation (3.69), we obtain $\alpha z_S^*(z_S^{*T}\widehat{z}_S - 1) + \Delta_{SS}(\widehat{z}_S - u) = \gamma_1 \widehat{z}_S - \alpha u$. Adding and subtracting $\gamma_1 u$ on the right-hand side and dividing by α, we have

$$z_S^*(z_S^{*T}\widehat{z}_S - 1) + \alpha^{-1}\Delta_{SS}(\widehat{z}_S - u) = \gamma_1\alpha^{-1}(\widehat{z}_S - u) + (\gamma_1\alpha^{-1} - 1)u,$$

which implies

$$\begin{aligned}\|\widehat{z}_S - u\|_\infty &\leq \Big(|\gamma_1\alpha^{-1}| - \alpha^{-1}|||\Delta_{SS}|||_{\infty,\infty}\Big)^{-1}\Big\{|z_S^{*T}\widehat{z}_S - 1| \cdot \|z_S^*\|_\infty + |\gamma_1\alpha^{-1} - 1| \cdot \|u\|_\infty\Big\} \\ &\leq (1-2\varepsilon)^{-1}[\varepsilon + \varepsilon(1-\varepsilon)^{-1}] \cdot \|z_S^*\|_\infty\end{aligned}$$

where the last inequality follows from equations (3.68), (3.70) and (3.72). Combining with the bound (3.73) on $\|uz_S^*\|_\infty$ yields

$$\begin{aligned}\frac{\|\widehat{z}_S - z_S^*\|_\infty}{\|z_S^*\|_\infty} &\leq \frac{\varepsilon}{1-2\varepsilon} + \frac{\varepsilon}{(1-2\varepsilon)(1-\varepsilon)} + \frac{\varepsilon}{1-\varepsilon} \\ &\leq \frac{3\varepsilon}{(1-2\varepsilon)^2}.\end{aligned}$$

Finally, we take $\varepsilon = \frac{1}{10}$ to conclude $\|\widehat{z}_S - z_S^*\|_\infty \leq \frac{1}{2}\|z_S^*\|_\infty = \frac{1}{2\sqrt{k}}$ a.a.s., as claimed.

3.C.2 Proof of Lemma 9

Recall that by definition, $\widetilde{z}_S = \widehat{z}_S/\|\widehat{z}_S\|_1$. Using the identity $\operatorname{sign}(\widehat{z}_S)^T\widehat{z}_S = \|\widehat{z}_S\|_1$ yields $\widehat{U}_{S^cS}\widehat{z}_S = \lambda_n^{-1}\Delta_{S^cS}\widehat{z}_S$, which is the desired equation. It only remains to prove that $\widehat{U}_{S^cS}$ is indeed a valid sign matrix.

First note that from equation (3.23) we have $|\widehat{z}_i| \in \left[\frac{1}{2\sqrt{k}}, \frac{3}{2\sqrt{k}}\right]$ for $i \in S$, which implies that $\|\widehat{z}_S\|_1 \in \left[\frac{\sqrt{k}}{2}, \frac{3\sqrt{k}}{2}\right]$. Thus, $\|\widetilde{z}_S\|_2 = 1/(\|\widehat{z}_S\|_1) \leq \frac{2}{\sqrt{k}}$.

Now we can write

$$\begin{aligned}\max_{i\in S^c, j\in S} |\widehat{U}_{ij}| &\leq \lambda_n^{-1}\|\Delta_{S^cS}\,\widetilde{z}_S\|_\infty \\ &\leq \lambda_n^{-1}|||\Delta_{S^cS}|||_{\infty,2}\,\|\widetilde{z}_S\|_2 \\ &\leq \frac{2k}{\beta}\frac{\delta}{\sqrt{k}}\frac{2}{\sqrt{k}} \\ &= \frac{4}{\beta}\,\delta,\end{aligned}$$

so that taking $\delta \leq \frac{\beta}{4}$ completes the proof.

3.C.3 Proof of Lemma 10

Here we provide the proof for the case $\Gamma_{p-k} = I_{p-k}$; necessary modifications for the general case are discussed in §3.4.4. First, let us bound the cross-term in equation (3.27). Recall that $\widetilde{z}_S = \widehat{z}_S/\|\widehat{z}_S\|_1$. Also, by our choice (3.25) of $\widehat{U}_{S^cS}$, we have

$$\Phi_{S^cS} = \Delta_{S^cS} - \lambda_n \widehat{U}_{S^cS} = \Delta_{S^cS} - \Delta_{S^cS}\widetilde{z}_S \operatorname{sign}(\widehat{z}_S)^T.$$

Now, using sub-multiplicative property of operator norms [see relation (2.11) in §2.1.3], we can write

$$\begin{aligned} |||\Phi_{S^cS}|||_{\infty,2} &= |||\Delta_{S^cS}\big(I_{p-k} - \widetilde{z}_S \operatorname{sign}(\widehat{z}_S)^T\big)|||_{\infty,2} \\ &\leq |||\Delta_{S^cS}|||_{\infty,2} \cdot |||I_{p-k} - \widetilde{z}_S \operatorname{sign}(\widehat{z}_S)^T|||_{2,2} \\ &\leq |||\Delta_{S^cS}|||_{\infty,2} \cdot \big(1 + \|\widetilde{z}_S\|_2 \, \|\operatorname{sign}(\widehat{z}_S)\|_2\big) \leq 3|||\Delta_{S^cS}|||_{\infty,2}, \end{aligned} \tag{3.74}$$

where we have also used the facts that $|||ab^T|||_{2,2} = \|a\|_2\|b\|_2$, and $\|\widetilde{z}_S\|_2 = 1/(\|\widehat{z}_S\|_1) \leq \frac{2}{\sqrt{k}}$, using the bound (28). Recall the decomposition $x = (u, v)$, where $u = \mu\widehat{z}_S + \widehat{z}_S^{\perp}$ with $\mu^2 + \|\widehat{z}_S^{\perp}\|_2^2 \leq 1$. Also, by our choice (3.25) of $\widehat{U}_{S^cS}$, we have $\Phi_{S^cS}u = \Phi_{S^cS}\widehat{z}_S^{\perp}$. Thus,

$$\max_u |2v^T\Phi_{S^cS}u| \;\leq \max_{\substack{\|\tilde{u}\|_2 \leq \sqrt{1-\mu^2}, \\ \tilde{u} \perp z_S}} |2v^T\Phi_{S^cS}\tilde{u}| \;\leq\; \sqrt{1-\mu^2} \max_{\|\tilde{u}\|_2 \leq 1} |2v^T\Phi_{S^cS}\tilde{u}|. \tag{3.75}$$

Using Hölder's inequality (cf. (2.4)), we have

$$\begin{aligned} \max_{\|\tilde{u}\|_2 \leq 1} |2v^T\Phi_{S^cS}\tilde{u}| &\leq 2\|v\|_1 \max_{\|\tilde{u}\|_2 \leq 1} \|\Phi_{S^cS}\tilde{u}\|_\infty \\ &\leq 2\|v\|_1 \, |||\Phi_{S^cS}|||_{\infty,2} \\ &\leq 6\|v\|_1 \frac{\delta}{\sqrt{k}} \end{aligned} \tag{3.76}$$

where we have used bound (3.74) and applied condition (3.26). We now turn to the last term in the decomposition (3.27), namely $v^T\Phi_{S^cS^c}v = v^T\Delta_{S^cS^c}v - \lambda_n\, v^T\widehat{U}_{S^cS^c}v$. In order to minimize this term, we use our freedom to choose $\widehat{U}_{S^cS^c}(x) = \operatorname{sign}(v)\operatorname{sign}(v)^T$, so that $-\lambda_n\, v^T\widehat{U}_{S^cS^c}v$ simply becomes $-\lambda_n\|v\|_1^2$.

Define the objective function $f^* := \max_x x^T\Phi x$. Also let $H = n^{-1/2}G_{S^c}$, where $G_{S^c} = (G_{ij})$ for $1 \leq i \leq n$ and $j \in S^c$. Noting that $\Delta_{S^cS^c} = H^TH - I_m$ (with $m = p-k$) and using the bounds (3.28), (3.75) and (3.76), we obtain the following bound on the objective

$$\begin{aligned} f^* &\leq \max_u \; u^T\Phi_{SS}u + \max_{u,v} \; 2v^T\Phi_{S^cS}u + \max_v v^T\Phi_{S^cS^c}v \\ &\leq \Big\{\mu^2\gamma_1 + (1-\mu^2)\gamma_2\Big\} + \; (1-\mu^2) \underbrace{\Big[\max_{\|v\|_2 \leq 1} \Big\{6\|v\|_1\frac{\delta}{\sqrt{k}} + \|Hv\|_2^2 - \|v\|_2^2 - \lambda_n\|v\|_1^2\Big\}\Big]}_{g^*}. \end{aligned} \tag{3.77}$$

In obtaining the last inequality, we have used the change of variable $v \to (\sqrt{1-\mu^2})\, v$, with some abuse of notation, and exploited the inequality $\|v\|_2 \le \sqrt{1-\mu^2}$. (Note that this bound follows from the identity $\|x\|_2^2 = 1 = \mu^2 + \|\widehat{z}_S^{\perp}\|_2^2 + \|v\|_2^2$.)

Let v^* be the optimal solution to problem g^* in equation (3.77); note that it is random due to the presence of H. Also, set $\mathbb{S} = \{(\eta_{ij}, \ell_{ij})\}$ where i and j range over $\{1, 2, \ldots, \lceil\sqrt{m}\,\rceil\}$ and

$$\eta_{ij} = \frac{i}{\sqrt{m}}, \quad \ell_{ij} = \frac{i}{\sqrt{m}}\, j.$$

Note that $\mathbb{S}$ satisfies the condition of the lemma, namely $|\mathbb{S}| = \lceil\sqrt{m}\,\rceil^2 = \mathcal{O}(m)$.

Since $\|v^*\|_2 \le 1$, and $\|v^*\|_2 \le \|v^*\|_1 \le \sqrt{m}\,\|v^*\|_2$, there exists[4] $(\eta^*, \ell^*) \in \mathbb{S}$ such that

$$\eta^* - \frac{1}{\sqrt{m}} < \|v^*\|_2 \le \eta^*$$
$$\ell^* - 3 < \|v^*\|_1 \le \ell^*$$

Thus, using condition (3.29), we have

$$\|Hv^*\|_2 \le \max_{\substack{\|v\|_2 \le \eta^*, \\ \|v\|_1 \le \ell^*}} \|Hv\|_2 \le \eta^* + \frac{\delta}{\sqrt{k}}\ell^* + \varepsilon \le \|v^*\|_2 + \frac{1}{\sqrt{m}} + \frac{\delta}{\sqrt{k}}(\|v^*\|_1 + 3) + \varepsilon.$$

To simplify notation, let

$$A = A(\varepsilon, \delta, m, k) := 1/\sqrt{m} + 3\,\delta/\sqrt{k} + \varepsilon, \tag{3.78}$$

so that the bound in the above display may be written as $\|v^*\|_2 + \delta\|v^*\|_1/\sqrt{k} + A$. Now, we have

$$\begin{aligned}\|Hv^*\|_2^2 - \|v^*\|_2^2 &\le 2\|v^*\|_2 \left(\delta\frac{\|v^*\|_1}{\sqrt{k}} + A\right) + \left(\delta\frac{\|v^*\|_1}{\sqrt{k}} + A\right)^2 \\ &\le 2\left(\delta\frac{\|v^*\|_1}{\sqrt{k}} + A\right) + \left(\delta\frac{\|v^*\|_1}{\sqrt{k}} + A\right)^2. \end{aligned} \tag{3.79}$$

[4] Let $i^* = \lceil\sqrt{m}\,\|v^*\|_2\rceil$ and $\eta^* = \frac{i^*}{\sqrt{m}}$. Using the fact that, for any $x \in \mathbb{R}$, $\lceil x\rceil - 1 < x \le \lceil x\rceil$, we have $\eta^* - 1/\sqrt{m} < \|v^*\|_2 \le \eta^*$ or, equivalently, $\|v^*\|_2 = \eta^* + \xi$ where $-1/\sqrt{m} < \xi \le 0$. Now let $j^* = \left\lceil \frac{\|v^*\|_1}{\|v^*\|_2} \right\rceil$. One has $(j^* - 1)\|v^*\|_2 < \|v^*\|_1 \le j^*\|v^*\|_2$ which, using the fact that $\|v^*\|_2 \le 1$, implies $j^*\|v^*\|_2 - 1 < \|v^*\|_1 \le j^*\|v^*\|_2$. This in turn implies

$$j^*\eta^* + j^*\xi - 1 < \|v^*\|_1 \le j^*\eta^*$$

Take $\ell^* = j^*\eta^*$ and note that $j^*\xi - 1 > -3$, since j^* is at most $\lceil\sqrt{m}\,\rceil$.

Using this in (3.77) and recalling from (3.21) that $\lambda_n = \beta/(2k)$, we obtain the following bound

$$g^* \leq 6\,\delta\frac{\|v^*\|_1}{\sqrt{k}} + 2\left(\delta\frac{\|v^*\|_1}{\sqrt{k}} + A\right) + \left(\delta\frac{\|v^*\|_1}{\sqrt{k}} + A\right)^2 - \frac{\beta}{2}\left(\frac{\|v^*\|_1}{\sqrt{k}}\right)^2.$$

Note that this is quadratic in $\|v^*\|_1/\sqrt{k}$, i.e.

$$g^* \leq a\left(\frac{\|v^*\|_1}{\sqrt{k}}\right)^2 + b\left(\frac{\|v^*\|_1}{\sqrt{k}}\right) + c$$

where

$$a = \delta^2 - \frac{\beta}{2}, \qquad b = 8\,\delta + 2\delta A, \quad \text{and} \quad c = 2A + A^2.$$

By choosing δ sufficiently small, say $\delta^2 \leq \beta/4$, we can make a negative. This makes the quadratic form $ax^2 + bx + c$ achieve a maximum of $c + b^2/4(-a)$, at the point $x^* = b/2(-a)$. Note that we have $b/2(-a) \to 0$ and $c \to 0$ as $\varepsilon, \delta \to 0$ and $m, k \to \infty$. Consequently, we can make this maximum (and hence g^*) arbitrarily small eventually, say less than $\alpha/2$, by choosing δ and ε sufficiently small.

Combining this bound on g^* with our bound (3.77) on f^*, and recalling that $\gamma_1 \to \alpha$ and $\gamma_2 \to 0$ by Lemma 8, we conclude that

$$f^* \;\leq\; \mu^2\big(\alpha + o(1)\big) + (1-\mu^2)\Big\{\frac{\alpha}{2} + o(1)\Big\} \;\leq\; \alpha + o(1),$$

as claimed.

Appendix 3.D Proof of Lemma 11

In this section, we prove Lemma 11, a general result on $\|\!|\cdot|\!\|_{\infty,2}$-norm of Wishart matrices. Some of the intermediate results are of independent interest and are stated as separate lemmas. Two sets of large deviation inequalities will be used, one for chi-squared RVs χ^2_n and one for "sums of Gaussian product" random variates. To define the latter precisely, let Z_1 and Z_2 be independent Gaussian RVs, and consider the sum $\sum_{i=1}^n X_i$ where $X_i \overset{\text{iid}}{\sim} Z_1 Z_2$, for $1 \leq i \leq n$. The following tail bounds are known [58, 19]:

$$\mathbb{P}\Big(\Big|n^{-1}\sum_{i=1}^n X_i\Big| > t\Big) \;\leq C\exp(-3nt^2/2), \quad \text{as } t \to 0 \tag{3.80}$$

$$\mathbb{P}\big(|n^{-1}\chi^2_n - 1| > t\big) \;\leq 2\exp(-3nt^2/16), \quad 0 \leq t < 1/2, \tag{3.81}$$

where C is some positive constant.

Let W be a $p \times p$ centered Wishart matrix as defined in (3.32). Consider the following linear combination of off-diagonal entries of the first row

$$\sum_{j=2}^{n} a_j W_{1j} = n^{-1} \sum_{i=1}^{n} g_{i1} \sum_{j=2}^{p} g_{ij} a_j$$

Let $\xi^i := \|a\|_2^{-1} \sum_{j=2}^{p} g_{ij} a_j$, where $a = (a_2, \ldots, a_p) \in \mathbb{R}^{p-1}$. Note that $\{\xi^i\}_{i=1}^n$ is a collection of independent standard Gaussian RVs. Moreover, $\{\xi^i\}_{i=1}^n$ is independent of $\{g_{i1}\}_{i=1}^n$. Now, we have

$$\sum_{j=2}^{p} a_j W_{1j} = n^{-1} \|a\|_2 \sum_{i=1}^{n} g_{i1} \xi^i,$$

which is a (scaled) sum of Gaussian products (as defined above). Using (3.80), we obtain

$$\mathbb{P}\Big(\Big| \sum_{j=2}^{p} a_j W_{1j} \Big| > t \Big) \leq C \exp\big(-3nt^2/2\|a\|_2^2 \big) \tag{3.82}$$

Combining the bounds in (3.82) and (3.81), we can bound a full linear combination of first-row entries. More specifically, let $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, with $x_1 \neq 0$ and $\sum_{j=2}^{p} x_j \neq 0$, and consider the linear combination $\sum_{j=1}^{p} x_j W_{1j}$. Noting that $W_{11} = n^{-1} \sum_i (g_{i1})^2 - 1$ is a centered χ_n^2, we obtain

$$\begin{aligned}
\mathbb{P}\Big[\Big| \sum_{j=1}^{p} x_j W_{1j} \Big| > t \Big] &\leq \mathbb{P}\Big(|x_1 W_{11}| + \Big| \sum_{j=2}^{p} x_j W_{1j} \Big| > t \Big) \\
&\leq \mathbb{P}\big[|x_1 W_{11}| > t/2 \big] + \mathbb{P}\Big[\Big| \sum_{j=2}^{p} x_j W_{1j} \Big| > t/2 \Big] \\
&\leq 2 \exp\left(-\frac{3nt^2}{16 \cdot 4 x_1^2} \right) + C \exp\Big(-\frac{3nt^2}{2 \cdot 4 \sum_{j=2}^{p} x_j^2} \Big) \\
&\leq 2 \max\{2, C\} \exp\Big(-\frac{3nt^2}{16 \cdot 4 \sum_{j=1}^{p} x_j^2} \Big).
\end{aligned}$$

Note that the last inequality holds, in general, for $x \neq 0$. Since there is nothing special about the "first" row, we can conclude the following.

Lemma 21. *For $t > 0$ small enough, there are (numerical constants) $c > 0$ and $C > 0$ such that for all $x \in \mathbb{R}^p \backslash \{0\}$,*

$$\mathbb{P}\Big(\Big| \sum_{j=1}^{p} x_j W_{ij} \Big| > t \Big) \leq C \exp\big(-c\, nt^2/\|x\|_2^2 \big), \tag{3.83}$$

for $1 \leq i \leq p$.

Now, let $I, J \subset \{1, \dots, p\}$ be index sets[5], both allowed to depend on p (though we have omitted the dependence for brevity). Choose x such that $x_j = 0$ for $j \notin J$ and $\|x_J\|_2 = 1$. Note that $\|W_{I,J} x_J\|_\infty = \max_{i \in I} |\sum_{j \in J} W_{ij} x_j| = \max_{i \in I} |\sum_{j=1}^p W_{ij} x_j|$, suggesting the following lemma

Lemma 22. *Consider some index set I such that $|I| \to \infty$ and $n^{-1} \log |I| \to 0$ as $n, p \to \infty$ and some $x_J \in S^{|J|-1}$. Then, there exists an absolute constant $B > 0$ such that*

$$\|W_{I,J} x_J\|_\infty \le B \sqrt{\frac{\log |I|}{n}} \tag{3.84}$$

as $n, p \to \infty$, with probability 1.

Proof. Applying the union bound in conjunction with the bound (3.83) yields

$$\mathbb{P}\Big(\max_{i \in I} \Big| \sum_{j \in J} W_{ij} x_j \Big| > t \Big) \le |I|\, C \exp(-cnt^2). \tag{3.85}$$

Letting $t = B\sqrt{n^{-1} \log |I|}$, the right-hand side simplifies to $C \exp\left(-(cB^2 - 1) \log |I|\right)$. Taking $B > \sqrt{2c^{-1}}$ and applying Borel–Cantelli lemma completes the proof. □

Note that as a corollary, setting $x_J = (1, 0, \dots, 0)$ yields bounds on the ∞-norm of columns (or, equivalently, rows) of Wishart matrices.

Lemma 22 may be used to obtain the desired bound on $|||W_{I,J}|||_{\infty,2}$. For simplicity, let $y \in \mathbb{R}^{|J|}$ represent a generic $|J|$-vector. Recall that $|||W_{I,J}|||_{\infty,2} = \max_{y \in S^{|J|-1}} \|W_{I,J} y\|_\infty$. We use a standard discretization argument, covering the unit ℓ^2-ball of $\mathbb{R}^{|J|}$ using an ε-net, say $\mathcal{N}$. It can be shown [68] that there exists such a net with cardinality $|\mathcal{N}| < (3/\varepsilon)^{|J|}$. For every $y \in S^{|J|-1}$, let $u_y \in \mathcal{N}$ be the point such that $\|y - u_y\|_2 \le \varepsilon$. Then

$$\|W_{I,J} y\|_\infty \le |||W_{I,J}|||_{\infty,2} \|y - u_y\|_2 + \|W_{I,J} u_y\|_\infty \le |||W_{I,J}|||_{\infty,2}\, \varepsilon + \|W_{I,J} u_y\|_\infty.$$

Taking the maximum over $y \in S^{|J|-1}$ and rearranging yields the inequality

$$|||W_{I,J}|||_{\infty,2} \le (1 - \varepsilon)^{-1} \max_{u \in \mathcal{N}} \|W_{I,J} u\|_\infty. \tag{3.86}$$

Using this bound (3.86), we can now provide the proof of Lemma 11 as follows. Let $\mathcal{N} = \{u_1, \dots, u_{|\mathcal{N}|}\}$ be a $\frac{1}{2}$-net of the ball $S^{|J|-1}$, with cardinality $|\mathcal{N}| < 6^{|J|}$. Then from our bound (3.86), we have

$$\begin{aligned} \mathbb{P}\big(|||W_{I,J}|||_{\infty,2} > t\big) &\le \mathbb{P}\big(2 \max_{u \in \mathcal{N}} \|W_{I,J} u\|_\infty > t\big) \\ &\le |\mathcal{N}| \cdot \mathbb{P}\big(\|W_{I,J} u_1\|_\infty > t/2\big) \\ &\le 6^{|J|} \cdot C|I| \exp(-c\, nt^2/4). \end{aligned}$$

[5]We always assume that these index sets form an increasing sequence of sets. More precisely, with $I = I_p$, we assume $I_1 \subset I_2 \subset \cdots$. We also assume $|I_p| \to \infty$ as $p \to \infty$.

In the last line, we used (3.85). Taking $t = D''\frac{\sqrt{|J|}+\sqrt{\log|I|}}{\sqrt{n}}$ with D'' large enough and using Borel-Cantelli lemma completes the proof.

Appendix 3.E Proof of Lemma 16

The mixture covariance can be expressed as

$$\begin{aligned}\Sigma_M := \mathbb{E}[xx^T] &= \mathbb{E}\left[\mathbb{E}\left[xx^T \,|\, U\right]\right] \\ &= \sum_{S\in\widetilde{\mathbb{S}}} \frac{1}{|\widetilde{\mathbb{S}}|}\,\mathbb{E}\left[xx^T \,|\, U = S\right] \\ &= \sum_{S\in\widetilde{\mathbb{S}}} \frac{1}{|\widetilde{\mathbb{S}}|}\left(I_p + \beta z^*(S)z^*(S)^T\right) \\ &= I_p + \frac{\beta}{|\widetilde{\mathbb{S}}|}\sum_{S\in\widetilde{\mathbb{S}}} z^*(S)z^*(S)^T =: I_p + \frac{\beta}{k\,|\widetilde{\mathbb{S}}|}\,Y,\end{aligned}$$

where

$$Y_{ij} = \sum_{S\in\widetilde{\mathbb{S}}}[\sqrt{k}\,z^*(S)]_i[\sqrt{k}\,z^*(S)^T]_j = \sum_{S\in\widetilde{\mathbb{S}}}\mathbb{I}\{i\in S\}\mathbb{I}\{j\in S\} = \sum_{S\in\widetilde{\mathbb{S}}}\mathbb{I}\{\{i,j\}\subset S\}.$$

Let $R := \{1,\ldots,k-1\}$ and $R^c := \{k,\ldots,p\}$. Note that we always have $R \subset S$ for $S \in \widetilde{\mathbb{S}}$. In general, we have

$$Y_{ij} = \begin{cases} |\widetilde{\mathbb{S}}|, & \text{if both } i,j \in R, \\ 1, & \text{if exactly one of } i \text{ or } j \in R, \\ 0, & \text{if both } i,j \notin R. \end{cases}$$

Consequently, Y takes the form

$$Y = \left(\begin{array}{ccc|cccc} |\widetilde{\mathbb{S}}| & \ldots & |\widetilde{\mathbb{S}}| & 1 & 1 & \ldots & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ |\widetilde{\mathbb{S}}| & \ldots & |\widetilde{\mathbb{S}}| & 1 & 1 & \ldots & 1 \\ \hline 1 & \ldots & 1 & 1 & 0 & \ldots & 0 \\ 1 & \ldots & 1 & 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \ldots & 1 & 0 & 0 & \ldots & 1 \end{array}\right) \quad \text{or} \quad Y = \begin{pmatrix} |\widetilde{\mathbb{S}}|\vec{1}_R\vec{1}_R^T & \vec{1}_R\vec{1}_{R^c}^T \\ \vec{1}_{R^c}\vec{1}_R^T & I_{R^c\times R^c} \end{pmatrix}$$

where $\vec{1}_R$, for example, denotes the vector of all ones over the index set R. We conjecture an eigenvector of the form

$$v = \begin{pmatrix} \vec{1}_R \\ b\vec{1}_{R^c} \end{pmatrix}$$

and let us denote the associated eigenvalue as λ. Thus, we assume $Yv = \lambda v$, or, in more detail,

$$\begin{aligned} |\widetilde{\mathbb{S}}||R|\,\vec{1}_R + b\,|R^c|\,\vec{1}_R &= \lambda\,\vec{1}_R, \\ |R|\,\vec{1}_{R^c} + b\,\vec{1}_{R^c} &= \lambda\, b\,\vec{1}_{R^c} \end{aligned}$$

where we have used, for example, $\vec{1}_R^T\vec{1}_R = |R|$. Note that $|R^c| = |\widetilde{\mathbb{S}}| = p - k + 1$. Rewriting in terms of $|\widetilde{\mathbb{S}}|$, we get

$$\begin{aligned} |\widetilde{\mathbb{S}}|\,(|R| + b) &= \lambda, \\ |R| + b &= \lambda\, b \end{aligned}$$

from which we conclude, assuming $\lambda \neq 0$, that $b = \frac{1}{|\widetilde{\mathbb{S}}|}$. This, in turn, implies $\lambda = |\widetilde{\mathbb{S}}|\,|R| + 1$.

Thus far, we have determined an eigenpair. We can now subtract $\lambda\,(v/\|v\|_2)(v/\|v\|_2)^T = (\lambda/\|v\|_2^2)\,vv^T$ and search for the rest of the eigenvalues in the remainder. Note that

$$\frac{\lambda}{\|v\|_2^2} = \frac{\lambda}{|R| + b^2\,|R^c|} = \frac{|\widetilde{\mathbb{S}}|\,|R| + 1}{|R| + |\widetilde{\mathbb{S}}|^{-1}} = |\widetilde{\mathbb{S}}|.$$

Thus, we have

$$\frac{\lambda}{\|v\|_2^2}\,vv^T = \begin{pmatrix} |\widetilde{\mathbb{S}}|\,\vec{1}_R\vec{1}_R^T & \vec{1}_R\vec{1}_{R^c}^T \\ \vec{1}_{R^c}^T\vec{1}_R & \frac{1}{|\widetilde{\mathbb{S}}|}\,\vec{1}_{R^c}\vec{1}_{R^c}^T \end{pmatrix}$$

implying

$$Y - \frac{\lambda}{\|v\|_2^2}\,vv^T = \begin{pmatrix} 0 & 0 \\ 0 & I - \frac{1}{|\widetilde{\mathbb{S}}|}\,\vec{1}_{R^c}\vec{1}_{R^c}^T \end{pmatrix}.$$

The nonzero block of the remainder has one eigenvalue equal to $1 - \frac{|R^c|}{|\widetilde{\mathbb{S}}|} = 0$ and the rest of $|R^c| - 1$ of its eigenvalues equal to 1. Thus, the remainder has $|R| + 1$ of its eigenvalues equal to zero and $|R^c| - 1$ of them equal to one.

Overall, we conclude that eigenvalues of Y are as follows:

$$\begin{cases} |\widetilde{\mathbb{S}}||R| + 1, & 1 \text{ time}, \\ 1, & |R^c| - 1 \text{ times}, \\ 0, & |R| \text{ times}, \end{cases} \quad \text{or} \quad \begin{cases} (p-k+1)(k-1) + 1, & 1 \text{ time}, \\ 1, & p - k \text{ times}, \\ 0, & k - 1 \text{ times}. \end{cases}$$

The eigenvalues of Y are mapped to those of Σ_M by the affine map $x \to 1 + \frac{\beta}{k\,|\widetilde{\mathbb{S}}|}\,x$, so that Σ_M has eigenvalues

$$1 + \frac{\beta(k-1)}{k} + \frac{\beta}{k(p-k+1)}, \quad 1 + \frac{\beta}{k(p-k+1)}, \quad 1 \tag{3.87}$$

with multiplicities 1, $p-k$ and $k-1$, respectively. The log determinant stated in the lemma then follows by straightforward calculation.

Appendix 3.F Proof of Theorem 7(a)

Since in part (a) of the theorem we are using the weaker scaling $n > \theta_{\mathrm{wr}} k^2 \log(p-k)$, we have more freedom in choosing the sign matrix $\widehat{U}$. We choose the upper-left block $\widehat{U}_{SS}$ as in part (b) so that Lemma 8 applies. Also let $\widehat{z} := (\widehat{z}_S, \vec{0}_{S^c})$ as in (3.24), where $\widehat{z}_S$ is the (unique) maximal eigenvector of the $k \times k$ block Φ_{SS}; it has the correct sign by Lemma 8. We set the off-diagonal and lower-right blocks of the sign matrix to

$$\widehat{U}_{S^cS} = \frac{1}{\lambda_n}\Delta_{S^cS}, \quad \widehat{U}_{S^cS^c} = \frac{1}{\lambda_n}\Delta_{S^cS^c} \tag{3.88}$$

so that $\Phi_{S^cS} = 0$ and $\Phi_{S^cS^c} = 0$. With these blocks of Φ being zero, $\widehat{z}$ is the maximal eigenvector of Φ, hence an optimal solution of (3.9), if and only if $\widehat{z}_S$ is the maximal eigenvector of Φ_{SS}; the latter is true by definition. Note that this argument is based on the remark following Lemma 7. It only remains to show that the choices of (3.88) lead to valid sign matrices.

Recalling that vector ∞-norm of a matrix A is $\|A\|_\infty := \max_{i,j} |A_{i,j}|$ (see § 2.1.3), we need to show $\|\widehat{U}_{S^cS}\|_\infty \leq 1$ and $\|\widehat{U}_{S^cS^c}\|_\infty \leq 1$. Using the notation of section 3.4.4 and the mixed-norm inequality (2.14), we have

$$\begin{aligned}
\|\widehat{U}_{S^cS}\|_\infty = \frac{\sqrt{\beta}}{\lambda_n}\|\tilde{h}_{S^c} z_S^{*T}\|_\infty &\leq \frac{\sqrt{\beta}}{\lambda_n}|\!|\!|\tilde{h}_{S^c}|\!|\!|_{\infty,\infty}\|z_S^{*T}\|_\infty \\
&= \frac{\sqrt{\beta}}{\lambda_n}\|\tilde{h}_{S^c}\|_\infty\|z_S^*\|_\infty \\
&\leq \frac{\sqrt{\beta}}{\lambda_n}|\!|\!|\Gamma_{p-k}^{1/2}|\!|\!|_{\infty,\infty}\|h_{S^c}\|_\infty\|z_S^*\|_\infty \\
&= \frac{2k}{\sqrt{\beta}}\,\mathcal{O}(1)\,\mathcal{O}\left(\sqrt{\frac{\log(p-k)}{n}}\right)\frac{1}{\sqrt{k}} = \mathcal{O}(1)\frac{1}{\sqrt{k}} \to 0,
\end{aligned}$$

where the last line follows under the scaling assumed and assumption (3.2a) on $|\!|\!|\Gamma_{p-k}^{1/2}|\!|\!|_{\infty,\infty}$. For the lower-right block, we use the mixed-norm inequality (2.14) twice together with symmetry to obtain

$$\begin{aligned}
\|\widehat{U}_{S^cS^c}\|_\infty = \frac{1}{\lambda_n}\|\widetilde{W}_{S^cS^c}\|_\infty &= \frac{1}{\lambda_n}\|\Gamma_{p-k}^{1/2} W_{S^cS^c}\Gamma_{p-k}^{1/2}\|_\infty \\
&\leq \frac{1}{\lambda_n}|\!|\!|\Gamma_{p-k}^{1/2}|\!|\!|_{\infty,\infty}^2\|W_{S^cS^c}\|_\infty \\
&= \frac{2k}{\beta}\,\mathcal{O}(1)\,\mathcal{O}\left(\sqrt{\frac{\log(p-k)}{n}}\right)
\end{aligned}$$

which can be made less than one by choosing θ_{wr} large enough. The bound on $\|W_{S^cS^c}\|_\infty$ used in the last line can be obtained using arguments similar to those of Lemma 11. The proof is complete.

Appendix 3.G Proofs of §3.6

We start with an inequality relating ℓ_1, ℓ_∞ norms and ℓ_q F-norm for $q \in (0,1)$; recall that for $q \in (0,1)$, we let $\|z\|_q := \sum_j |z_j|^q$. We have the following

$$\|z\|_1 \le \|z\|_\infty^{1-q} \|z\|_q, \quad q \in (0,1) \tag{3.89}$$

for $z \in \mathbb{R}^p$. For the proof, first assume $\|z\|_\infty \le 1$, that is, $|z_j| \le 1$ for $j \in [p]$. For any $q \in (0,1)$, the function $x \mapsto x^{1-q}$ is increasing on $[0,1]$. That is, for $x \in [0,1]$, we have $x^{1-q} \le 1$ or equivalently $x \le x^q$. In particular, $\sum_{j=1}^p |z_j| \le \sum_{j=1}^p |z_j|^q$ which is the desired inequality. The general form is obtained by applying the inequality to $z/\|z\|_\infty$.

The following lemma collects some observations regarding ℓ_1 norm of z^* and its subvectors and some relations between k, τ and κ.

Lemma 23. *Assume $\|z^*\|_q \le \kappa$ and $\|z^*\|_2 = 1$. Let $S := \{i : |z_j^*| \ge \tau\}$ and $\kappa := \mathrm{card}(S)$. Then,*

$$\|z^*\|_1 \le \kappa, \quad \|z^*_{S^c}\|_1 \le \tau^{1-q}\kappa, \quad k \le \tau^{-q}\kappa \tag{3.90}$$

Proof. We have $\|z^*\|_\infty \le \|z^*\|_2 = 1$ and $\|z^*_{S^c}\|_\infty < \tau$. The first two assertions in (3.90) now follow from (3.89). For the third assertion, we note that since $x \mapsto x^q$ is increasing for $q > 0$, we have

$$k = \sum_{j \in S} 1 \le \sum_{j \in S} \Big(\frac{|z_j^*|}{\tau}\Big)^q \le \tau^{-q}\kappa. \tag{3.91}$$

□

We can now give a bound on $\|Z^*_{S^c}\|_1$. For simplicity, let $a = \|z^*_S\|_1$ and $b = \|z^*_{S^c}\|_1$. Then,

$$\|Z^*_{S^c}\|_1 = \sum_{(i,j) \notin S \times S} |z_i^* z_j^*| = b^2 + 2ab \le 2b(a+b) \le 2\tau^{1-q}\kappa^2 \tag{3.92}$$

by (3.90) and $a + b = \|z^*\|_1$.

3.G.1 Proof of Lemma 19

We have $\|h(z^*)^T\|_\infty = \max_{i,j} |h_i||z_j^*| \le \|h\|_\infty \|z^*\|_\infty \le \|h\|_\infty$. Both $\|h\|_\infty$ and $\|n^{-1}G^TG - I_p\|_\infty$ are bounded by a constant multiple of $\sqrt{\frac{\log p}{n}}$ with stated probability.

3.G.2 Proof of Lemma 20

By triangle inequality, $\|Z_S^* + E_S\|_1 \geq \|Z_S^*\|_1 - \|E_S\|_1$. We have

$$\begin{aligned} T_1 &\geq -\|\widetilde{\Delta}_S\|_\infty \|E_S\|_1 - \lambda_n \|E_S\|_1 \\ &\geq -2\lambda_n \|E_S\|_1 \\ &\geq -2\lambda_n k \|E_S\|_2 \end{aligned}$$

where the first line follows from Hölder inequality (2.4) in addition to triangle, the second line from assumption $\|\widetilde{\Delta}\|_\infty < \lambda_n$ and the last line from (2.5) . Similarly, by triangle inequality, $\|Z_{S^c}^* + E_{S^c}\|_1 \geq \|E_{S^c}\|_1 - \|Z_{S^c}^*\|_1$. Then,

$$\begin{aligned} T_2 &\geq -\|\widetilde{\Delta}_{S^c}\|_\infty \|E_{S^c}\|_1 + \lambda_n \|E_{S^c}\|_1 - 2\lambda_n \|Z_{S^c}^*\|_1 \\ &\geq -2\lambda_n \|Z_{S^c}^*\|_1 \\ &\geq -2\lambda_n (2\tau^{1-q}\kappa^2) \end{aligned}$$

where the last line follows from (3.92).

Chapter 4

Approximation properties of certain operator-induced norms on Hilbert spaces

4.1 Introduction

This chapter serves as an interlude to Chapter 5 where we study effects of sampling in functional PCA. The focus here is on a class of approximation-theoretic issues that arise frequently in the analysis of functional estimators in statistics and statistical learning theory. In particular, we will see how the results established here will assist us in determining the functional rate of convergence for the estimators of Chapter 5 .

To set the stage, let $\mathbb{P}$ be a probability measure supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$ and consider the function class

$$L^2(\mathbb{P}) := \big\{ f : \mathcal{X} \to \mathbb{R} \;\big|\; \|f\|_{L^2(\mathbb{P})} < \infty \big\}, \tag{4.1}$$

where $\|f\|_{L^2(\mathbb{P})} := \sqrt{\int_{\mathcal{X}} f^2(x)\, d\mathbb{P}(x)}$ is the usual L^2 norm[1] defined with respect to the measure $\mathbb{P}$. It is often of interest to construct approximations to this L^2 norm that are "finite-dimensional" in nature, and to study the quality of approximation over the unit ball of some Hilbert space $\mathcal{H}$ that is continuously embedded within L^2. For example, in approximation theory and mathematical statistics, a collection of n design points in $\mathcal{X}$ is often used to define a surrogate for the L^2 norm. In other settings, one is given some orthonormal basis of $L^2(\mathbb{P})$, and defines an approximation based on the sum of squares of the first n (generalized) Fourier coefficients. For problems of this type, it is of interest to gain a precise understanding of the approximation accuracy in terms of its dimension n and other problem parameters.

[1] We also use $L^2(\mathcal{X})$ or simply L^2 to refer to the space (4.1), with corresponding conventions for its norm. Also, one can take $\mathcal{X}$ to be a compact subset of any separable metric space and $\mathbb{P}$ a (regular) Borel measure.

The goal in this chapter is to study such questions in reasonable generality for the case of Hilbert spaces $\mathcal{H}$. We let $\Phi_n : \mathcal{H} \to \mathbb{R}^n$ denote a continuous linear operator on the Hilbert space, which acts by mapping any $f \in \mathcal{H}$ to the n-vector $\begin{pmatrix}[\Phi_n f]_1 & [\Phi_n f]_2 & \cdots & [\Phi_n f]_n\end{pmatrix}$. This operator defines the Φ_n-semi-norm

$$\|f\|_{\Phi_n} := \sqrt{\sum_{i=1}^{n} [\Phi_n f]_i^2}. \tag{4.2}$$

In the sequel, with a minor abuse of terminology,[2] we refer to $\|f\|_{\Phi_n}$ as the Φ_n-norm of f. Our goal is to study how well $\|f\|_{\Phi_n}$ approximates $\|f\|_{L^2}$ over the unit ball of $\mathcal{H}$ as a function of n, and other problem parameters. We provide a number of examples of the *sampling operator* Φ_n in §4.2.2. Since the dependence on the parameter n should be clear, we frequently omit the subscript to simplify notation.

In order to measure the quality of approximation over $\mathcal{H}$, we consider the quantity

$$R_\Phi(\varepsilon) := \sup \left\{ \|f\|_{L^2}^2 \mid f \in B_{\mathcal{H}}, \ \|f\|_\Phi^2 \leq \varepsilon^2 \right\}, \tag{4.3}$$

where $B_{\mathcal{H}} := \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$ is the unit ball of $\mathcal{H}$. The goal in this chapter is to obtain sharp upper bounds on R_Φ. As discussed in Appendix 4.C, a relatively straightforward argument can be used to translate such upper bounds into lower bounds on the related quantity

$$\underline{T}_\Phi(\varepsilon) := \inf \left\{ \|f\|_\Phi^2 \mid f \in B_{\mathcal{H}}, \ \|f\|_{L^2}^2 \geq \varepsilon^2 \right\}. \tag{4.4}$$

We also note that, for a complete picture of the relationship between the semi-norm $\|\cdot\|_\Phi$ and the L^2 norm, one can also consider the related pair

$$T_\Phi(\varepsilon) := \sup \left\{ \|f\|_\Phi^2 \mid f \in B_{\mathcal{H}}, \ \|f\|_{L^2}^2 \leq \varepsilon^2 \right\}, \quad \text{and} \tag{4.5a}$$

$$\underline{R}_\Phi(\varepsilon) := \inf \left\{ \|f\|_{L^2}^2 \mid f \in B_{\mathcal{H}}, \ \|f\|_\Phi^2 \geq \varepsilon^2 \right\}. \tag{4.5b}$$

Our methods are also applicable to these quantities, but we limit our treatment to $(R_\Phi, \underline{T}_\Phi)$ so as to keep the contribution focused.

Certain special cases of linear operators Φ, and associated functionals have been studied in past work. In the special case $\varepsilon = 0$, we have

$$R_\Phi(0) = \sup \left\{ \|f\|_{L^2}^2 \mid f \in B_{\mathcal{H}}, \ \Phi(f) = 0 \right\},$$

a quantity that corresponds to the squared diameter of $B_{\mathcal{H}} \cap \operatorname{Ker}(\Phi)$, measured in the L^2-norm. Quantities of this type are standard in approximation theory (e.g., [35, 78, 79]), for instance in the context of Kolmogorov and Gelfand widths. Our primary interest in this

[2] This can be justified by identifying f and g if $\Phi f = \Phi g$, i.e. considering the quotient $\mathcal{H}/\operatorname{Ker}\Phi$.

chapter is the more general setting with $\varepsilon > 0$, for which additional factors are involved in controlling $R_\Phi(\varepsilon)$. In statistics, there is a literature on the case in which Φ is a sampling operator, which maps each function f to a vector of n samples, and the norm $\|\cdot\|_\Phi$ corresponds to the empirical L^2-norm defined by these samples. When these samples are chosen randomly, then techniques from empirical process theory [93] can be used to relate the two terms. As discussed in the sequel, our results have consequences for this setting of random sampling.

As an example of a problem in which an upper bound on R_Φ is useful, let us consider a general linear inverse problem, in which the goal is to recover an estimate of the function f^* based on the noisy observations

$$y_i = [\Phi f^*]_i + w_i, \quad i = 1, \ldots, n,$$

where $\{w_i\}$ are zero-mean noise variables, and $f^* \in B_{\mathcal{H}}$ is unknown. An estimate $\widehat{f}$ can be obtained by solving a least-squares problem over the unit ball of the Hilbert space—that is, to solve the convex program

$$\widehat{f} := \arg\min_{f \in B_{\mathcal{H}}} \sum_{i=1}^{n} (y_i - [\Phi f]_i)^2.$$

For such estimators, there are fairly standard techniques for deriving upper bounds on the Φ-semi-norm of the deviation $\widehat{f} - f^*$. Our results in this chapter on R_Φ can then be used to translate this to a corresponding upper bound on the L^2-norm of the deviation $\widehat{f} - f^*$, which is often a more natural measure of performance.

As an example where the dual quantity $\underline{T}_\Phi$ might be helpful, consider the packing problem for a subset $\mathcal{D} \subset B_{\mathcal{H}}$ of the Hilbert ball. Let $M(\varepsilon; \mathcal{D}, \|\cdot\|_{L^2})$ be the ε-packing number of $\mathcal{D}$ in $\|\cdot\|_{L^2}$, i.e., the maximal number of function $f_1, \ldots, f_M \in \mathcal{D}$ such that $\|f_i - f_j\|_{L^2} \geq \varepsilon$ for all $i, j = 1, \ldots, M$. Similarly, let $M(\varepsilon; \mathcal{D}, \|\cdot\|_\Phi)$ be the ε-packing number of $\mathcal{D}$ in $\|\cdot\|_\Phi$ norm. Now, suppose that for some fixed ε, $\underline{T}_\Phi(\varepsilon) > 0$. Then, if we have a collection of functions $\{f_1, \ldots, f_M\}$ which is an ε-packing of $\mathcal{D}$ in $\|\cdot\|_{L^2}$ norm, then the same collection will be a $\sqrt{\underline{T}_\Phi(\varepsilon)}$-packing of $\mathcal{D}$ in $\|\cdot\|_\Phi$. This implies the following useful relationship between packing numbers

$$M(\varepsilon\,; \mathcal{D}, \|\cdot\|_{L^2}) \leq M(\sqrt{\underline{T}_\Phi(\varepsilon)}\,; \mathcal{D}, \|\cdot\|_\Phi).$$

The remainder of this chapter is organized as follows. We begin in §4.2 with background on the Hilbert space set-up, and provide various examples of the linear operators Φ to which our results apply. §4.3 contains the statement of our main result, and illustration of some its consequences for different Hilbert spaces and linear operators. Finally, §4.4 is devoted to the proofs of our results.

4.1.1 Notation

For the convenience of the reader, we review some notations used in this chapter. More details can be found in Chapter 2. For any positive integer p, we use $\mathbb{S}_+^p$ to denote the

cone of $p \times p$ positive semidefinite matrices. For $A, B \in \mathbb{S}_+^p$, we write $A \succeq B$ or $B \preceq A$ to mean $A - B \in \mathbb{S}_+^p$. For any square matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its minimal and maximal eigenvalues, respectively. We will use both $\sqrt{A}$ and $A^{1/2}$ to denote the symmetric square root of $A \in \mathbb{S}_+^p$. We will use $\{x_k\} = \{x_k\}_{k=1}^\infty$ to denote a (countable) sequence of objects (e.g. real-numbers and functions). Occasionally we might denote an n-vector as $\{x_1, \ldots, x_n\}$. The context will determine whether the elements between braces are ordered. The symbols $\ell_2 = \ell_2(\mathbb{N})$ are used to denote the Hilbert sequence space consisting of real-valued sequences equipped with the inner product $\langle \{x_k\}, \{y_k\} \rangle_{\ell_2} := \sum_{k=1}^\infty x_i y_i$. The corresponding norm is denoted as $\|\cdot\|_{\ell_2}$.

4.2 Background and setup

We begin with some background on the class of Hilbert spaces of interest in this paper and then proceed to provide some examples of the sampling operators of interest. For a general review of the functional-analytic concepts used here, one can refer to §2.5.

4.2.1 Hilbert spaces

We consider a class of Hilbert function spaces contained within $L^2(\mathcal{X})$, and defined as follows. Let $\{\psi_k\}_{k=1}^\infty$ be an orthonormal sequence (not necessarily a basis) in $L^2(\mathcal{X})$ and let $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots > 0$ be a sequence of positive weights decreasing to zero. Given these two ingredients, we can consider the class of functions

$$\mathcal{H} := \left\{ f \in L^2(\mathbb{P}) \,\middle|\, f = \sum_{k=1}^\infty \sqrt{\sigma_k} \alpha_k \psi_k, \quad \text{for some } \{\alpha_k\}_{k=1}^\infty \in \ell_2(\mathbb{N}) \right\}, \tag{4.6}$$

where the series in (4.6) is assumed to converge in L^2. (The series converges since $\sum_{k=1}^\infty (\sqrt{\sigma_k}\alpha_k)^2 \leq \sigma_1 \|\{\alpha_k\}\|_{\ell_2} < \infty$.) We refer to the sequence $\{\alpha_k\}_{k=1}^\infty \in \ell_2$ as the representative of f. Note that this representation is unique due to σ_k being strictly positive for all $k \in \mathbb{N}$.

If f and g are two members of $\mathcal{H}$, say with associated representatives $\alpha = \{\alpha_k\}_{k=1}^\infty$ and $\beta = \{\beta_k\}_{k=1}^\infty$, then we can define the inner product

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{k=1}^\infty \alpha_k \beta_k \;=\; \langle \alpha, \beta \rangle_{\ell_2}. \tag{4.7}$$

With this choice of inner product, it can be verified that the space $\mathcal{H}$ is a Hilbert space. (In fact, $\mathcal{H}$ inherits all the required properties directly from ℓ_2.) For future reference, we note that for two functions $f, g \in \mathcal{H}$ with associated representatives $\alpha, \beta \in \ell_2$, their L^2-based inner product is given by[3] $\langle f, g \rangle_{L^2} = \sum_{k=1}^\infty \sigma_k \alpha_k \beta_k$.

[3] In particular, for $f \in \mathcal{H}$, $\|f\|_{L^2} \leq \sqrt{\sigma_1} \|f\|_{\mathcal{H}}$ which shows that the inclusion $\mathcal{H} \subset L^2$ is continuous.

We note that each ψ_k is in $\mathcal{H}$, as it is represented by a sequence with a single nonzero element, namely, the k-th element which is equal to $\sigma_k^{-1/2}$. It follows from (4.7) that $\langle\sqrt{\sigma_k}\psi_k, \sqrt{\sigma_j}\psi_j\rangle_{\mathcal{H}} = \delta_{kj}$. That is, $\{\sqrt{\sigma_k}\psi_k\}$ is an orthonormal sequence in $\mathcal{H}$. Now, let $f \in \mathcal{H}$ be represented by $\alpha \in \ell_2$. We claim that the series in (4.6) also converges in $\mathcal{H}$ norm. In particular, $\sum_{k=1}^{N}\sqrt{\sigma_k}\alpha_k\psi_k$ is in $\mathcal{H}$, as it is represented by the sequence $\{\alpha_1, \ldots, \alpha_N, 0, 0, \ldots\} \in \ell_2$. It follows from (4.7) that $\|f - \sum_{k=1}^{N}\sqrt{\sigma_k}\alpha_k\psi_k\|_{\mathcal{H}} = \sum_{k=N+1}^{\infty}\alpha_k^2$ which converges to 0 as $N \to \infty$. Thus, $\{\sqrt{\sigma_k}\psi_k\}$ is in fact an orthonormal basis for $\mathcal{H}$.

We now turn to a special case of particular importance to us, namely the reproducing kernel Hilbert space (RKHS) of a continuous kernel. Consider a symmetric bivariate function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact[4]. Furthermore, assume $\mathbb{K}$ to be positive semidefinite and continuous. Consider the integral operator $I_{\mathbb{K}}$ mapping a function $f \in L^2$ to the function $I_{\mathbb{K}}f := \int \mathbb{K}(\cdot, y)f(y)d\mathbb{P}(y)$. As a consequence of Mercer's theorem [83, 42], $I_{\mathbb{K}}$ is a compact operator from L^2 to $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$ equipped with the uniform norm[5]. Let $\{\sigma_k\}$ be the sequence of nonzero eigenvalues of $I_{\mathbb{K}}$, which are positive, can be ordered in nonincreasing order and converge to zero. Let $\{\psi_k\}$ be the corresponding eigenfunctions which are continuous and can be taken to be orthonormal in L^2. With these ingredients, the space $\mathcal{H}$ defined in equation (4.6) is the RKHS of the kernel function $\mathbb{K}$. This can be verified as follows.

As another consequence of the Mercer's theorem, $\mathbb{K}$ has the decomposition

$$\mathbb{K}(x, y) := \sum_{k=1}^{\infty}\sigma_k\psi_k(x)\psi_k(y) \tag{4.8}$$

where the convergence is absolute and uniform (in x and y). In particular, for any fixed $y \in \mathcal{X}$, the sequence $\{\sqrt{\sigma_k}\psi_k(y)\}$ is in ℓ_2. (In fact, $\sum_{k=1}^{\infty}(\sqrt{\sigma_k}\psi_k(y))^2 = \mathbb{K}(y, y) < \infty$.) Hence, $\mathbb{K}(\cdot, y)$ is in $\mathcal{H}$, as defined in (4.6), with representative $\{\sqrt{\sigma_k}\,\psi_k(y)\}$. Furthermore, it can be verified that the convergence in (4.6) can be taken to be also pointwise[6]. To be more specific, for any $f \in \mathcal{H}$ with representative $\{\alpha_k\}_{k=1}^{\infty} \in \ell_2$, we have $f(y) = \sum_{k=1}^{\infty}\sqrt{\sigma_k}\alpha_k\psi_k(y)$, for all $y \in \mathcal{X}$. Consequently, by definition of the inner product (4.7), we have

$$\langle f, \mathbb{K}(\cdot, y)\rangle_{\mathcal{H}} = \sum_{k=1}^{\infty}\alpha_k\sqrt{\sigma_k}\psi_k(y) \;=\; f(y),$$

so that $\mathbb{K}(\cdot, y)$ acts as the representer of evaluation. This argument shows that for any fixed $y \in \mathcal{X}$, the linear functional on $\mathcal{H}$ given by $f \mapsto f(y)$ is bounded, since we have

$$|f(y)| = \left|\langle f, \mathbb{K}(\cdot, y)\rangle_{\mathcal{H}}\right| \;\le\; \|f\|_{\mathcal{H}}\|\mathbb{K}(\cdot, y)\|_{\mathcal{H}},$$

[4]Also assume that $\mathbb{P}$ assign positive mass to every open Borel subset of $\mathcal{X}$.

[5]In fact, $I_{\mathbb{K}}$ is well defined over $L^1 \supset L^2$ and the conclusions about $I_{\mathbb{K}}$ hold as a operator from L^1 to $C(\mathcal{X})$.

[6]The convergence is actually even stronger, namely it is absolute and uniform, as can be seen by noting that $\sum_{k=n+1}^{m}|\alpha_k\sqrt{\sigma_k}\psi_k(y)| \le (\sum_{k=n+1}^{m}\alpha_k^2)^{1/2}(\sum_{k=n+1}^{m}\sigma_k\psi_k^2(y))^{1/2} \le (\sum_{k=n+1}^{m}\alpha_k^2)^{1/2}\max_{y\in\mathcal{X}}k(y, y)$.

hence $\mathcal{H}$ is indeed the RKHS of the kernel $\mathbb{K}$. This fact plays an important role in the sequel, since some of the linear operators that we consider involve pointwise evaluation.

A comment regarding the scope: our general results hold for the basic setting introduced in equation (4.6). For those examples that involve pointwise evaluation, we assume the more refined case of the RKHS described above.

4.2.2 Linear operators, semi-norms and examples

Let $\Phi : \mathcal{H} \to \mathbb{R}^n$ be a continuous linear operator, with co-ordinates $[\Phi f]_i$ for $i = 1, 2, \ldots, n$. It defines the (semi)-inner product

$$\langle f, g \rangle_\Phi := \langle \Phi f, \Phi g \rangle_{\mathbb{R}^n}, \tag{4.9}$$

which induces the semi-norm $\| \cdot \|_\Phi$. By the Riesz representation theorem, for each $i = 1, \ldots, n$, there is a function $\varphi_i \in \mathcal{H}$ such that $[\Phi f]_i = \langle \varphi_i, f \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$.

Let us illustrate the preceding definitions with some examples.

Example 1 (Generalized Fourier truncation)**.** Recall the orthonormal basis $\{\psi_i\}_{i=1}^\infty$ underlying the Hilbert space. Consider the linear operator $\mathbb{T}_{\psi_1^n} : \mathcal{H} \to \mathbb{R}^n$ with coordinates

$$[\mathbb{T}_{\psi_1^n} f]_i := \langle \psi_i, f \rangle_{L^2}, \quad \text{for } i = 1, 2, \ldots, n. \tag{4.10}$$

We refer to this operator as the *(generalized) Fourier truncation operator,* since it acts by truncating the (generalized) Fourier representation of f to its first n co-ordinates. More precisely, by construction, if $f = \sum_{k=1}^\infty \sqrt{\sigma_k} \alpha_k \psi_k$, then

$$[\Phi f]_i = \sqrt{\sigma_i} \alpha_i, \qquad \text{for } i = 1, 2, \ldots, n. \tag{4.11}$$

By definition of the Hilbert inner product, we have $\alpha_i = \langle \psi_i, f \rangle_{\mathcal{H}}$, so that we can write $[\Phi f]_i = \langle \varphi_i, f \rangle_{\mathcal{H}}$, where $\varphi_i := \sqrt{\sigma_i} \psi_i$. ♢

Example 2 (Domain sampling)**.** A collection $x_1^n := \{x_1, \ldots, x_n\}$ of points in the domain $\mathcal{X}$ can be used to define the (scaled) *sampling operator* $\mathbb{S}_{x_1^n} : \mathcal{H} \to \mathbb{R}^n$ via

$$\mathbb{S}_{x_1^n} f := n^{-1/2} \begin{pmatrix} f(x_1) & \ldots & f(x_n) \end{pmatrix}, \quad \text{for } f \in \mathcal{H}. \tag{4.12}$$

As previously discussed, when $\mathcal{H}$ is a reproducing kernel Hilbert space (with kernel $\mathbb{K}$), the (scaled) evaluation functional $f \mapsto n^{-1/2} f(x_i)$ is bounded, and its Riesz representation is given by the function $\varphi_i = n^{-1/2} \mathbb{K}(\cdot, x_i)$. ♢

Example 3 (Weighted domain sampling)**.** Consider the setting of the previous example. A slight variation on the sampling operator (4.12) is obtained by adding some weights to the samples

$$\mathbb{W}_{x_1^n, w_1^n} f := n^{-1/2} \begin{pmatrix} w_1 f(x_1) & \ldots & w_n f(x_n) \end{pmatrix}, \quad \text{for } f \in \mathcal{H}. \tag{4.13}$$

where $w_1^n = (w_1, \ldots, w_n)$ is chosen such that $\sum_{k=1}^n w_k^2 = 1$. Clearly, $\varphi_i = n^{-1/2} w_i \, \mathbb{K}(\cdot, x_i)$.

[As an example of how this might arise, consider approximating $f(t)$ by $\sum_{k=1}^n f(x_k) G_n(t, x_k)$ where $\{G_n(\cdot\,, x_k)\}$ is a collection of functions in $L^2(\mathcal{X})$ such that $\langle G_n(\cdot\,, x_k), G_n(\cdot\,, x_j)\rangle_{L^2} = n^{-1} w_k^2 \, \delta_{kj}$. Proper choices of $\{G_n(\cdot, x_i)\}$ might produce better approximations to the L^2 norm in the cases where one insists on choosing elements of x_1^n to be uniformly spaced, while $\mathbb{P}$ in (4.1) is not a uniform distribution. Another slightly different but closely related case is when one approximates $f^2(t)$ over $\mathcal{X} = [0, 1]$, by say $n^{-1} \sum_{k=1}^{n-1} f^2(x_k) W(n(t - x_k))$ for some function $W : [-1, 1] \to \mathbb{R}_+$ and $x_k = k/n$. Again, non-uniform weights are obtained when $\mathbb{P}$ is nonuniform.]

$\diamondsuit$

4.3 Main result and some consequences

We now turn to the statement of our main result, and the development of some its consequences for various models.

4.3.1 General upper bounds on $R_\Phi(\varepsilon)$

We now turn to upper bounds on $R_\Phi(\varepsilon)$ which was defined previously in (4.3). Our bounds are stated in terms of a real-valued function defined as follows: for matrices $D, M \in \mathbb{S}_+^p$,

$$\mathcal{L}(t, M, D) := \max\Big\{ \lambda_{\max}\big(D - t\sqrt{D}\, M \sqrt{D}\big),\ 0 \Big\}, \qquad \text{for } t \geq 0. \tag{4.14}$$

Here $\sqrt{D}$ denotes the matrix square root, valid for positive semidefinite matrices.

The upper bounds on $R_\Phi(\varepsilon)$ involve principal submatrices of certain infinite-dimensional matrices—or equivalently linear operators on $\ell_2(\mathbb{N})$—that we define here. Let Ψ be the infinite-dimensional matrix with entries

$$[\Psi]_{jk} := \langle \psi_j, \psi_k \rangle_\Phi, \quad \text{for } j, k = 1, 2, \ldots, \tag{4.15}$$

and let $\Sigma = \operatorname{diag}\{\sigma_1, \sigma_2, \ldots, \}$ be a diagonal operator. For any $p = 1, 2, \ldots$, we use Ψ_p and $\Psi_{\widetilde{p}}$ to denote the principal submatrices of Ψ on rows and columns indexed by $\{1, 2, \ldots, p\}$ and $\{p+1, p+2, \ldots\}$, respectively. A similar notation will be used to denote submatrices of Σ.

Theorem 10. *For all $\varepsilon \geq 0$, we have:*

$$R_\Phi(\varepsilon) \ \leq \ \inf_{p \in \mathbb{N}} \inf_{t \geq 0} \Big\{ \mathcal{L}(t, \Psi_p, \Sigma_p) + t \left(\varepsilon + \sqrt{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2})} \right)^2 + \sigma_{p+1} \Big\}. \tag{4.16}$$

Moreover, for any $p \in \mathbb{N}$ such that $\lambda_{\min}(\Psi_p) > 0$, we have

$$R_\Phi(\varepsilon) \;\leq\; \Big(1 - \frac{\sigma_{p+1}}{\sigma_1}\Big)\, \frac{1}{\lambda_{\min}(\Psi_p)} \Big(\varepsilon + \sqrt{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2})}\Big)^2 + \sigma_{p+1}. \tag{4.17}$$

Remark (a): These bounds cannot be improved in general. This is most easily seen in the special case $\varepsilon = 0$. Setting $p = n$, bound (4.17) implies that $R_\Phi(0) \leq \sigma_{n+1}$ whenever Ψ_n is strictly positive definite and $\Psi_{\widetilde{n}} = 0$. This bound is sharp in a "minimax sense", meaning that equality holds if we take the infimum over all bounded linear operators $\Phi : \mathcal{H} \to \mathbb{R}^n$. In particular, it is straightforward to show that

$$\inf_{\substack{\Phi\colon \mathcal{H}\to\mathbb{R}^n \\ \Phi \text{ surjective}}} R_\Phi(0) \;=\; \inf_{\substack{\Phi\colon \mathcal{H}\to\mathbb{R}^n \\ \Phi \text{ surjective}}} \sup_{f \in B_{\mathcal{H}}} \big\{ \|f\|_{L^2}^2 \;\big|\; \Phi f = 0 \big\} \;=\; \sigma_{n+1}, \tag{4.18}$$

and moreover, this infimum is in fact achieved by some linear operator. Such results are known from the general theory of n-widths for Hilbert spaces (e.g., see Chapter IV in Pinkus [78] and Chapter 3 of [41].)

In the more general setting of $\varepsilon > 0$, there are operators for which the bound (4.17) is met with equality. As a simple illustration, recall the (generalized) Fourier truncation operator $\mathbb{T}_{\psi_1^n}$ from Example 1. First, it can be verified that $\langle \psi_k, \psi_j \rangle_{\mathbb{T}_{\psi_1^n}} = \delta_{jk}$ for $j, k \leq n$ and $\langle \psi_k, \psi_j \rangle_{\mathbb{T}_{\psi_1^n}} = 0$ otherwise. Taking $p = n$, we have $\Psi_n = I_n$, that is, the n-by-n identity matrix, and $\Psi_{\widetilde{n}} = 0$. Taking $p = n$ in (4.17), it follows that for $\varepsilon^2 \leq \sigma_1$,

$$R_{\mathbb{T}_{\psi_1^n}}(\varepsilon) \;\leq\; \Big(1 - \frac{\sigma_{n+1}}{\sigma_1}\Big)\varepsilon^2 + \sigma_{n+1}, \tag{4.19}$$

As shown in Appendix 4.E, the bound (4.19) in fact holds with equality. In other words, the bounds of Theorems 10 are tight in this case. Also, note that (4.19) implies $R_{\mathbb{T}_{\psi_1^n}}(0) \leq \sigma_{n+1}$ showing that the (generalized) Fourier truncation operator achieves the minimax bound of (4.18). Fig 4.1 provides a geometric interpretation of these results.

Remark (b): In general, it might be difficult to obtain a bound on $\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2})$ as it involves the infinite dimensional matrix $\Psi_{\widetilde{p}}$. One may obtain a simple (although not usually sharp) bound on this quantity by noting that for a positive semidefinite matrix, the maximal eigenvalue is bounded by the trace, that is,

$$\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\big) \leq \operatorname{tr}\big(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\big) = \sum_{k>p} \sigma_k [\Psi]_{kk}. \tag{4.20}$$

Another relatively easy-to-handle upper bound is

$$\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\big) \leq \|\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\|\infty = \sup_{k>p} \sum_{r>p} \sqrt{\sigma_k}\sqrt{\sigma_r} \big|[\Psi]_{kr}\big|. \tag{4.21}$$

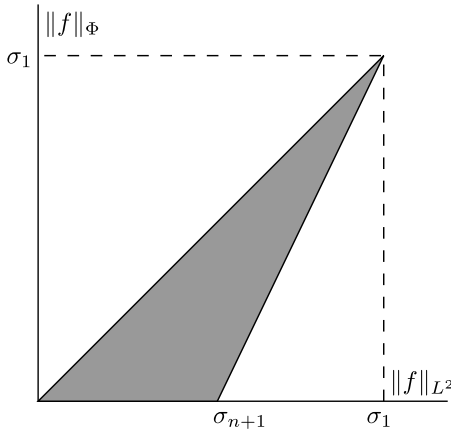


Figure 4.1: Geometry of Fourier truncation. The plot shows the set $\{(\|f\|_{L^2}, \|f\|_\Phi) : \|f\|_{\mathcal{H}} \leq 1\} \subset \mathbb{R}^2$ for the case of (generalized) Fourier truncation operator $\mathbb{T}_{\psi_1^n}$.

These bounds can be used, in combination with appropriate block partitioning of $\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}$, to provide sharp bounds on the maximal eigenvalue. Block partitioning is useful due to the following: for a positive semidefinite matrix $M = \left(\begin{smallmatrix} A_1 & C \\ C^T & A_2 \end{smallmatrix}\right)$, we have $\lambda_{\max}(M) \leq \lambda_{\max}(A_1) + \lambda_{\max}(A_2)$. We leave the the details on the application of these ideas to examples in §4.3.2.

4.3.2 Some illustrative examples

Theorem 10 has a number of concrete consequences for different Hilbert spaces and linear operators, and we illustrate a few of them in the following subsections.

Random domain sampling

We begin by stating a corollary of Theorem 10 in application to random time sampling in a reproducing kernel Hilbert space (RKHS). Recall from equation (4.12) the time sampling operator $\mathbb{S}_{x_1^n}$, and assume that the sample points $\{x_1, \ldots, x_n\}$ are drawn in an i.i.d. manner according to some distribution $\mathbb{P}$ on $\mathcal{X}$. Let us further assume that the eigenfunctions ψ_k, $k \geq 1$ are uniformly bounded[7] on $\mathcal{X}$, meaning that

$$\sup_{k \geq 1} \sup_{x \in \mathcal{X}} |\psi_k(x)| \leq C_\psi. \tag{4.22}$$

Finally, we assume that $\|\sigma\|_1 := \sum_{k=1}^{\infty} \sigma_k < \infty$, and that

$$\sigma_{pk} \leq C_\sigma \, \sigma_k \, \sigma_p, \quad \text{for some positive constant } C_\sigma \text{ and for all large } p, \tag{4.23}$$

$$\textstyle\sum_{k > p^m} \sigma_k \leq \sigma_p, \quad \text{for some positive integer } m \text{ and for all large } p. \tag{4.24}$$

[7] One can replace $\sup_{x \in \mathcal{X}}$ with essential supremum with respect to $\mathbb{P}$.

Let m_σ be the smallest m for which (4.24) holds. These conditions on $\{\sigma_k\}$ are satisfied, for example, for both a polynomial decay $\sigma_k = \mathcal{O}(k^{-\alpha})$ with $\alpha > 1$ and an exponential decay $\sigma_k = \mathcal{O}(\rho^k)$ with $\rho \in (0,1)$. In particular, for the polynomial decay, using the tail bound (4.63) in Appendix 4.B, we can take $m_\sigma = \lceil \frac{\alpha}{\alpha-1} \rceil$ to satisfy (4.24). For the exponential decay, we can take $m_\sigma = 1$ for $\rho \in (0, \frac{1}{2})$ and $m_\sigma = 2$ for $\rho \in (\frac{1}{2}, 1)$ to satisfy (4.24).

Define the function

$$\mathcal{G}_n(\varepsilon) := \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{\infty} \min\{\sigma_j, \varepsilon^2\}}, \tag{4.25}$$

as well as the *critical radius*

$$r_n := \inf\{\varepsilon > 0 \,:\, \mathcal{G}_n(\varepsilon) \le \varepsilon^2\}. \tag{4.26}$$

Corollary 4. *Suppose that* $r_n > 0$ *and* $64\, C_\psi^2\, m_\sigma\, r_n^2 \log(2nr_n^2) \le 1$. *Then for any* $\varepsilon^2 \in [r_n^2, \sigma_1)$, *we have*

$$\mathbb{P}\Big[R_{\mathbb{S}_{x_1^n}}(\varepsilon) > (\widetilde{C}_\psi + \widetilde{C}_\sigma)\,\varepsilon^2\Big] \le 2\exp\Big(-\frac{1}{64\, C_\psi^2\, r_n^2}\Big), \tag{4.27}$$

where $\widetilde{C}_\psi := 2(1 + C_\psi)^2$ *and* $\widetilde{C}_\sigma := 3(1 + C_\psi^{-1}) C_\sigma \|\sigma\|_1 + 1$.

We provide the proof of this corollary in Appendix 4.A. As a concrete example consider a polynomial decay $\sigma_k = \mathcal{O}(k^{-\alpha})$ for $\alpha > 1$, which satisfies assumptions on $\{\sigma_k\}$. Using the tail bound (4.63) in Appendix 4.B, one can verify that $r_n^2 = \mathcal{O}(n^{-\alpha/(\alpha+1)})$. Note that, in this case,

$$r_n^2 \log(2nr_n^2) = \mathcal{O}(n^{-\frac{\alpha}{\alpha+1}} \log n^{\frac{1}{\alpha+1}}) = \mathcal{O}(n^{-\frac{\alpha}{\alpha+1}} \log n) \to 0, \quad n \to \infty.$$

Hence conditions of Corollary 4 are met for sufficiently large n. It follows that for some constants C_1, C_2 and C_3, we have

$$R_{\mathbb{S}_{x_1^n}}(C_1 n^{-\frac{\alpha}{2(\alpha+1)}}) \le C_2\, n^{-\frac{\alpha}{\alpha+1}}$$

with probability $1 - 2\exp(-C_3 n^{\frac{\alpha}{\alpha+1}})$ for sufficiently large n.

Sobolev kernel

Consider the kernel $\mathbb{K}(x,y) = \min(x,y)$ defined on $\mathcal{X}^2$ where $\mathcal{X} = [0,1]$. The corresponding RKHS is of Sobolev type and can be expressed as

$$\big\{f \in L^2(\mathcal{X}) \;\big|\; f \text{ is absolutely continuous, } f(0) = 0 \text{ and } f' \in L^2(\mathcal{X})\big\}.$$

Also consider a uniform domain sampling operator $\mathbb{S}_{x_1^n}$, that is, that of (4.12) with $x_i = i/n, i \le n$ and let $\mathbb{P}$ be uniform (i.e., the Lebesgue measure restricted to $[0,1]$).

This setting has the benefit that many interesting quantities can be computed explicitly, while also having some practical appeal. The following can be shown about the eigen-decomposition of the integral operator $I_{\mathbb{K}}$ introduced in §4.2,

$$\sigma_k = \Big[\frac{(2k-1)\pi}{2}\Big]^{-2}, \quad \psi_k(x) = \sqrt{2}\sin\big(\sigma_k^{-1/2}x\big), \quad k = 1, 2, \dots .$$

In particular, the eigenvalues decay as $\sigma_k = \mathcal{O}(k^{-2})$.

To compute the Ψ, we write

$$[\Psi]_{kr} = \langle \psi_k, \psi_r \rangle_\Phi = \frac{1}{n}\sum_{\ell=1}^{n}\Big\{\cos\frac{(k-r)\ell\pi}{n} - \cos\frac{(k+r-1)\ell\pi}{n}\Big\}. \tag{4.28}$$

We note that Ψ is periodic in k and r with period $2n$. It is easily verified that $n^{-1}\sum_{\ell=1}^{n}\cos(q\ell\pi/n)$ is equal to -1 for odd values of q and zero for even values, other than $q = 0, \pm 2n, \pm 4n, \dots$. It follows that

$$[\Psi]_{kr} = \begin{cases} 1 + \frac{1}{n} & \text{if } k - r = 0, \\ -1 - \frac{1}{n} & \text{if } k + r = 2n + 1\ , \\ \frac{1}{n}(-1)^{k-r} & \text{otherwise} \end{cases} \tag{4.29}$$

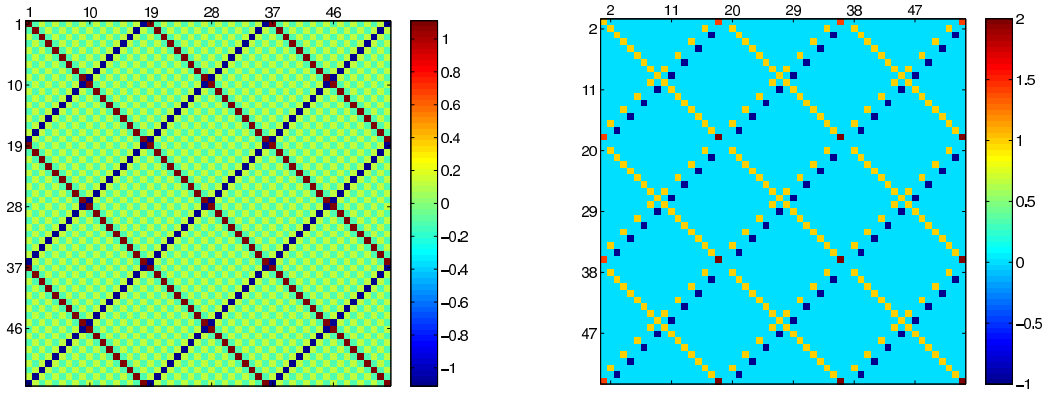
for $1 \le k, r \le 2n$. Letting $\mathbb{I}_s \in \mathbb{R}^n$ be the vector with entries, $(\mathbb{I}_s)_j = (-1)^{j+1}, j \le n$, we observe that $\Psi_n = I_n + \frac{1}{n}\mathbb{I}_s\mathbb{I}_s^T$. It follows that $\lambda_{\min}(\Psi_n) = 1$. It remains to bound the terms in (4.17) involving the infinite sub-block $\Psi_{\widetilde{n}}$.

The Ψ matrix of this example, given by (4.29), shares certain properties with the Ψ obtained in other situations involving periodic eigenfunctions $\{\psi_k\}$. We abstract away these properties by introducing a class of periodic Ψ matrices. We call $\Psi_{\widetilde{n}}$ a *sparse periodic* matrix, if each row (or column) is periodic and in each period only a vanishing fraction of elements are large. More precisely, $\Psi_{\widetilde{n}}$ is *sparse periodic* if there exist positive integers γ and η, and positive constants c_1 and c_2, all independent of n, such that each row of $\Psi_{\widetilde{n}}$ is periodic with period γn. and for any row k, there exits a subset of elements $S_k = \{\ell_1, \dots, \ell_\eta\} \subset \{1, \dots, \gamma n\}$ such that

$$\big|[\Psi]_{k,n+r}\big| \le c_1, \qquad r \in S_k, \tag{4.30a}$$

$$\big|[\Psi]_{k,n+r}\big| \le c_2\, n^{-1}, \qquad r \in \{1, \dots, \gamma n\} \setminus S_k, \tag{4.30b}$$

The elements of S_k could depend on k, but the cardinality of this set should be the constant η, independent of k and n. Also, note that we are indexing rows and columns of $\Psi_{\widetilde{n}}$ by $\{n+1, n+2, \dots\}$; in particular, $k \ge n+1$. For this class, we have the following whose proof can be found in Appendix 4.B.



(a) (b)

Figure 4.2: Sparse periodic Ψ matrices. Display (a) is a plot of the N-by-N leading principal submatrix of Ψ for the Sobolev kernel $(s,t) \mapsto \min\{s,t\}$. Here $n = 9$ and $N = 6n$; the period is $2n = 18$. Display (b) is a the same plot for a Fourier-type kernel. The plots exhibit sparse periodic patterns as defined in §4.3.2.

Lemma 24. *Assume $\Psi_{\tilde{n}}$ to be sparse periodic as defined above and $\sigma_k = \mathcal{O}(k^{-\alpha})$, $\alpha \geq 2$. Then,*

(a) for $\alpha > 2$, $\lambda_{\max}\big(\Sigma_{\tilde{n}}^{1/2}\Psi_{\tilde{n}}\Sigma_{\tilde{n}}^{1/2}\big) = \mathcal{O}(n^{-\alpha})$, $n \to \infty$,

(b) for $\alpha = 2$, $\lambda_{\max}\big(\Sigma_{\tilde{n}}^{1/2}\Psi_{\tilde{n}}\Sigma_{\tilde{n}}^{1/2}\big) = \mathcal{O}(n^{-2}\log n)$, $n \to \infty$.

In particular (4.29) implies that $\Psi_{\tilde{n}}$ is sparse periodic with parameters $\gamma = 2$, $\eta = 2$, $c_1 = 2$ and $c_2 = 1$. Hence, part (b) of Lemma 24 applies. Now, we can use (4.17) with $p = n$ to obtain

$$R_{\mathbb{S}_{x_1^n}}(\varepsilon) \leq 2\varepsilon^2 + \mathcal{O}\big(n^{-2}\log n\big) \tag{4.31}$$

where we have also used $(a+b)^2 \leq 2a^2 + 2b^2$.

Fourier-type kernels

In this example, we consider an RKHS of functions on $\mathcal{X} = [0,1] \subset \mathbb{R}$, generated by a *Fourier-type* kernel defined as $\mathbb{K}(x,y) := \kappa(x-y)$, $x, y \in [0,1]$, where

$$\kappa(x) = \zeta_0 + \sum_{k=1}^{\infty} 2\zeta_k \cos(2\pi k x), \quad x \in [-1,1]. \tag{4.32}$$

We assume that (ζ_k) is a $\mathbb{R}_+$-valued nonincreasing sequence in ℓ_1, i.e. $\sum_k \zeta_k < \infty$. Thus, the trigonometric series in (4.32) is absolutely (and uniformly) convergent. As for the operator Φ, we consider the uniform time sampling operator $\mathbb{S}_{x_1^n}$, as in the previous example. That is, the operator defined in (4.12) with $x_i = i/n, i \leq n$. We take $\mathbb{P}$ to be uniform.

This setting again has the benefit of being simple enough to allow for explicit computations while also practically important. One can argue that the eigen-decomposition of the kernel integral operator is given by

$$\psi_1 = \psi_0^{(c)}, \quad \psi_{2k} = \psi_k^{(c)}, \quad \psi_{2k+1} = \psi_k^{(s)}, \quad k \geq 1 \tag{4.33}$$

$$\sigma_1 = \zeta_0, \quad \sigma_{2k} = \zeta_k, \quad \sigma_{2k+1} = \zeta_k, \quad k \geq 1 \tag{4.34}$$

where $\psi_0^{(c)}(x) := 1$, $\psi_k^{(c)}(x) := \sqrt{2}\cos(2\pi kx)$ and $\psi_k^{(s)}(t) := \sqrt{2}\sin(2\pi kx)$ for $k \geq 1$.

For any integer k, let $(\!(k)\!)_n$ denote k modulo n. Also, let $k \mapsto \delta_k$ be the function defined over integers which is 1 at $k = 0$ and zero elsewhere. Let $\iota := \sqrt{-1}$. Using the identity $n^{-1}\sum_{\ell=1}^{n} \exp(\iota 2\pi k\ell/n) = \delta_{(\!(k)\!)_n}$, one obtains the following,

$$\langle \psi_k^{(c)}, \psi_j^{(c)} \rangle_\Phi = \left[\delta_{(\!(k-j)\!)_n} + \delta_{(\!(k+j)\!)_n}\right] \left(\frac{1}{\sqrt{2}}\right)^{\delta_k + \delta_j}, \tag{4.35a}$$

$$\langle \psi_k^{(s)}, \psi_j^{(s)} \rangle_\Phi = \delta_{(\!(k-j)\!)_n} - \delta_{(\!(k+j)\!)_n}, \tag{4.35b}$$

$$\langle \psi_k^{(c)}, \psi_j^{(s)} \rangle_\Phi = 0, \qquad \text{valid for all } j, k \geq 0. \tag{4.35c}$$

It follows that $\Psi_n = I_n$ if n is odd and $\Psi_n = \mathrm{diag}\{1, 1, \ldots, 1, 2\}$ if n is even. In particular, $\lambda_{\min}(\Psi_n) = 1$ for all $n \geq 1$. It is also clear that the principal submatrix of Ψ on indices $\{2, 3, \ldots\}$ has periodic rows and columns with period $2n$. If follows that Ψ_n is sparse periodic as defined in §4.3.2 with parameters $\gamma = 2$, $\eta = 2$, $c_1 = 2$ and $c_2 = 0$.

Suppose for example that the eigenvalues decay polynomially, say as $\zeta_k = \mathcal{O}(k^{-\alpha})$ for $\alpha > 2$. Then, applying (4.17) with $p = n$, in combination with Lemma 24 part (a), we get

$$R_{\mathbb{S}_{x_1^n}}(\varepsilon) \leq 2\varepsilon^2 + \mathcal{O}(n^{-\alpha}). \tag{4.36}$$

As another example, consider the exponential decay $\zeta_k = \rho^k$, $k \geq 1$ for some $\rho \in (0, 1)$, which corresponds to the Poisson kernel. In this case, the tail sum of $\{\sigma_k\}$ decays as the sequence itself, namely, $\sum_{k>n} \sigma_k \leq 2\sum_{k>n} \rho^k = \frac{2\rho}{1-\rho}\rho^k$. Hence, we can simply use the trace bound (4.20) together with (4.17) to obtain

$$R_{\mathbb{S}_{x_1^n}}(\varepsilon) \leq 2\varepsilon^2 + \mathcal{O}(\rho^n). \tag{4.37}$$

4.4 Proof of Theorem 10

We now turn to the proof of our main theorem. Recall from §4.2.1 the correspondence between any $f \in \mathcal{H}$ and a sequence $\alpha \in \ell_2$; also, recall the diagonal operator $\Sigma : \ell_2 \to \ell_2$ defined by the matrix $\mathrm{diag}\{\sigma_1, \sigma_2, \ldots\}$. Using the definition of (4.15) of the Ψ matrix, we have

$$\|f\|_\Phi^2 = \langle \alpha, \Sigma^{1/2}\Psi\Sigma^{1/2}\alpha \rangle_{\ell_2},$$

By definition (4.6) of the Hilbert space $\mathcal{H}$, we have $\|f\|_{\mathcal{H}}^2 = \sum_{k=1}^{\infty} \alpha_k^2$ and $\|f\|_{L^2}^2 = \sum_k \sigma_k \alpha_k^2$. Letting $B_{\ell_2} = \big\{\alpha \in \ell_2 \;\big|\; \|\alpha\|_{\ell_2} \leq 1\big\}$ be the unit ball in ℓ_2, we conclude that R_Φ can be written as

$$R_\Phi(\varepsilon) = \sup_{\alpha \in B_{\ell_2}} \big\{Q_2(\alpha) \;\big|\; Q_\Phi(\alpha) \leq \varepsilon^2\big\}, \tag{4.38}$$

where we have defined the quadratic functionals

$$Q_2(\alpha) := \langle \alpha, \Sigma\alpha\rangle_{\ell_2}, \quad \text{and} \quad Q_\Phi(\alpha) := \langle \alpha, \Sigma^{1/2}\Psi\Sigma^{1/2}\alpha\rangle_{\ell_2}. \tag{4.39}$$

Also let us define the symmetric bilinear form

$$B_\Phi(\alpha, \beta) := \langle \alpha, \Sigma^{1/2}\Psi\Sigma^{1/2}\beta\rangle_{\ell_2}, \quad \alpha, \beta \in \ell^2, \tag{4.40}$$

whose diagonal is $B_\Phi(\alpha, \alpha) = Q_\Phi(\alpha)$.

We now upper bound $R_\Phi(\varepsilon)$ using a truncation argument. Define the set

$$\mathcal{C} := \{\alpha \in B_{\ell_2} \;|\; Q_\Phi(\alpha) \leq \varepsilon^2\}, \tag{4.41}$$

corresponding to the feasible set for the optimization problem (4.38). For each integer $p = 1, 2, \ldots$, consider the following truncated sequence spaces

$$\begin{aligned} \mathcal{T}_p &:= \big\{\alpha \in \ell_2 \;|\; \alpha_i = 0, \quad \text{for all } i > p\big\}, \quad \text{and} \\ \mathcal{T}_p^\perp &:= \big\{\alpha \in \ell_2 \;|\; \alpha_i = 0, \quad \text{for all } i = 1, 2, \ldots p\big\}. \end{aligned}$$

Note that ℓ_2 is the direct sum of $\mathcal{T}_p$ and $\mathcal{T}_p^\perp$. Consequently, any fixed $\alpha \in \mathcal{C}$ can be decomposed as $\alpha = \xi + \gamma$ for some (unique) $\xi \in \mathcal{T}_p$ and $\gamma \in \mathcal{T}_p^\perp$. Since Σ is a diagonal operator, we have

$$Q_2(\alpha) = Q_2(\xi) + Q_2(\gamma).$$

Moreover, since any $\alpha \in \mathcal{C}$ is feasible for the optimization problem (4.38), we have

$$Q_\Phi(\alpha) \;=\; Q_\Phi(\xi) + 2B_\Phi(\xi, \gamma) + Q_\Phi(\gamma) \;\leq\; \varepsilon^2. \tag{4.42}$$

Note that since $\gamma \in \mathcal{T}_p^\perp$, it can be written as $\gamma = (0_p, c)$, where 0_p is a vector of p zeroes, and $c = (c_1, c_2, \ldots) \in \ell_2$. Similarly, we can write $\xi = (x, 0)$ where $x \in \mathbb{R}^p$. Then, each of the terms $Q_\Phi(\xi)$, $B_\Phi(\xi, \gamma)$, $Q_\Phi(\gamma)$ can be expressed in terms of block partitions of $\Sigma^{1/2}\Psi\Sigma^{1/2}$. For example,

$$Q_\Phi(\xi) = \langle x, Ax\rangle_{\mathbb{R}^p}, \quad Q_\Phi(\gamma) = \langle y, Dy\rangle_{\ell_2}, \tag{4.43}$$

where $A := \Sigma_p^{1/2}\Psi_p\Sigma_p^{1/2}$ and $D := \Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}$, in correspondence with the block partitioning notation of Appendix 4.F. We now apply inequality (4.85) derived in Appendix 4.F. Fix some $\rho^2 \in (0,1)$ and take

$$\kappa^2 := \rho^2 \lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}), \tag{4.44}$$

so that condition (4.88) is satisfied. Then, (4.85) implies

$$Q_\Phi(\xi) + 2B_\Phi(\xi,\gamma) + Q_\Phi(\gamma) \geq \rho^2 Q_\Phi(\xi) - \frac{\kappa^2}{1-\rho^2}\|\gamma\|_2^2. \tag{4.45}$$

Combining (4.42) and (4.45), we obtain

$$Q_\Phi(\xi) \leq \frac{\varepsilon^2}{\rho^2} + \frac{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2})}{1-\rho^2}\|\gamma\|_2^2. \tag{4.46}$$

We further note that $\|\gamma\|_2^2 \leq \|\gamma\|_2^2 + \|\xi\|_2^2 = \|\alpha\|_2^2 \leq 1$. It follows that

$$Q_\Phi(\xi) \;\leq\; \widetilde{\varepsilon}^2, \quad \text{where} \quad \widetilde{\varepsilon}^2 := \frac{\varepsilon^2}{\rho^2} + \frac{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2})}{1-\rho^2}. \tag{4.47}$$

Let us define

$$\widetilde{\mathcal{C}} := \{\xi \in B_{\ell_2} \cap \mathcal{T}_p \mid Q_\Phi(\xi) \leq \widetilde{\varepsilon}^2\}. \tag{4.48}$$

Then, our arguments so far show that for $\alpha \in \mathcal{C}$,

$$Q_2(\alpha) = Q_2(\xi) + Q_2(\gamma) \;\leq\; \underbrace{\sup_{\xi \in \widetilde{\mathcal{C}}} Q_2(\xi)}_{S_p} \;+\; \underbrace{\sup_{\gamma \in B_{\ell_2} \cap \mathcal{T}_p^\perp} Q_2(\gamma)}_{S_p^\perp}. \tag{4.49}$$

Taking the supremum over $\alpha \in \mathcal{C}$ yields the upper bound

$$R_\Phi(\varepsilon) \leq S_p + S_p^\perp.$$

It remains to bound each of the two terms on the right-hand side. Beginning with the term $S_p^\perp$ and recalling the decomposition $\gamma = (0_p, c)$, we have $Q_2(\gamma) = \sum_{k=1}^\infty \sigma_{k+p} c_k^2$, from which it follows that

$$S_p^\perp = \sup\Big\{ \sum_{k=1}^\infty \sigma_{k+p}\, c_k^2 \;\mid\; \sum_{k=1}^\infty c_k^2 \leq 1 \Big\} \;=\; \sigma_{p+1},$$

since $\{\sigma_k\}_{k=1}^\infty$ is a nonincreasing sequence by assumption.

We now control the term S_p. Recalling the decomposition $\xi = (x, 0)$ where $x \in \mathbb{R}^p$, we have

$$\begin{aligned} S_p = \sup_{\xi \in \widetilde{\mathcal{C}}} Q_2(\xi) &= \sup\big\{\langle x, \Sigma_p\, x\rangle \;:\; \langle x, x\rangle \le 1,\ \langle x, \Sigma_p^{1/2}\Psi_p\Sigma_p^{1/2}\, x\rangle \le \widetilde{\varepsilon}^2\big\} \\ &= \sup_{\langle x,x\rangle \le 1} \inf_{t \ge 0} \big\{\langle x, \Sigma_p x\rangle + t\big(\widetilde{\varepsilon}^2 - \langle x, \Sigma_p^{1/2}\Psi_p\Sigma_p^{1/2}\, x\rangle\big)\big\} \\ &\overset{(a)}{\le} \inf_{t \ge 0}\Big\{ \sup_{\langle x,x\rangle \le 1} \langle x, \Sigma_p^{1/2}(I_p - t\Psi_p)\Sigma_p^{1/2}\, x\rangle + t\,\widetilde{\varepsilon}^2\Big\} \end{aligned}$$

where inequality (a) follows by Lagrange (weak) duality. It is not hard to see that for any symmetric matrix M, one has

$$\sup\big\{\langle x, Mx\rangle \;:\; \langle x, x\rangle \le 1\big\} = \max\big\{0, \lambda_{\max}(M)\big\}.$$

Putting the pieces together and optimizing over ρ^2, noting that

$$\inf_{r \in (0,1)} \left\{\frac{a}{r} + \frac{b}{1-r}\right\} = (\sqrt{a} + \sqrt{b})^2$$

for any $a, b > 0$, completes the proof of the bound (4.16).

We now prove bound (4.17), using the same decomposition and notation established above, but writing an upper bound on $Q_2(\alpha)$slightly different form (4.49). In particular, the argument leading to (4.49), also shows that

$$R_\Phi(\varepsilon) \le \sup_{\xi \in \mathcal{T}_p,\, \gamma \in \mathcal{T}_p^\perp} \big\{Q_2(\xi) + Q_2(\gamma) \;\mid\; \xi + \gamma \in B_{\ell_2},\ Q_\Phi(\xi) \le \widetilde{\varepsilon}^2\big\}. \tag{4.50}$$

Recalling the expression (4.39) for $Q_\Phi(\xi)$ and noting that $\Psi_p \succeq \lambda_{\min}(\Psi_p) I_p$ implies $A = \Sigma_p^{1/2}\Psi_p\Sigma_p^{1/2} \succeq \lambda_{\min}(\Psi_p)\Sigma_p$, we have

$$Q_\Phi(\xi) \;\ge\; \lambda_{\min}(\Psi_p)\, Q_2(\xi). \tag{4.51}$$

Now, since we are assuming $\lambda_{\min}(\Psi_p) > 0$, we have

$$R_\Phi(\varepsilon) \le \sup_{\xi \in \mathcal{T}_p,\, \gamma \in \mathcal{T}_p^\perp} \left\{Q_2(\xi) + Q_2(\gamma) \;\middle|\; \xi + \gamma \in B_{\ell_2},\ Q_2(\xi) \le \frac{\widetilde{\varepsilon}^2}{\lambda_{\min}(\Psi_p)}\right\}. \tag{4.52}$$

The RHS of the above is an instance of the Fourier truncation problem with ε^2 replaced with $\widetilde{\varepsilon}^2/\lambda_{\min}(\Psi_p)$. That problem is workout in detail in Appendix 4.E. In particular, applying equation (4.83) in Appendix 4.E with ε^2 changed to $\widetilde{\varepsilon}^2/\lambda_{\min}(\Psi_p)$ completes the proof of (4.17). Figure 4.3 provides a graphical representation of the geometry of the proof.

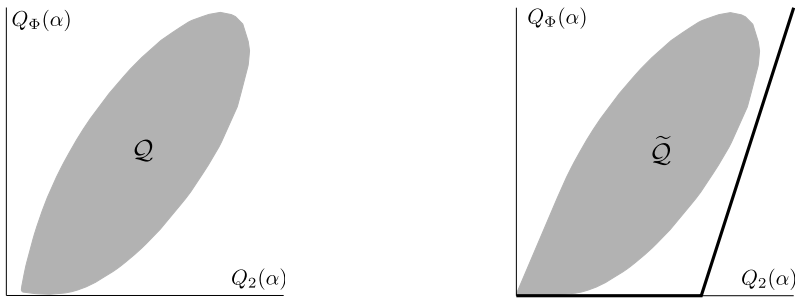


Figure 4.3: Geometry of the proof of (4.17). Display (a) is a plot of the set $\mathcal{Q} := \{(Q_2(\alpha), Q_\Phi(\alpha)) : \|\alpha\|_{\ell_2} = 1\} \subset \mathbb{R}^2$. This is a convex set as a consequence of Hausdorff-Toeplitz theorem on convexity of the numerical range and preservation of convexity under projections. Display (b) shows the set $\widetilde{\mathcal{Q}} := \mathrm{conv}(0, \mathcal{Q})$, i.e., the convex hull of $\{0\} \cup \mathcal{Q}$. Observe that $R_\Phi(\varepsilon) = \sup\{x : (x, y) \in \widetilde{\mathcal{Q}},\ y \leq \varepsilon^2\}$. For any fixed $r \in (0, 1)$, the bound of (4.17) is a piecewise linear approximation to one side of $\widetilde{\mathcal{Q}}$ as shown in Display (b).

4.5 Conclusion

We considered the problem of bounding (squared) L^2 norm of functions in a Hilbert unit ball, based on restrictions on an operator-induced norm acting as a surrogate for the L^2 norm. In particular, given that $f \in B_\mathcal{H}$ and $\|f\|_\Phi^2 \leq \varepsilon^2$, our results enable us to obtain, by estimating norms of certain finite and infinite dimensional matrices, inequalities of the form

$$\|f\|_{L^2}^2 \leq c_1 \varepsilon^2 + h_{\Phi,\mathcal{H}}(\sigma_n)$$

where $\{\sigma_n\}$ are the eigenvalues of the operator embedding $\mathcal{H}$ in L^2, $h_{\Phi,\mathcal{H}}(\cdot)$ is an increasing function (depending on Φ and $\mathcal{H}$) and $c_1 \geq 1$ is some constant. We considered examples of operators Φ (uniform time sampling and Fourier truncation) and Hilbert spaces $\mathcal{H}$ (Sobolev, Fourier-type RKHSs) and showed that it is possible to obtain optimal scaling $h_{\Phi,\mathcal{H}}(\sigma_n) = \mathcal{O}(\sigma_n)$ in most of those cases. We also considered random time sampling, under polynomial eigen-decay $\sigma_n = \mathcal{O}(n^{-\alpha})$, and effectively showed that $h_{\Phi,\mathcal{H}}(\sigma_n) = \mathcal{O}(n^{-\alpha/(\alpha+1)})$ (for ε small enough), with high probability as $n \to \infty$. This last result complements those on related quantities obtained by techniques form empirical process theory, and we conjecture it to be sharp.

Appendix 4.A Analysis of random time sampling

This section is devoted to the proof of Corollary 4 on random time sampling in reproducing kernel Hilbert spaces. The proof is based on an auxiliary result, which we begin by stating. Fix some positive integer m and define

$$\nu(\varepsilon) = \nu(\varepsilon; m) := \inf\Big\{p : \sum_{k > p^m} \sigma_k \leq \varepsilon^2\Big\}. \tag{4.53}$$

With this notation, we have

Lemma 25. *Assume* $\varepsilon^2 < \sigma_1$ *and* $32\, C_\psi^2\, m\, \nu(\varepsilon) \log \nu(\varepsilon) \le n$. *Then,*

$$\mathbb{P}\big\{ R_{\mathbb{S}_{x_1^n}}(\varepsilon) > \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_{\nu(\varepsilon)} \big\} \le 2 \exp\Big(- \frac{1}{32 C_\psi^2} \frac{n}{\nu(\varepsilon)} \Big). \tag{4.54}$$

We prove this claim in §4.A.2 below.

4.A.1 Proof of Corollary 4

To apply the lemma, recall that we assume that there exists m such that for all (large) p, one has

$$\sum_{k > p^m} \sigma_k \le \sigma_p. \tag{4.55}$$

and we let m_σ be the smallest such m. We define

$$\mu(\varepsilon) := \inf \big\{ p : \sigma_p \le \varepsilon^2 \big\}, \tag{4.56}$$

and note that by (4.55), we have $\nu(\varepsilon; m_\sigma) \le \mu(\varepsilon)$. Then, Lemma 25 states that as long as $\varepsilon^2 < \sigma_1$ and $32 C_\psi^2 m_\sigma \mu(\varepsilon) \log \mu(\varepsilon) \le n$, we have

$$\mathbb{P}\big\{ R_{\mathbb{S}_{x_1^n}}(\varepsilon) > (\widetilde{C}_\psi + \widetilde{C}_\sigma)\varepsilon^2 \big\} \le 2 \exp\Big(- \frac{1}{32 C_\psi^2} \frac{n}{\mu(\varepsilon)} \Big). \tag{4.57}$$

Now by the definition of $\mu(\varepsilon)$, we have $\sigma_j > \varepsilon^2$ for $j < \mu(\varepsilon)$, and hence

$$\mathcal{G}_n^2(\varepsilon) \ge \frac{1}{n} \sum_{j\, <\, \mu(\varepsilon)} \min\{\sigma_j, \varepsilon^2\} = \frac{\mu(\varepsilon) - 1}{n}\, \varepsilon^2 \ge \frac{\mu(\varepsilon)}{2n}\, \varepsilon^2,$$

since $\mu(\varepsilon) \ge 2$ when $\varepsilon^2 < \sigma_1$. One can argue that $\varepsilon \mapsto \mathcal{G}_n(\varepsilon)/\varepsilon$ is nonincreasing. It follows from definition (4.26) that for $\varepsilon \ge r_n$, we have

$$\mu(\varepsilon) \le 2n \Big(\frac{\mathcal{G}(\varepsilon)}{\varepsilon} \Big)^2 \le 2n \Big(\frac{\mathcal{G}(r_n)}{r_n} \Big)^2 \le 2n r_n^2,$$

which completes the proof of Corollary 4.

4.A.2 Proof of Lemma 25

For $\xi \in \mathbb{R}^p$, let $\xi \otimes \xi$ be the rank-one operator on $\mathbb{R}^p$ given by $\eta \mapsto \langle \xi, , \eta \rangle_2 \, \xi$. For an operator A on $\mathbb{R}^p$, let $|\!|\!| A |\!|\!| 2$ denote its usual operator norm, $|\!|\!| A |\!|\!| 2 := \sup_{\|x\|_2 \le 1} \|Ax\|_2$. Recall that for a symmetric (i.e., real self-adjoint) operator A on $\mathbb{R}^p$, $|\!|\!| A |\!|\!| 2 = \sup\{|\lambda| : \lambda \text{ an eigenvalue of } A\}$. It follows that $|\!|\!| A |\!|\!| 2 \le \alpha$ is equivalent to $-\alpha I_p \preceq A \preceq \alpha I_p$.

Our approach is to first show that $|\!|\!| \Psi_p - I_p |\!|\!| 2 \le \frac{1}{2}$ for some properly chosen p with high probability. It then follows that $\lambda_{\min}(\Psi_p) \ge \frac{1}{2}$ and we can use bound (4.17) for that value of p. Then, we need to control $\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\big)$. To do this, we further partition $\Psi_{\widetilde{p}}$ into blocks. In order to have a consistent notation, we look at the whole matrix Ψ and let $\Psi^{(k)}$ be the principal submatrix indexed by $\{(k-1)p+1, \dots, (k-1)p+p\}$, for $k = 1, 2, \dots, p^{m-1}$. Throughout the proof, m is assumed to be a fixed positive integer. Also, let $\Psi^{(\infty)}$ be the principal submatrix of Ψ indexed by $\{p^m + 1, p^m + 2, \dots\}$. This provides a full partitioning of Ψ for which $\Psi^{(1)}, \dots, \Psi^{(p^{m-1})}$ and $\Psi^{(\infty)}$ are the diagonal blocks, the first p^{m-1} of which are p-by-p matrices and the last an infinite matrix. To connect with our previous notations, we note that $\Psi^{(1)} = \Psi_p$ and that $\Psi^{(2)}, \dots, \Psi^{(p^{m-1})}, \Psi^{(\infty)}$ are diagonal blocks of $\Psi_{\widetilde{p}}$. Let us also partition the Σ matrix and name its diagonal blocks similarly.

We will argue that, in fact, we have $|\!|\!| \Psi^{(k)} - I_p |\!|\!| 2 \le \frac{1}{2}$ for all $k = 1, \dots, p^{m-1}$, with high probability. Let $\mathcal{A}_p$ denote the event on which this claim holds. In particular, on event $\mathcal{A}_p$, we have $\Psi^{(k)} \preceq \frac{3}{2} I_p$ for $k = 2, \dots, p^{m-1}$; hence, we can write

$$\begin{aligned}
\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\big) &\le \sum_{k=2}^{p^{m-1}} \lambda_{\max}\Big(\sqrt{\Sigma^{(k)}}\, \Psi^{(k)} \sqrt{\Sigma^{(k)}}\Big) + \lambda_{\max}\Big(\sqrt{\Sigma^{(\infty)}}\, \Psi^{(\infty)} \sqrt{\Sigma^{(\infty)}}\Big) \\
&\le \frac{3}{2} \sum_{k=2}^{p^{m-1}} \lambda_{\max}\big(\Sigma^{(k)}\big) + \operatorname{tr}\Big(\sqrt{\Sigma^{(\infty)}}\, \Psi^{(\infty)} \sqrt{\Sigma^{(\infty)}}\Big) \\
&= \frac{3}{2} \sum_{k=2}^{p^{m-1}} \sigma_{(k-1)p+1} + \sum_{k > p^m} \sigma_k [\Psi]_{kk}. \qquad (4.58)
\end{aligned}$$

Using assumptions (4.23) on the sequence $\{\sigma_k\}$, the first sum can be bounded as

$$\sum_{k=2}^{p^{m-1}} \sigma_{(k-1)p+1} \le \sum_{k=2}^{p^{m-1}} \sigma_{(k-1)p} \le \sum_{k=2}^{p^{m-1}} C_\sigma \sigma_{k-1} \sigma_p \le C_\sigma \|\sigma\|_1 \sigma_p$$

Using the uniform boundedness assumption (4.53), we have $[\Psi]_{kk} = n^{-1} \sum_{i=1}^n \psi_k^2(x_i) \le C_\psi^2$. Hence the second sum in (4.58) is bounded above by $C_\psi^2 \sum_{k > p^m} \sigma_k$.

We can now apply Theorem 10. Assume for the moment that $\varepsilon^2 \ge \sum_{k > p^m} \sigma_k$ so that the right-hand side of (4.58) is bounded above by $\frac{3}{2} C_\sigma \|\sigma\|_1 \sigma_p + C_\psi^2 \varepsilon^2$. Applying bound (4.17),

on event $\mathcal{A}_p$, with[8] $r = (1 + C_\psi)^{-1}$, we get

$$\begin{aligned} R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) &\le 2\Big\{ r^{-1}\varepsilon^2 + (1-r)^{-1}\Big(\frac{3}{2} C_\sigma \|\sigma\|_1 \sigma_p + C_\psi^2 \varepsilon^2\Big)\Big\} + \sigma_{p+1} \\ &= 2(1+C_\psi)^2 \varepsilon^2 + 3(1 + C_\psi^{-1}) C_\sigma \|\sigma\|_1 \sigma_p + \sigma_{p+1}. \\ &\le \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p \end{aligned}$$

where $\widetilde{C}_\psi := 2(1 + C_\psi)^2$ and $\widetilde{C}_\sigma := 3(1 + C_\psi^{-1}) C_\sigma \|\sigma\|_1 + 1$. To summarize, we have shown the following

$$\text{Event } \mathcal{A}_p \quad \text{and} \quad \varepsilon^2 \ge \sum_{k > p^m} \sigma_k \implies R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) \le \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p. \tag{4.59}$$

It remains to control the probability of $\mathcal{A}_p := \bigcap_{k=1}^{p^{m-1}} \big\{ \|\!| \Psi^{(k)} - I_p |\!\| 2 \le \frac{1}{2} \big\}$. We start with the deviation bound on $\Psi^{(1)} - I_p$, and then extend by union bound. We will use the following lemma which follows, for example, from the Ahlswede-Winter bound [1], or from [85]. (See also [98, 91, 102].)

Lemma 26. *Let $\xi_1, \dots, \xi_n$ be i.i.d. random vectors in $\mathbb{R}^p$ with $\mathbb{E}\, \xi_1 \otimes \xi_1 = I_p$ and $\|\xi_1\|_2 \le C_p$ almost surely for some constant C_p. Then, for $\delta \in (0,1)$,*

$$\mathbb{P}\Big\{ \Big\| n^{-1} \sum_{i=1}^n \xi_i \otimes \xi_i - I_p \Big\|_2 > \delta \Big\} \le p \exp\big(- \frac{n\delta^2}{4C_p^2} \big). \tag{4.60}$$

Recall that for the time sampling operator, $[\Phi\, \psi_k]_i = \frac{1}{\sqrt{n}} \psi_k(x_i)$ so that from (4.15),

$$\Psi_{k\ell} = \frac{1}{n} \sum_{i=1}^n \psi_k(x_i) \psi_\ell(x_i)$$

Let $\xi_i := (\psi_k(x_i), 1 \le k \le p) \in \mathbb{R}^p$ for $i = 1, \dots, n$. Then, $\{\xi_i\}$ satisfy the conditions of Lemma 26. In particular, letting e_k denote the k-th standard basis vector of $\mathbb{R}^p$, we note that

$$\langle e_k, \mathbb{E}(\xi_i \otimes\, \xi_i) e_\ell \rangle_2 = \mathbb{E} \langle e_k, \xi_i \rangle_2 \langle e_\ell, \xi_i \rangle_2 = \langle \psi_k, \psi_\ell \rangle_{L^2} = \delta_{k\ell}$$

and $\|\xi_i\|_2 \le \sqrt{p}\, C_\psi$, where we have used uniform boundedness of $\{\psi_k\}$ as in (4.22). Furthermore, we have $\Psi^{(1)} = n^{-1} \sum_{i=1}^n \xi_i \otimes \xi_i$. Applying Lemma 26 with $C_p = \sqrt{p} C_\psi$ yields,

$$\mathbb{P}\big\{ \|\!| \Psi^{(1)} - I_p |\!\| 2 > \delta \big\} \;\le\; p \exp\big(- \frac{\delta^2}{4C_\psi^2} \frac{n}{p} \big). \tag{4.61}$$

[8] We are using the alternate form of the bound based on $(\sqrt{A} + \sqrt{B})^2 = \inf_{r \in (0,1)} \big\{ A r^{-1} + B(1-r)^{-1} \big\}$.

Similar bounds hold for $\Psi^{(k)}$, $k = 2, \dots, p^{m-1}$. Applying the union bound, we get

$$\mathbb{P} \bigcup_{k=1}^{p^{m-1}} \{ \|\!| \Psi^{(k)} - I_p |\!\| 2 > \delta \} \;\le\; \exp\big(m \log p - \frac{\delta^2}{4C_\psi^2} \frac{n}{p} \big).$$

For simplicity, let $A = A_{n,p} := n/(4C_\psi^2\, p)$. We impose $m \log p \le \frac{A}{2}\, \delta^2$ so that the exponent in (4.61) is bounded above by $-\frac{A}{2}\delta^2$. Furthermore, for our purpose, it is enough to take $\delta = \frac{1}{2}$. It follows that

$$\mathbb{P}(\mathcal{A}_p^c) = \mathbb{P} \bigcup_{k=1}^{p^{m-1}} \{ \|\!| \Psi^{(k)} - I_p |\!\| 2 > \frac{1}{2} \} \;\le\; \exp\big(- \frac{1}{32C_\psi^2} \frac{n}{p} \big), \tag{4.62}$$

if $32C_\psi^2\, m\, p \log p \le n$. Now, by (4.59), under $\varepsilon^2 \ge \sum_{k>p^m} \sigma_k$, $R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) > \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p$ implies $\mathcal{A}_p^c$. Thus, the exponential bound in (4.62) holds for $\mathbb{P}\{R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) > \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p\}$ under the assumptions. We are to choose p and the bound is optimized by making p as small as possible. Hence, we take p to be $\nu(\varepsilon) := \inf\{p : \; \varepsilon^2 \ge \sum_{k>p^m} \sigma_k\}$ which proves Lemma 25. (Note that, in general, $\nu(\varepsilon)$ takes its values in $\{0, 1, 2, \dots\}$. The assumption $\varepsilon^2 < \sigma_1$ guarantees that $\nu(\varepsilon) \neq 0$.)

Appendix 4.B Proof of Lemma 24

Assume $\sigma_k = Ck^{-\alpha}$, for some $\alpha \ge 2$. First, note the following upper bound on the tail sum

$$\sum_{k>p} \sigma_k \le C \int_p^\infty x^{-\alpha}\, dx = C_1(\alpha)\, p^{1-\alpha}. \tag{4.63}$$

Furthermore, from the bounds (4.30a) and (4.30b), we have, for $k \ge n+1$,

$$[\Psi]_{kk} \le \min\{c_1, c_2\}. \tag{4.64}$$

To simplify notation, let us define $I_n := \{1, 2, \dots, \gamma n\}$.

Consider the case $\alpha > 2$. We will use the ℓ_∞–ℓ_∞ upper bound of (4.21), with $p = n$. Fix some $k \ge n+1$. Note that $\sigma_k \le \sigma_{n+1}$. Then, recalling the assumptions on Ψ and the definition of S_k, we have

$$\begin{aligned}
\sum_{\ell \ge n+1} \sqrt{\sigma_k}\sqrt{\sigma_\ell}\, \big|[\Psi]_{k,\ell}\big| &\le \sqrt{\sigma_{n+1}} \sum_{q=0}^{\infty} \sum_{r=1}^{\gamma n} \sqrt{\sigma_{n+r+q\gamma n}} \big|[\Psi]_{k,n+r+q\gamma n}\big| \\
&= \sqrt{\sigma_{n+1}} \sum_{q=0}^{\infty} \sum_{r=1}^{\gamma n} \sqrt{\sigma_{n+r+q\gamma n}} \big|[\Psi]_{k,n+r}\big| \\
&\le \sqrt{\sigma_{n+1}} \sum_{q=0}^{\infty} \Big\{ c_1 \sum_{r \in S_k} \sqrt{\sigma_{n+r+q\gamma n}} + \frac{c_2}{n} \sum_{r \in I_n \setminus S_k} \sqrt{\sigma_{n+r+q\gamma n}} \Big\}.
\end{aligned} \tag{4.65}$$

Using (4.63), the second double sum in (4.65) is bounded by

$$\sum_{q=0}^{\infty} \sum_{r \in I_n \setminus S_k} \sqrt{\sigma_{n+r+q\gamma n}} \;\le\; \sum_{\ell > n} \sqrt{\sigma_\ell} \;\le\; C_2(\alpha)\, n^{1-\alpha/2}. \tag{4.66}$$

Recalling that $S_k \subset I_n$ and $|S_k| = \eta$, the first double sum in (4.65) can be bounded as follows

$$\begin{aligned}
\sum_{q=0}^{\infty} \sum_{r \in S_k} \sqrt{\sigma_{n+r+q\gamma n}} &= \sqrt{C} \sum_{q=0}^{\infty} \sum_{r \in S_k} (n + r + q\gamma n)^{-\alpha/2} \\
&\le \sqrt{C} \sum_{q=0}^{\infty} \sum_{r \in S_k} (n + q\gamma n)^{-\alpha/2} \\
&\le \sqrt{C}\, \eta \sum_{q=0}^{\infty} (1 + q\gamma)^{-\alpha/2} n^{-\alpha/2} \\
&\le \sqrt{C}\, \eta \Big(1 + \gamma^{-\alpha/2} \sum_{q=1}^{\infty} q^{-\alpha/2}\Big) n^{-\alpha/2} \\
&= C_3(\alpha, \gamma, \eta)\, n^{-\alpha/2}
\end{aligned} \tag{4.67}$$

where in the last line we have used $\sum_{q=1}^{\infty} q^{-\alpha/2} < \infty$ due to $\alpha/2 > 1$. Combining (4.65), (4.66) and (4.67) and noting that $\sqrt{\sigma_{n+1}} \le \sqrt{C} n^{-\alpha/2}$, we obtain

$$\sum_{\ell \ge n+1} \sqrt{\sigma_k}\sqrt{\sigma_\ell}\, \big|[\Psi]_{k,\ell}\big| \le \sqrt{C} n^{-\alpha/2} \Big\{ c_1 C_3(\alpha, \gamma, \eta)\, n^{-\alpha/2} + \frac{c_2}{n}\, C_2(\alpha)\, n^{1-\alpha/2} \Big\} = C_4(\alpha, \eta, \gamma)\, n^{-\alpha}. \tag{4.68}$$

Taking supremum over $k \ge 1$ and applying the ℓ_∞–ℓ_∞ bound of (4.21), with $p = n$, concludes the proof of part (a).

Now, consider the case $\alpha = 2$. The above argument breaks down in this case because $\sum_{q=1}^{\infty} q^{-\alpha/2}$ does not converge for $\alpha = 2$. A remedy is to further partition the matrix $\Sigma_{\widetilde{n}}^{1/2} \Psi_{\widetilde{n}} \Sigma_{\widetilde{n}}^{1/2}$. Recall that the rows and columns of this matrix are indexed by $\{n+1, n+2, \dots\}$. Let A be the principal submatrix indexed by $\{n+1, n+2, \dots, n^2\}$ and D be the principal submatrix indexed by $\{n^2+1, n^2+2, \dots\}$. We will use a combination of the bounds (4.30a) and (4.30b), and the well-known perturbation bound $\lambda_{\max}\big[\left(\begin{smallmatrix} A & C \\ C^T & D \end{smallmatrix}\right)\big] \le \lambda_{\max}(A) + \lambda_{\max}(D)$, to write

$$\lambda_{\max}\big(\Sigma_{\widetilde{n}}^{1/2} \Psi_{\widetilde{n}} \Sigma_{\widetilde{n}}^{1/2}\big) \le \lambda_{\max}(A) + \lambda_{\max}(D) \le |\!|\!| A |\!|\!|_\infty + \mathrm{tr}(D). \tag{4.69}$$

The second term is bounded as

$$\mathrm{tr}(D) = \sum_{k > n^2} \sigma_k\, [\Psi]_{kk} \le \min\{c_1, c_2\} \sum_{k > n^2} \sigma_k = \min\{c_1, c_2\}\, (n^2)^{1-2} = C_5(\gamma)\, n^{-2}, \tag{4.70}$$

where we have used (4.63) and (4.64). To bound the first term, fix $k \in \{n+1, \dots, n^2\}$. By an argument similar to that of part (a) and noting that $\gamma \geq 1$, hence $\gamma n^2 \geq n^2$, we have

$$\begin{aligned}\sum_{\ell=n+1}^{n^2} \sqrt{\sigma_k}\sqrt{\sigma_\ell}\,\big|[\Psi]_{k,\ell}\big| &\leq \sqrt{\sigma_{n+1}} \sum_{q=0}^{n} \sum_{r=1}^{\gamma n} \sqrt{\sigma_{n+r+q\gamma n}}\big|[\Psi]_{k,n+r}\big| \\ &\leq \sqrt{\sigma_{n+1}} \sum_{q=0}^{n} \Big\{ c_1 \sum_{r\,\in\, S_k} \sqrt{\sigma_{n+r+q\gamma n}} + \frac{c_2}{n} \sum_{r\,\in\, I_n \setminus S_k} \sqrt{\sigma_{n+r+q\gamma n}} \Big\}. \end{aligned} \tag{4.71}$$

Using $\gamma \geq 1$ again, the second double sum in (4.71) is bounded as

$$\sum_{q=0}^{n} \sum_{r\,\in\, I_n \setminus S_k} \sqrt{\sigma_{n+r+q\gamma n}} \leq \sum_{\ell=n+1}^{3\gamma n^2} \sqrt{\sigma_\ell} \leq \sqrt{C} \sum_{\ell=2}^{3\gamma n^2} \frac{1}{\ell} \leq \sqrt{C}\log(3\gamma n^2) \leq C_6(\gamma)\log n, \tag{4.72}$$

for sufficiently large n. Note that we have used the bound $\sum_{\ell=2}^{p} \ell^{-1} \leq \int_1^p x^{-1}\,dx = \log p$. The first double sum in (4.71) is bounded as follows

$$\begin{aligned}\sum_{q=0}^{\infty} \sum_{r\,\in\, S_k} \sqrt{\sigma_{n+r+q\gamma n}} &= \sqrt{C} \sum_{q=0}^{n} \sum_{r\,\in\, S_k} (n+r+q\gamma n)^{-1} \\ &\leq \sqrt{C}\,\eta \sum_{q=0}^{n} (1+q\gamma)^{-1} n^{-1} \\ &\leq \sqrt{C}\,\eta \Big(1 + \gamma^{-1} + \gamma^{-1} \sum_{q=2}^{n} q^{-1}\Big) n^{-1} \\ &= C_7(\gamma,\eta)\, n^{-1} \log n, \end{aligned} \tag{4.73}$$

for n sufficiently large. Combining (4.71), (4.72) and (4.73), taking supremum over k and using the simple bound $\sqrt{\sigma_{n+1}} \leq \sqrt{C} n^{-1}$, we get

$$\|\!|A|\!\|_\infty \leq \sqrt{C} n^{-1} \Big\{ c_1 C_7(\gamma,\eta)\, \frac{\log n}{n} + \frac{c_2}{n}\, C_6(\gamma)\, \log n \Big\} = C_8(\gamma,\eta)\, \frac{\log n}{n^2} \tag{4.74}$$

which in view of (4.70) and (4.69) completes the proof of part (b).

Appendix 4.C Relationship between $R_\Phi(\varepsilon)$ and $\underline{T}_\Phi(\varepsilon)$

In this appendix, we prove the claim made in §4.1 about the relation between the upper quantities R_Φ and T_Φ and the lower quantities $\underline{T}_\Phi$ and $\underline{R}_\Phi$. We only carry out the proof for

R_Φ; the dual version holds for T_Φ. To simplify the argument, we look at slightly different versions of R_Φ and $\underline{T}_\Phi$, defined as

$$R_\Phi^\circ(\varepsilon) := \sup\left\{\|f\|_{L^2}^2 : f \in B_\mathcal{H},\ \|f\|_\Phi^2 < \varepsilon^2\right\}, \tag{4.75}$$

$$\underline{T}_\Phi^\circ(\delta) := \inf\left\{\|f\|_\Phi^2 : f \in B_\mathcal{H},\ \|f\|_{L^2}^2 > \delta^2\right\} \tag{4.76}$$

and prove the following

$$R_\Phi^{\circ\,-1}(\delta) = \underline{T}_\Phi^\circ(\delta) \tag{4.77}$$

where $R_\Phi^{\circ\,-1}(\delta) := \inf\{\varepsilon^2 : R_\Phi^\circ(\varepsilon) > \delta^2\}$ is a generalized inverse of R_Φ°. To see (4.77), we note that $R_\Phi(\varepsilon) > \delta^2$ iff there exists $f \in B_\mathcal{H}$ such that $\|f\|_\Phi^2 < \varepsilon^2$ and $\|f\|_{L^2}^2 > \delta^2$. But this last statement is equivalent to $\underline{T}_\Phi^\circ(\delta) < \varepsilon^2$. Hence,

$$R_\Phi^{\circ\,-1}(\delta) = \inf\{\varepsilon^2 : \underline{T}_\Phi^\circ(\delta) < \varepsilon^2\} \tag{4.78}$$

which proves (4.77).

Using the following lemma, we can use relation (4.77) to convert upper bounds on R_Φ to lower bounds on $\underline{T}_\Phi$.

Lemma 27. *Let $t \mapsto p(t)$ be a nondecreasing function (defined on the real line with values in the extended real line.). Let q be its generalized inverse defined as $q(s) := \inf\{t : p(t) > s\}$. Let r be a properly invertible (i.e., one-to-one) function such that $p(t) \le r(t)$, for all t. Then,*

(a) $q(p(t)) \ge t$, for all t,

(b) $q(s) \ge r^{-1}(s)$, for all s.

Proof. Assume (a) does not hold, that is, $\inf\{\alpha : p(\alpha) > p(t)\} < t$. Then, there exists α_0 such that $p(\alpha_0) > p(t)$ and $\alpha_0 < t$. But this contradicts $p(t)$ being nondecreasing. For part (b), note that (a) implies $t \le q(p(t)) \le q(r(t))$, since q is nondecreasing by definition. Letting $t := r^{-1}(s)$ and noting that $r(r^{-1}(s)) = s$, by assumption, proves (b). □

Let $p = R_\Phi^\circ$, $q = \underline{T}_\Phi^\circ$ and $r(t) = At + B$ for some constant $A > 0$. Noting that $R_\Phi^\circ \le R_\Phi$ and $\underline{T}_\Phi(\cdot + \gamma) \ge \underline{T}_\Phi^\circ$ for any $\gamma > 0$, we obtain from Lemma 27 and (4.77) that

$$R_\Phi(\varepsilon) \le A\,\varepsilon^2 + B \implies \underline{T}_\Phi(\delta+) \ge \frac{\delta^2}{A} - B, \tag{4.79}$$

where $\underline{T}_\Phi(\delta+)$ denotes the right limit of $\underline{T}_\Phi$ as δ^2. This may be used to translate an upper bound of the form (4.17) on R_Φ to a corresponding lower bound on $\underline{T}_\Phi$.

Appendix 4.D The 2×2 subproblem

The following subproblem arises in the proof of Theorem 10.

$$F(\varepsilon^2) := \sup \Big\{ \underbrace{\begin{pmatrix} r & s \end{pmatrix} \begin{pmatrix} u^2 & 0 \\ 0 & v^2 \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix}}_{=: x(r,s)} : \ r^2 + s^2 \le 1, \ \underbrace{\begin{pmatrix} r & s \end{pmatrix} \begin{pmatrix} a^2 & 0 \\ 0 & d^2 \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix}}_{=: y(r,s)} \le \varepsilon^2 \Big\}, \quad (4.80)$$

where u^2, v^2, a^2 and d^2 are given constants and the optimization is over (r, s). Here, we discuss the solution in some detail; in particular, we provide explicit formulas for $F(\varepsilon^2)$. Without loss of generality assume $u^2 \ge v^2$. Then, it is clear that $F(\varepsilon^2) \le u^2$ and $F(\varepsilon^2) = u^2$ for $\varepsilon^2 \ge u^2$. Thus, we are interested in what happens when $\varepsilon^2 < u^2$.

The problem is easily solved by drawing a picture. Let $x(r, s)$ and $y(r, s)$ be as denoted in the last display. Consider the set

$$\begin{aligned} \mathcal{S} &:= \big\{ \big(x(r,s),\, y(r,s)\big) : \ r^2 + s^2 \le 1 \big\} \\ &= \big\{ r^2(u^2, a^2) + s^2(v^2, d^2) + q^2(0,0) : \ r^2 + s^2 + q^2 = 1 \big\} \\ &= \operatorname{conv} \big\{ (u^2, a^2),\, (v^2, d^2),\, (0,0) \big\}. \end{aligned} \quad (4.81)$$

That is, $\mathcal{S}$ is the convex hull of the three points (u^2, a^2), (v^2, d^2) and the origin $(0, 0)$.

Then, two (or maybe three) different pictures arise depending on whether $a^2 > d^2$ (and whether $d^2 \ge v^2$ or $d^2 < v^2$) or $a^2 \le d^2$; see Fig. 4.4. It follows that we have two (or three) different pictures for the function $\varepsilon^2 \mapsto F(\varepsilon^2)$. In particular, for $a^2 > d^2$ and $d^2 < v^2$,

$$F(\varepsilon^2) = v^2 \min \Big\{ \frac{\varepsilon^2}{d^2}, 1 \Big\} + (u^2 - v^2) \max \Big\{ 0, \frac{\varepsilon^2 - d^2}{a^2 - d^2} \Big\}, \quad (4.82)$$

for $a^2 > d^2$ and $d^2 \ge v^2$, $F(\varepsilon^2) = \varepsilon^2$, and for $a^2 \le d^2$,

$$F(\varepsilon^2) = u^2 \min \Big\{ \frac{\varepsilon^2}{a^2}, 1 \Big\}.$$

All the equations above are valid for $\varepsilon^2 \in [0, \sigma_1]$.

Appendix 4.E Details of Fourier truncation example

Here we establish the claim that the bound (4.19) holds with equality. Recall that for the (generalized) Fourier truncation operator $\mathbb{T}_{\psi_1^n}$, we have

$$R_{\mathbb{T}_{\psi_1^n}}(\varepsilon^2) = \sup \Big\{ \sum_{k=1}^{\infty} \sigma_k \alpha_k^2 : \ \sum_{k=1}^{\infty} \alpha_k^2 \le 1, \ \sum_{k=1}^{n} \sigma_k \alpha_k^2 \le \varepsilon^2 \Big\}$$

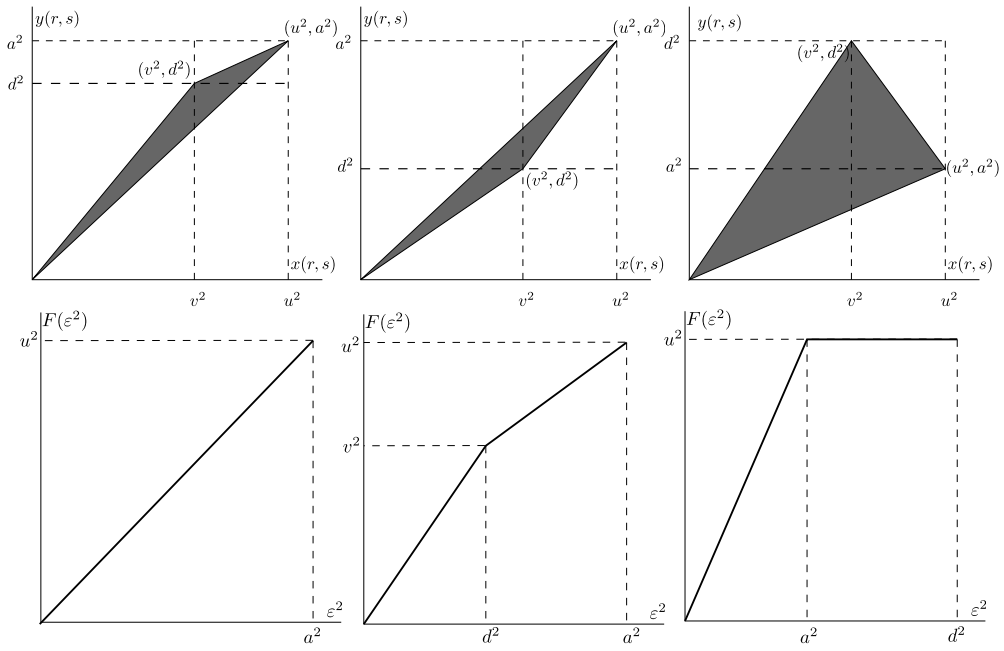


Figure 4.4: Top plots illustrate the set $\mathcal{S}$ as defined in (4.81), in various cases. The bottom plots are the corresponding $\varepsilon^2 \mapsto F(\varepsilon^2)$.

Let $\alpha = (t\xi, s\gamma)$, where $t, s \in \mathbb{R}$, $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$, $\gamma = (\gamma_1, \gamma_2 \dots) \in \ell_2$ and $\|\xi\|_2 = 1 = \|\gamma\|_2$. Let $u^2 = u^2(\xi) := \sum_{k=1}^n \sigma_k \xi_k^2$ and $v^2 = v^2(\gamma) := \sum_{k>n} \sigma_k \gamma_k^2$.

Let us fix ξ and γ for now and try to optimize over t and s. That is, we look at

$$G(\varepsilon^2; \xi, \gamma) := \sup \Big\{ t^2 u^2 + s^2 v^2 : \ t^2 + s^2 \le 1, \ t^2 u^2 \le \varepsilon^2 \Big\}.$$

This is an instance of the 2-by-2 problem (4.80), with $a^2 = u^2$ and $d^2 = 0$. Note that our assumption that $u^2 \ge v^2$ holds in this case, for all ξ and γ, because $\{\sigma_k\}$ is a nonincreasing sequence. Hence, we have, for $\varepsilon^2 \le \sigma_1$,

$$G(\varepsilon^2; \xi, \gamma) = v^2 + (u^2 - v^2)\frac{\varepsilon^2}{u^2} = v^2(\gamma) + \Big(1 - \frac{v^2(\gamma)}{u^2(\xi)}\Big)\varepsilon^2.$$

Now we can maximize $G(\varepsilon^2; \xi, \gamma)$ over ξ and then γ. Note that G is increasing in u^2. Thus, the maximum is achieved by selecting u^2 to be $\sup_{\|\xi\|_2=1} u^2(\xi) = \sigma_1$. Thus,

$$\sup_{\xi} G(\varepsilon^2; \xi, \gamma) = \Big(1 - \frac{\varepsilon^2}{\sigma_1}\Big) v^2(\gamma) + \varepsilon^2.$$

For $\varepsilon^2 < \sigma_1$, the above is increasing in v^2. Hence the maximum is achieved by setting v^2 to be $\sup_{\|\gamma\|_2=1} v^2(\gamma) = \sigma_{n+1}$. Hence, for $\varepsilon^2 \le \sigma_1$

$$R_{\mathbb{T}_{\psi_1^n}}(\varepsilon^2) := \sup_{\xi, \gamma} G(\varepsilon^2; \xi, \gamma) = \Big(1 - \frac{\sigma_{n+1}}{\sigma_1}\Big)\varepsilon^2 + \sigma_{n+1}. \tag{4.83}$$

Appendix 4.F A quadratic inequality

In this appendix, we derive an inequality which will be used in the proof of Theorem 10. Consider a positive semidefinite matrix M (possibly infinite-dimensional) partitioned as

$$M = \begin{pmatrix} A & C \\ C^T & D \end{pmatrix}.$$

Assume that there exists $\rho^2 \in (0,1)$ and $\kappa^2 > 0$ such that

$$\begin{pmatrix} A & C \\ C^T & (1-\rho^2)D + \kappa^2 I \end{pmatrix} \succeq 0. \tag{4.84}$$

Let (x, y) be a vector partitioned to match the block structure of M. Then we have the following.

Lemma 28. *Under (4.84), for all x and y,*

$$x^T A x + 2x^T C y + y^T D y \;\geq\; \rho^2 x^T A x - \frac{\kappa^2}{1-\rho^2}\|y\|_2^2. \tag{4.85}$$

Proof. By assumption (4.84), we have

$$\begin{pmatrix} \sqrt{1-\rho^2}\, x^T & \frac{1}{\sqrt{1-\rho^2}}\, y^T \end{pmatrix} \begin{pmatrix} A & C \\ C^T & (1-\rho^2)D + \kappa^2 I \end{pmatrix} \begin{pmatrix} \sqrt{1-\rho^2}\, x \\ \frac{1}{\sqrt{1-\rho^2}}\, y \end{pmatrix} \;\geq\; 0. \tag{4.86}$$

□

Writing (4.84) as a perturbation of the original matrix,

$$\begin{pmatrix} A & C \\ C^T & D \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -\rho^2 D + \kappa^2 I \end{pmatrix} \succeq 0, \tag{4.87}$$

we observe that a sufficient condition for (4.84) to hold is $\rho^2 D \preceq \kappa^2 I$. That is, it is sufficient to have

$$\rho^2 \lambda_{\max}(D) \leq \kappa^2. \tag{4.88}$$

Rewriting (4.84) differently, as

$$\begin{pmatrix} (1-\rho^2)A & 0 \\ 0 & (1-\rho^2)D \end{pmatrix} + \begin{pmatrix} \rho^2 A & C \\ C^T & \kappa^2 I \end{pmatrix} \succeq 0, \tag{4.89}$$

we find another sufficient condition for (4.84), namely, $\rho^2 A - \kappa^{-2} C C^T \succeq 0$. In particular, it is also sufficient to have

$$\kappa^{-2}\lambda_{\max}(CC^T) \leq \rho^2 \lambda_{\min}(A). \tag{4.90}$$

Chapter 5

Sampled forms of functional PCA in reproducing kernel Hilbert spaces

As mentioned in §1.4, the aim of this chapter is to study effects of "sampling" on functional PCA (fPCA). We recall our functional-theoretic take on sampling, namely a continuous linear operator $\Phi : \mathcal{H} \mapsto \mathbb{R}^m$ acting on some Hilbert subspace $\mathcal{H}$ of L^2 which usually represents some smooth subclass of functions in L^2. We also recall the basic setup: there are functions $x_i(t)$, $t \in [0,1]$ in $\mathcal{H}$ for $i = 1, \ldots, n$, generated i.i.d. from a probabilistic model (to be discussed below) which are acted upon by Φ. We observe the collection $\{\Phi x_i\}_{i=1}^n \subset \mathbb{R}^m$ in noise. The index n is referred to as the number of *statistical samples*, and the index m as the number of *functional samples*.

Our model for the functions $\{x_i\}$ will be an extension to function spaces of the *spiked covariance model* introduced by Johnstone and his collaborators [55, 58], and studied by various authors [58, 75, 3]. We consider such models with r components, each lying within the Hilbert ball $\mathbb{B}_{\mathcal{H}}(\rho)$ of radius ρ, with the goal of recovering the r-dimensional subspace spanned by the spiked components in this functional model. We analyze our M-estimators within a high-dimensional framework that allows both the number of statistical samples n and the number of functional samples m to diverge together. Our theoretical contribution is to provide non-asymptotic bounds on the estimation error as a function of the pair (m, n). Although our rates also explicitly track the number of components r and the smoothness parameter ρ, we do not make any effort to obtain optimal dependence on these parameters.

The general asymptotic properties of PCA in function spaces have been investigated by various authors (e.g., [33, 22, 49].) Accounting for smoothness of functions by introducing various roughness/smoothness penalties is a standard approach, used in the papers [82, 77, 86, 20] among others. The problem of principal component analysis for sampled functions, with a functional-theoretic take on sampling as ours, is discussed in [14], for the noiseless case. A more recent line of work is devoted to the case of functional PCA with noisy sampled functions [27, 105, 51]. Cardot [27] considers estimation via spline-based approximation, and derives MISE rates in terms of various parameters of the model. Hall et

al. [51] study estimation via local linear smoothing, and establish minimax-optimality in certain settings that involve a fixed number of functional samples. Both of these papers [27, 51] have studied trade-offs between the numbers of statistical and functional samples; we refer the reader to Hall et al. [51] for an illuminating discussion of connections between FDA and LDA approaches (i.e. having full versus sampled functions), which inspired much of the present work. We note that the regularization present in our M-estimator is closely related to classical roughness penalties [82, 86] in the special case of spline kernels, although the discussion there applies to fully-observed functions, as opposed to the sampled models considered here.

As mentioned above, our sampled model resembles very much that of spiked covariance model for high-dimensional principal component analysis which was studied in some detail in Chapter 3; in contrast, here the smoothness condition on functional components translates into an ellipsoid condition on the vector principal components. Perhaps an even more significant difference is that, here, the effective scaling of noise in $\mathbb{R}^m$ is substantially smaller in some cases (e.g., the case of time sampling). This could explain why the difficulty of "high-dimensional" setting is not observed in such cases as one lets $m, n \to \infty$. On the other hand, a difficulty particular to our sampled model is the lack of orthonormality between components (after sampling) which leads to identifiability issues; it also makes recovering individual components difficult.

Elements of the technique we use to analyze the M-estimator (such as establishing a perturbation inequality and uniformly controlling the terms involving noise, etc.) have become more or less standard in recent years. We refer the reader to [93] for some general discussions. These techniques lead to finite-sample bounds which hold with high probability. We also draw on the recent work, namely [70], on bounding Gaussian complexities of balls in a RKHS. Techniques from non-asymptotic random matrix theory, for example as discussed in [34] and §2.4.2, are employed in bounding norms of random matrices. We provide a slight extension, in Appendix 5.G, of one such result. Results on controlling suprema of linear product-of-Gaussians processes are also established (cf. Appendix 5.F).

The remainder of this chapter is organized as follows. Section 5.1 is devoted to background material on reproducing kernel Hilbert spaces, adjoints of operators, as well as the class of sampled functional models that we study in this chapter. More details can be found in Chapter 4.2. In Section 5.2, we describe M-estimators for sampled functional PCA, and discuss various implementation details. Section 5.3 is devoted to the statements of our main results, and discussion of their consequences for particular sampling models. In subsequent sections, we provide the proofs of our results, with some more technical aspects deferred to the appendices. Section 5.4 is devoted to bounds on the subspace-based error, whereas Section 5.5 is devoted to bounds on error in the function space. Section 5.6 provides matching lower bounds on the minimax error, showing that our analysis is sharp. We conclude with a discussion in Section 5.7.

5.1 Background and problem set-up

In this section, we begin by introducing background on reproducing kernel Hilbert spaces, as well as linear operators and their adjoints. We then introduce the functional and observation model that we study in this chapter, and conclude with discussion of some approximation-theoretic issues that play an important role in parts of our analysis.

5.1.1 Reproducing Kernel Hilbert Spaces

We begin with a quick overview of some standard properties of reproducing kernel Hilbert spaces; we refer the reader to the books [99, 47] and §2.5 for more details. A reproducing kernel Hilbert space (or RKHS for short) is a Hilbert space $\mathcal{H}$ of functions $f : T \to \mathbb{R}$ that is equipped with an associated kernel $\mathbb{K} : T \times T \to \mathbb{R}$. We assume the kernel to be continuous and $T \subset \mathbb{R}^d$ to be compact. For concreteness, we think of $T = [0, 1]$ throughout this chapter, but any compact set of $\mathbb{R}^d$ suffices. For each $t \in T$, the function $R_t := \mathbb{K}(\cdot, t)$ belongs to the Hilbert space $\mathcal{H}$, and it acts as the *representer of evaluation*, meaning that $\langle f, R_t \rangle_{\mathcal{H}} = f(t)$ for all $f \in \mathcal{H}$.

The kernel $\mathbb{K}$ defines an integral operator $\mathcal{T}_{\mathbb{K}}$ on $L^2(T)$, mapping the function f to the function $g(s) = \int_T K(s, t) f(t) dt$. By the spectral theorem in Hilbert spaces, this operator can be associated with a sequence of eigenfunctions $\psi_k, k = 1, 2, \ldots$ in $\mathcal{H}$, orthogonal in $\mathcal{H}$ and orthonormal in $L^2(T)$, and a sequence of non-negative eigenvalues $\mu_1 \geq \mu_2 \geq \cdots$. Most useful for this chapter is the fact that any function $f \in \mathcal{H}$ has an expansion in terms of these eigenfunctions and eigenvalues, namely

$$f = \sum_{k=1}^{\infty} \sqrt{\mu_k} \alpha_k \psi_k \tag{5.1}$$

for some $(\alpha_k) \in \ell^2$. In terms of this expansion, we have the representations $\|f\|_{\mathcal{H}}^2 = \sum_{k=1}^{\infty} \alpha_k^2$ and $\|f\|_{L^2}^2 = \sum_{k=1}^{\infty} \mu_k \alpha_k^2$. Many of our results involve the decay rate of these eigenvalues: in particular, for some parameter $\alpha > 1/2$, we say that the kernel operator has eigenvalues with *polynomial-α decay* if there is a constant $c > 0$ such that

$$\mu_k \leq \frac{c}{k^{2\alpha}} \quad \text{for all } k = 1, 2, \ldots. \tag{5.2}$$

Let us consider an example to illustrate.

Example 4 (Sobolev class with smoothness $\alpha = 1$)**.** In the case $T = [0, 1]$ and $\alpha = 1$, we can consider the kernel function $\mathbb{K}(s, t) = \min\{s, t\}$. As discussed in Appendix 5.A, this kernel generates the class of functions

$$\mathcal{H} := \big\{ f \in L^2([0, 1]) \;\mid\; f(0) = 0, \; f \text{ absolutely continuous and } f' \in L^2([0, 1]) \big\}.$$

The class $\mathcal{H}$ is an RKHS with inner product $\langle f, g\rangle_{\mathcal{H}} = \int_0^1 f'(t)g'(t)dt$, and the ball $\mathbb{B}_{\mathcal{H}}(\rho)$ corresponds to a Sobolev space with smoothness $\alpha = 1$. The eigen-decomposition of the kernel integral operator is

$$\mu_k = \left[\frac{(2k-1)\pi}{2}\right]^{-2}, \quad \psi_k(t) = \sqrt{2}\sin\left(\mu_k^{-1/2}t\right), \quad k = 1, 2, \ldots. \tag{5.3}$$

Consequently, this class has polynomial decay with parameter $\alpha = 1$.

We note that there are natural generalizations of this example to $\alpha = 2, 3, \ldots$, corresponding to the Sobolev classes of α-times differentiable functions (e.g., see the books [47, 13]).

In this chapter, the operation of generalized sampling is defined in terms of a bounded linear operator $\Phi : \mathcal{H} \to \mathbb{R}^m$ on the Hilbert space. Its adjoint is a mapping $\Phi^* : \mathbb{R}^m \to \mathcal{H}$, defined by the relation $\langle \Phi f, a\rangle_{\mathbb{R}^m} = \langle f, \Phi^* a\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $a \in \mathbb{R}^m$. In order to compute a representation of the adjoint, we note that by the Riesz representation theorem, the j-th coordinate of this mapping—namely, $f \mapsto [\Phi f]_j$—can be represented as an inner product $\langle \phi_j, f\rangle_{\mathcal{H}}$, for some element $\phi_j \in \mathcal{H}$, and we can write

$$\Phi f = \begin{bmatrix}\langle \phi_1, f\rangle_{\mathcal{H}} & \langle \phi_2, f\rangle_{\mathcal{H}} & \cdots & \langle \phi_m, f\rangle_{\mathcal{H}}\end{bmatrix}^T. \tag{5.4}$$

Consequently, we have $\langle \Phi(f), a\rangle_{\mathbb{R}^m} = \sum_{j=1}^m a_j \langle \phi_j, f\rangle_{\mathcal{H}} = \langle \sum_{j=1}^m a_j \phi_j, f\rangle_{\mathcal{H}}$, so that for any $a \in \mathbb{R}^m$, the adjoint can be written as

$$\Phi^* a = \sum_{j=1}^m a_j \phi_j. \tag{5.5}$$

This adjoint operator plays an important role in our analysis.

5.1.2 Functional model and observations

Let $s_1 \geq s_2 \geq s_3 \geq \cdots \geq s_r > 0$ be a fixed sequence of positive numbers, and let $\{f_j^*\}_{j=1}^r$ be a fixed sequence of functions orthonormal in $L^2[0, 1]$. Consider a collection of n i.i.d. random functions $\{x_1, \ldots, x_n\}$, generated according to the model

$$x_i(t) = \sum_{j=1}^r s_j \beta_{ij} f_j^*(t), \quad \text{for } i = 1, \ldots, n, \tag{5.6}$$

where $\{\beta_{ij}\}$ are i.i.d. $N(0, 1)$ across all pairs (i, j). This model corresponds to a finite-rank instantiation of functional PCA, in which the goal is to estimate the span of the unknown eigenfunctions $\{f_j^*\}_{j=1}^r$. Typically, these eigenfunctions are assumed to satisfy certain

smoothness conditions; in this chapter, we model such conditions by assuming that the eigenfunctions belong to a reproducing kernel Hilbert space $\mathcal{H}$ embedded within $L^2[0,1]$; more specifically, they lie in some ball in $\mathcal{H}$,

$$\|f_j^*\|_{\mathcal{H}} \leq \rho, \quad j = 1, \ldots, r. \tag{5.7}$$

For statistical problems involving estimation of functions, the random functions might only be observed at certain times $(t_1, \ldots, t_m)$, such as in longitudinal data analysis, or we might collect only projections of each x_i in certain directions, such as in tomographic reconstruction. More concretely, in a *time-sampling model*, we observe m-dimensional vectors of the form

$$y_i = \begin{bmatrix} x_i(t_1) & x_i(t_2) & \cdots & x_i(t_m) \end{bmatrix}^T + \sigma_0 w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{5.8}$$

where $\{t_1, t_2, \ldots, t_m\}$ is a fixed collection of design points, and $w_i \in \mathbb{R}^m$ is a noise vector. Another observation model is the *basis truncation model* in which we observe the projections of f onto the first m basis functions $\{\psi_j\}_{j=1}^m$ of the kernel operator—namely,

$$y_i = \begin{bmatrix} \langle \psi_1, x_i \rangle_{L^2} & \langle \psi_2, x_i \rangle_{L^2} & \cdots & \langle \psi_m, x_i \rangle_{L^2} \end{bmatrix}^T + \sigma_0 w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{5.9}$$

where $\langle \cdot, \cdot \rangle_{L^2}$ represents the inner product in $L^2[0,1]$.

In order to model these and other scenarios in a unified manner, we introduce a linear operator Φ_m that maps any function x in the Hilbert space to a vector $\Phi_m(x)$ of m samples, and then consider the linear observation model

$$y_i = \Phi_m(x_i) + \sigma_m w_i, \qquad \text{for } i = 1, 2, \ldots, n. \tag{5.10}$$

This model (5.10) can be viewed as a functional analog of the spiked covariance models introduced by Johnstone [55, 58] as an analytically-convenient model for studying high-dimensional effects in classical PCA.

Both the time-sampling (5.8) and frequency truncation (5.9) models can be represented in this way, for appropriate choices of the operator Φ_m. Recall the representation (5.4) of Φ_m in terms of the functions $\{\phi_j\}_{j=1}^m$.

- For the time sampling model (5.8), we set $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$, so that by the reproducing property of the kernel, we have $\langle \phi_j, f \rangle_{\mathcal{H}} = f(t_j)/\sqrt{m}$ for all $f \in \mathcal{H}$, and $j = 1, 2, \ldots m$. With these choices, the operator Φ_m maps each $f \in \mathcal{H}$ to the m-vector of rescaled samples $\frac{1}{\sqrt{m}} \begin{bmatrix} f(t_1) & \cdots & f(t_m) \end{bmatrix}^T$. Defining the rescaled noise $\sigma_m = \frac{\sigma_0}{\sqrt{m}}$ yields an instantiation of the model (5.10) which is equivalent to time-sampling (5.8).

- For the basis truncation model (5.9), we set $\phi_j = \mu_j \psi_j$ so that the operator Φ maps each function $f \in \mathcal{H}$ to the vector of basis coefficients $\begin{bmatrix} \langle \psi_1, f \rangle_{L^2} & \cdots & \langle \psi_m, f \rangle_{L^2} \end{bmatrix}^T$. Setting $\sigma_m = \sigma_0$ then yields another instantiation of the model (5.10), this one equivalent to basis truncation (5.9).

A remark on notation before proceeding: in the remainder of the chapter, we use (Φ, σ) as shorthand notation for (Φ_m, σ_m), since the index m should be implicitly understood throughout our analysis.

In this chapter, we provide and analyze estimators for the r-dimensional eigen-subspace spanned by $\{f_j^*\}$, in both the sampled domain $\mathbb{R}^m$, and in the functional domain. To be more specific, for $j = 1, \ldots, r$, define the vectors $z_j^* := \Phi f_j^* \in \mathbb{R}^m$, and the subspaces

$$\mathfrak{Z}^* := \operatorname{span}\{z_1^*, \ldots, z_r^*\} \subset \mathbb{R}^m, \quad \text{and} \quad \mathfrak{F}^* := \operatorname{span}\{f_1^*, \ldots, f_r^*\} \subset \mathcal{H},$$

and let $\widehat{\mathfrak{Z}}$ and $\widehat{\mathfrak{F}}$ denote the corresponding estimators. In order to measure the performance of the estimators, we will use projection-based distances between subspaces. In particular, let $P_{\mathfrak{Z}^*}$ and $P_{\widehat{\mathfrak{Z}}}$ be orthogonal projection operators into $\mathfrak{Z}^*$ and $\widehat{\mathfrak{Z}}$, respectively, considered as subspaces of $\ell_2^m := (\mathbb{R}^m, \|\cdot\|_2)$. Similarly, let $P_{\mathfrak{F}^*}$ and $P_{\widehat{\mathfrak{F}}}$ be orthogonal projection operators into $\mathfrak{F}^*$ and $\widehat{\mathfrak{F}}$, respectively, considered as subspaces of $(\mathcal{H}, \|\cdot\|_{L^2})$. We are interested in bounding the deviations

$$\mathrm{d}_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) := |||P_{\widehat{\mathfrak{Z}}} - P_{\mathfrak{Z}^*}|||_{HS}, \quad \text{and} \quad \mathrm{d}_{HS}(\widehat{\mathfrak{F}}, \mathfrak{F}^*) := |||P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*}|||_{HS}, \tag{5.11}$$

where $|||\cdot|||_{HS}$ is the Hilbert-Schmidt norm of an operator (or matrix).

5.1.3 Approximation-theoretic quantities

One object that plays an important role in our analysis is the matrix $K := \Phi\Phi^* \in \mathbb{R}^{m \times m}$. From the form of the adjoint, it can be seen that $[K]_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{H}}$. For future reference, let us compute this matrix for the two special cases of linear operators considered thus far.

- For the time sampling model (5.8), we have $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$ for all $j = 1, \ldots, m$, and hence $[K]_{ij} = \frac{1}{m}\langle \mathbb{K}(\cdot, t_i), \mathbb{K}(\cdot, t_j) \rangle_{\mathcal{H}} = \frac{1}{m}\mathbb{K}(t_i, t_j)$, using the reproducing property of the kernel.
- For the basis truncation model (5.9), we have $\phi_j = \mu_j \psi_j$, and hence $[K]_{ij} = \langle \mu_i \psi_i, \mu_j \psi_j \rangle_{\mathcal{H}} = \mu_i \delta_{ij}$. Thus, in this special case, we have $K = \operatorname{diag}(\mu_1, \ldots, \mu_m)$.

In general, the matrix K is a type of Gram matrix, and so is symmetric and positive semidefinite. We assume throughout this chapter that the functions $\{\phi_j\}_{j=1}^m$ are linearly independent in $\mathcal{H}$, which implies that K is strictly positive definite. Consequently, it has a set of eigenvalues which can be ordered as

$$\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \ldots \geq \widehat{\mu}_m > 0. \tag{5.12}$$

Under this condition, we may use K to define a norm on $\mathbb{R}^m$ via $\|z\|_K^2 := z^T K^{-1} z$. Moreover, we have the following interpolation lemma, which is proved Appendix 5.B.1:

Lemma 29. *For any $f \in \mathcal{H}$, we have $\|\Phi f\|_K \leq \|f\|_{\mathcal{H}}$, with equality if and only if $f \in \mathrm{Ra}(\Phi^*)$. Moreover, for any $z \in \mathbb{R}^m$, the function $g = \Phi^* K^{-1} z$ has smallest Hilbert norm of all functions satisfying $\Phi g = z$, and is the unique function with this property.*

This lemma is useful in constructing a function-based estimator, as will be clarified in Section 5.2.

In our analysis of the functional error $\mathrm{d}_{HS}(\widehat{\mathfrak{F}}, \mathfrak{F}^*)$, a number of approximation-theoretic quantities play an important role. As a mapping from an infinite-dimensional space $\mathcal{H}$ to $\mathbb{R}^m$, the operator Φ has a non-trivial nullspace. Given the observation model (5.10), we receive no information about any component of a function f^* that lies within this nullspace. For this reason, we define the width of the nullspace in the L^2-norm, namely the quantity

$$N_m(\Phi) := \sup \left\{ \|f\|_{L^2}^2 \mid f \in \mathrm{Ker}(\Phi), \ \|f\|_{\mathcal{H}} \leq 1 \right\}. \tag{5.13}$$

In addition, the observation operator Φ induces a semi-norm on the space $\mathcal{H}$, defined by

$$\|f\|_\Phi^2 := \|\Phi f\|_2^2 = \sum_{j=1}^m [\Phi f]_j^2. \tag{5.14}$$

It is of interest to assess how well this semi-norm approximates the L^2-norm. Accordingly, we define the quantity

$$D_m(\Phi) := \sup_{\substack{f \in \mathrm{Ra}(\Phi^*) \\ \|f\|_{\mathcal{H}} \leq 1}} \left| \|f\|_\Phi^2 - \|f\|_{L^2}^2 \right|, \tag{5.15}$$

which measures the worst-case gap between these two (semi)-norms, uniformly over the Hilbert ball of radius one, restricted to the subspace of interest $\mathrm{Ra}(\Phi^*)$. Given knowledge of the linear operator Φ, the quantity $D_m(\Phi)$ can be computed in a relatively straightforward manner. In particular, recall the definition of the matrix K, and let us define a second matrix $\Theta \in \mathbb{S}_+^m$ with entries $\Theta_{ij} := \langle \varphi_i, \varphi_j \rangle_{L^2}$.

Lemma 30. *We have the equivalence*

$$D_m(\Phi) = |||K - K^{-1/2} \Theta K^{-1/2}|||_2, \tag{5.16}$$

where $|||\cdot|||_2$ denotes the ℓ_2-operator norm.

See Appendix 5.B.2 for the proof of this claim.

5.2 M-estimator and implementation

With this background in place, we now turn to the description of our M-estimator, as well as practical details associated with its implementation.

5.2.1 M-estimator

We begin with some preliminaries on notation, and our representation of subspaces. For each $j = 1, \ldots, m$, define the vector $z_j^* := \Phi f_j^*$, corresponding to the image of the function f_j^* under the observation operator. We let $\mathfrak{Z}^*$ denote the r-dimensional subspace of $\mathbb{R}^m$ spanned by $\{z_1^*, \ldots, z_r^*\}$, where $z_j^* = \Phi f_j^*$. Our initial goal is to construct an estimate $\widehat{\mathfrak{Z}}$, itself an r-dimensional subspace, of the unknown subspace $\mathfrak{Z}^*$.

We represent subspaces by elements of the Stiefel manifold $V_r(\mathbb{R}^m)$, which consists of of $m \times r$ matrices Z with orthonormal columns

$$V_r(\mathbb{R}^m) := \big\{ Z \in \mathbb{R}^{m \times r} \mid Z^T Z = I_r \big\}.$$

A given matrix Z acts as a representative of the subspace spanned by its columns, denoted by $\mathrm{col}(Z)$. For any $U \in V_r(\mathbb{R}^r)$, the matrix ZU also belongs to the Stiefel manifold, and since $\mathrm{col}(Z) = \mathrm{col}(ZU)$, we may call ZU a version of Z. We let $P_Z = ZZ^T \in \mathbb{R}^{m \times m}$ be the orthogonal projection onto $\mathrm{col}(Z)$. For two matrices $Z_1, Z_2 \in V_r(\mathbb{R}^m)$, we measure the distance between the associated subspaces via $\mathrm{d}_{HS}(Z_1, Z_2) := |||P_{Z_1} - P_{Z_2}|||_{HS}$, where $|||\cdot|||_{HS}$ is the Hilbert-Schmidt (or Frobenius) matirx norm.

Subspace-based estimator

With this notation, we now specify an M-estimator for the subspace $\mathfrak{Z}^* = \mathrm{span}\{z_1^*, \ldots, z_r^*\}$. Let us begin with some intuition. Given the n samples $\{y_1, \ldots, y_n\}$, let us define the $m \times m$ sample covariance matrix $\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n y_i y_i^T$. Given the observation model (5.10), a straightforward computation shows that

$$\mathbb{E}[\widehat{\Sigma}_n] = \sum_{j=1}^r s_j^2 z_j^* (z_j^*)^T + \sigma_m^2 I_m. \tag{5.17}$$

Thus, as n becomes large, we expect that the top r eigenvectors of $\widehat{\Sigma}_n$ might give a good approximation to $\mathrm{span}\{z_1^*, \ldots, z_r^*\}$. By the Courant-Fischer variational representation, these r eigenvectors can be obtained by maximizing the objective function

$$\langle\langle \widehat{\Sigma}_n, \; P_Z \rangle\rangle := \mathrm{tr}(\widehat{\Sigma}_n Z Z^T)$$

over all matrices $Z \in V_r(\mathbb{R}^m)$.

However, this approach fails to take into account the smoothness constraints that the vectors $z_j^* = \Phi f_j^*$ inherit from the smoothness of the eigenfunctions f_j^*. Since $\|f_j^*\|_{\mathcal{H}} \leq \rho$ by assumption, Lemma 29 implies that

$$\|z_j^*\|_K^2 = (z_j^*)^T K^{-1} z_j^* \;\leq\; \|f_j^*\|_{\mathcal{H}}^2 \;\leq\; \rho^2 \quad \text{for all } j = 1, 2, \ldots, r.$$

Consequently, if we define the matrix $Z^* := \begin{bmatrix} z_1^* & \cdots & z_r^* \end{bmatrix} \in \mathbb{R}^{m \times r}$, then it must satisfy the *trace smoothness condition*

$$\langle\langle K^{-1},\ Z^*(Z^*)^T \rangle\rangle = \sum_{j=1}^{r} (z_j^*)^T K^{-1} z_j^* \ \leq\ r\rho^2. \tag{5.18}$$

This calculation motivates the constraint $\langle\langle K^{-1},\ P_Z \rangle\rangle \leq 2r\rho^2$ in our estimation procedure.

Based on the preceding intuition, we are led to consider the optimization problem

$$\widehat{Z} \in \arg \max_{Z \in V_r(\mathbb{R}^m)} \Big\{ \langle\langle \widehat{\Sigma}_n,\ P_Z \rangle\rangle \ \mid\ \langle\langle K^{-1},\ P_Z \rangle\rangle \leq 2r\rho^2 \Big\}, \tag{5.19}$$

where we recall that $P_Z = ZZ^T \in \mathbb{R}^{m \times m}$. Given any optimal solution $\widehat{Z}$, we return the subspace $\widehat{\mathfrak{Z}} = \mathrm{col}(\widehat{Z})$ as our estimate of $\mathfrak{Z}^*$. As discussed at more length in Section 5.2.2, it is straightforward to compute $\widehat{Z}$ in polynomial time. The reader might wonder why we have included an additional factor of two in this trace smoothness condition. This slack is actually needed due to the potential infeasibility of the matrix Z^* for the program (5.19), which arises since the columns Z^* are not guaranteed to be orthonormal. As shown by our analysis, the additional slack allows us to find a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ that spans the same subspace as Z^*, and is also feasible for the program (5.19). More formally, we have:

Lemma 31. *Under condition (5.26b), there exists a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ such that*

$$\mathrm{Ra}(\widetilde{Z}^*) = \mathrm{Ra}(Z^*), \quad \text{and} \quad \langle\langle K^{-1},\ \widetilde{Z}^*(\widetilde{Z}^*)^T \rangle\rangle \leq 2r\rho^2. \tag{5.20}$$

See Appendix 5.B.3 for the proof of this claim.

The functional estimate $\widehat{\mathfrak{F}}$

Having obtained an estimate[1] $\widehat{\mathfrak{Z}} = \mathrm{span}\{\widehat{z}_1, \ldots, \widehat{z}_r\}$ of $\mathfrak{Z}^* = \mathrm{span}\{z_1^*, \ldots, z_r^*\}$, we now need to construct a r-dimensional subspace $\widehat{\mathfrak{F}}$ of the Hilbert space as an estimate of $\mathfrak{F}^* = \mathrm{span}\{f_1^*, \ldots, f_r^*\}$. We do so using the interpolation suggested by Lemma 29. For each $j = 1, \ldots, r$, define the function

$$\widehat{f}_j := \Phi^* K^{-1} \widehat{z}_j \ =\ \sum_{i=1}^{m} (K^{-1}\widehat{z}_j)_i\ \phi_i. \tag{5.21}$$

Since $K = \Phi\Phi^*$ by definition, this construction ensures that $\Phi \widehat{f}_j = \widehat{z}_j$. Moreover, Lemma 29 guarantees that $\widehat{f}_j$ has the minimal Hilbert norm (and hence is smoothest in a certain

[1]Here, $\{\widehat{z}_j\}_{j=1}^r \subset \mathbb{R}^m$ is any collection of vectors that span $\widehat{\mathfrak{Z}}$. As we are ultimately only interested in the resulting functional "subspace", it does not matter which particular collection we choose.

sense) over all functions that have this property. Finally, since Φ is assumed to be surjective (equivalently, K assumed invertible), $\Phi^* K^{-1}$ maps linearly independent vectors to linearly independent functions, and hence preserves dimension. Consequently, the space $\widehat{\mathfrak{F}} := \text{span}\{\widehat{f}_1, \ldots, \widehat{f}_r\}$ is an r-dimensional subspace of $\mathcal{H}$ which we take as our estimate of $\mathfrak{F}^*$.

5.2.2 Implementation details

In this section, we consider some practical aspects of implementing the M-estimator, and present some simulations to illustrate its qualitative properties. We begin by observing that once the subspace vectors $\{\widehat{z}_j\}_{j=1}^r$ have been computed, then it is straightforward to compute the function estimates $\{\widehat{f}_j\}_{j=1}^r$, as weighted combinations of the functions $\{\phi_j\}_{j=1}^m$. Accordingly, we focus our attention on solving the program (5.19).

On the surface, the problem (5.19) might appear non-convex, due to the Stiefel manifold constraint. However, it can be reformulated as a semidefinite program (SDP), a well-known class of convex programs, as clarified in the following:

Lemma 32. *The problem (5.19) is equivalent to solving the SDP*

$$\widehat{X} \in \arg\max_{X \succeq 0} \langle\langle \widehat{\Sigma}_n, \, X \rangle\rangle \quad \textit{such that } |||X|||_2 \leq 1, \, \text{tr}(X) = r, \, \textit{and } \langle\langle K^{-1}, \, X \rangle\rangle \leq 2r\rho^2, \quad (5.22)$$

for which there always exists an optimal rank r solution. Moreover, by Lagrangian duality, for some $\beta > 0$, the problem is equivalent to

$$\widehat{X} \in \arg\max_{X \succeq 0} \langle\langle \widehat{\Sigma}_n - \beta K^{-1}, \, X \rangle\rangle \quad \textit{such that } |||X|||_2 \leq 1 \textit{ and } \text{tr}(X) = r, \quad (5.23)$$

which can be solved by an eigendecomposition of $\widehat{\Sigma}_n - \beta K^{-1}$.

As a consequence, for a given Lagrange multiplier β, the regularized form of the estimator can be solved with the cost of solving an eigenvalue problem. For a given constraint $2r\rho^2$, the appropriate value of β can be found by a path-tracing algorithm, or a simple dyadic splitting approach.

In order to illustrate the estimator, we consider the time sampling model (5.8), with uniformly spaced samples, in the context of a first-order Sobolev RKHS (with kernel function $\mathbb{K}(s,t) = \min(s,t)$). The parameters of the model are taken to be $r = 4$, $(s_1, s_2, s_3, s_4) = (1, 0.5, 0.25, 0.125)$, $\sigma_0 = 1$, $m = 100$ and $n = 75$. The regularized form (5.23) of the estimator is applied and the results are shown in Fig. 5.1. The top row corresponds to the four "true" signals $\{f_j^*\}$, the leftmost being f_1^* (i.e. having the highest signal-to-noise ratio.) and the rightmost f_4^*. The subsequent rows show the corresponding estimates $\{\widehat{f}_j\}$, obtained using different values of β. The second, third and fourth rows correspond to $\beta = 0$, $\beta = 0.0052$ and $\beta = 0.83$.

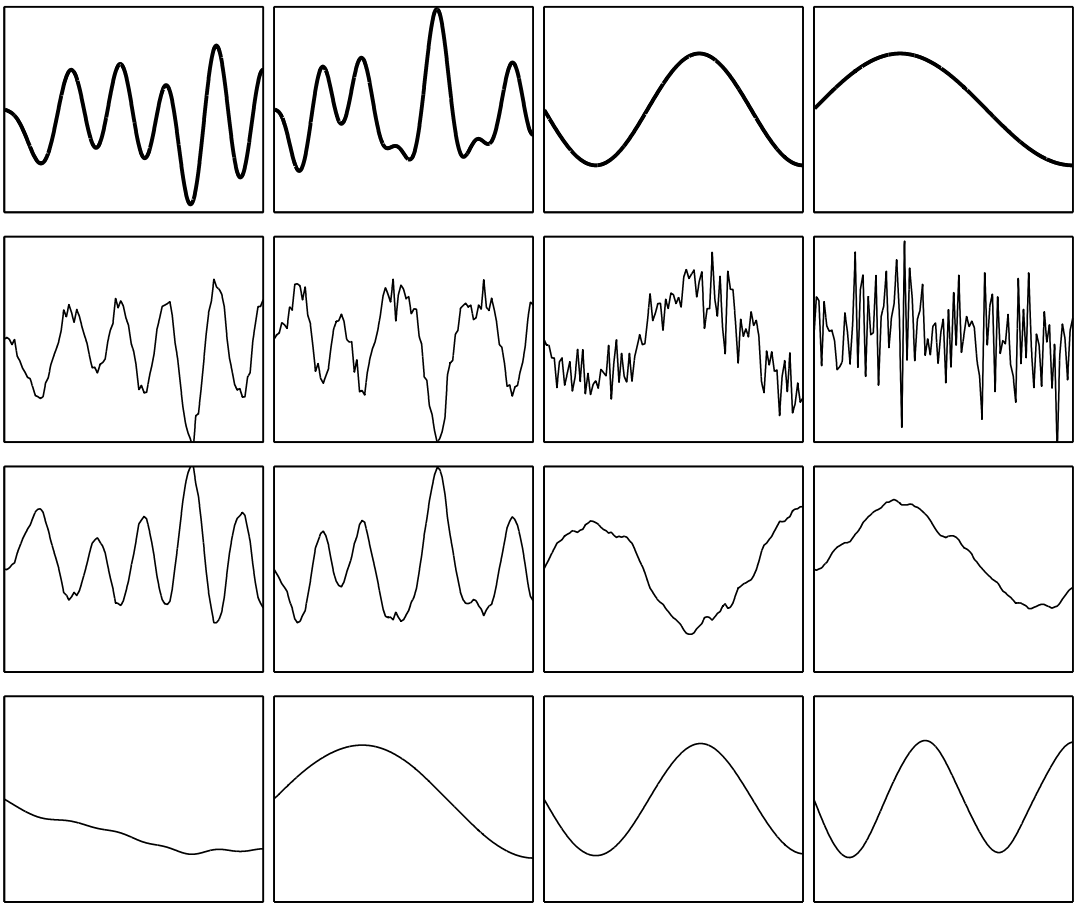


Figure 5.1: Regularized PCA for time sampling in first-order Sobolev RKHS. Top row shows, from left to right, plots of the $r = 4$ "true" principal components $f_1^*, \ldots, f_4^*$ with signal-to-noise ratios $s_1 = 1$, $s_2 = 0.5$, $s_3 = 0.25$ and $s_4 = 0.125$, respectively. The number of statistical and functional samples are $n = 75$ and $m = 100$. Subsequent rows show the corresponding estimators $\widehat{f}_1, \ldots, \widehat{f}_4$ obtained by applying the regularized form (5.23).

One observes that without regularization ($\beta = 0$), the estimates for two weakest signals (f_3^* and f_4^*) are poor. The case $\beta = 0.0052$ is roughly the one which achieves the minimum for the dual problem. One observes that the quality of the estimates of the signals, and in particular the weakest ones, are considerably improved. The optimal (oracle) value of β, that is the one which achieves the minimum error between $\{f_j^*\}$ and $\{\widehat{f}_j\}$, is $\beta = 0.0075$ in this problem. The corresponding estimates are qualitatively similar to those of $\beta = 0.0052$ and are not shown.

The case $\beta = 0.83$ shows the effect of over-regularization. It produces very smooth signals and although it fails to reveal f_1^* and f_2^*, it reveals highly accurate versions of f_3^* and f_4^*. It is also interesting to note that the smoothest signal, f_4^*, now occupies the position of the second (estimated) principal component. That is, the regularized PCA sees an effective signal-to-noise ratio which is influenced by smoothness. This suggests a rather practical

appeal of the method in revealing smooth signals embedded in noise. One can vary β from zero upward and if some patterns seem to be present for a wide range of β (and getting smoother as β is increased), one might suspect that they are indeed present in data but masked by noise.

5.3 Main results

We now turn to the statistical analysis of our estimators, in particular deriving high-probability upper bounds on the error of the subspace-based estimate $\widehat{\mathfrak{Z}}$, and the functional estimate $\widehat{\mathfrak{F}}$. In both cases, we begin by stating general theorems that applies to arbitrary linear operators Φ—Theorems 11 and 12 respectively—and then derive a number of corollaries for particular instantiations of the observation operator.

5.3.1 Subspace-based estimation rates (for $\widehat{\mathfrak{Z}}$)

We begin by stating high-probability upper bounds on the error $\mathrm{d}_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*)$ of the subspace-based estimates. Our rates are stated in terms of a function that involves the eigenvalues of the matrix $K = \Phi\Phi^* \in \mathbb{R}^m$, ordered as $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_m > 0$. Consider the function $\mathcal{F} : \mathbb{R}_+ \to \mathbb{R}_+$ given by

$$\mathcal{F}(t) := \Big[\sum_{j=1}^{m} \min\{t^2, r\rho^2 \widehat{\mu}_j\} \Big]^{1/2}. \tag{5.24}$$

As will be clarified in our proofs, this function provides a measure of the statistical complexity of the function class $\mathrm{Ra}(\Phi^*) = \{f \in \mathcal{H} \mid f = \sum_{j=1}^m a_j \phi_j \text{ for some } a \in \mathbb{R}^m\}$.

We require a few regularity assumptions. Define the quantity

$$C_m(f^*) := \max_{1 \leq i,j \leq r} \big| \langle f_i^*, f_j^* \rangle_\Phi - \delta_{ij} \big| \;=\; \max_{1 \leq i,j \leq r} \big| \langle z_i^*, z_j^* \rangle_{\mathbb{R}^m} - \delta_{ij} \big|, \tag{5.25}$$

which measures the departure from orthonormality of the vectors $z_j^* := \Phi f_j^*$ in $\mathbb{R}^m$. A straightforward argument using a polarization identity shows that $C_m(f^*)$ is upper bounded (up to a constant factor) by the uniform quantity $D_m(\Phi)$, as defined in equation (5.15). Recall that the random functions are generated according to the model $x_i = \sum_{j=1}^r s_j \beta_{ij} f_j^*$, where the signal strengths are ordered as $1 = s_1 \geq s_2 \geq \cdots \geq s_r > 0$, and that σ_m denotes the noise standard deviation in the observation model (5.10).

In terms of these quantities, we require the following assumptions:

$$\textbf{(A1)} \quad \frac{s_r^2}{s_1^2} \geq \frac{1}{2}, \quad \text{and} \quad \sigma_0^2 := \sup_m \sigma_m^2 \leq \kappa s_1^2, \tag{5.26a}$$

$$\textbf{(A2)} \quad C_m(f^*) \leq \frac{1}{2r}, \quad \text{and} \tag{5.26b}$$

$$\textbf{(A3)} \quad \frac{\sigma_m}{\sqrt{n}} \mathcal{F}(t) \leq \sqrt{\kappa} t \quad \text{for the same constant } \kappa \text{ as in (A1).} \tag{5.26c}$$

$$\textbf{(A4)} \quad r \leq \min\left\{\frac{m}{2}, \frac{n}{4}, \kappa \frac{\sqrt{n}}{\sigma_m}\right\}. \tag{5.26d}$$

Remarks: The first part of condition (A1) is to prevent the ratio s_r/s_1 from going to zero as the pair (m, n) increases, where the constant $1/2$ is chosen for convenience. Such a lower bound is necessary for consistent estimation of the eigen-subspace corresponding to $\{s_1, \ldots, s_r\}$. The second part of condition (A1), involving the constant κ, provides a lower bound on the signal-to-noise ratio s_r/σ_m. Condition (A2) is required to prevent degeneracy among the vectors $z_j^* = \Phi f_j^*$ obtained by mapping the unknown eigenfunctions to the observation space $\mathbb{R}^m$. (In the ideal setting, we would have $C_m(f^*) = 0$, but our analysis shows that the upper bound in (A2) is sufficient.) Condition (A3) is required so that the critical tolerance $\epsilon_{m,n}$ specified below is well-defined; as will be clarified, it is always satisfied for the time-sampling model, and holds for the basis truncation model whenever $n \geq m$. Condition (A4) is easily satisfied, since the RHS of (5.26d) goes to ∞ while we usually take r to be fixed. Our results, however, hold if r grows slowly with m and n subject to (5.26d).

Theorem 11. *Under conditions (A1)—(A3) for a sufficiently small constant κ, let $\epsilon_{m,n}$ be the smallest positive number satisfying the inequality*

$$\frac{\sigma_m}{\sqrt{n}} r^{3/2} \mathcal{F}(\epsilon) \leq \kappa \epsilon^2. \tag{5.27}$$

Then there are universal positive constants (c_0, c_1, c_2) such that

$$\mathbb{P}\big[\mathrm{d}_{HS}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0 \, \epsilon_{m,n}^2 \big] \geq 1 - \varphi(n, \epsilon_{m,n}), \tag{5.28}$$

where $\varphi(n, \epsilon_{m,n}) := c_1 \big\{ r^2 \exp\big(- c_2 r^{-3} \frac{n}{\sigma_m^2} (\epsilon_{m,n} \wedge \epsilon_{m,n}^2) \big) + r \exp(-\frac{n}{64}) \big\}$.

Theorem 11 is a general result, applying to an arbitrary bounded linear operator Φ. However, we can obtain a number of concrete results by making specific choices of this sampling operator, as we explore in the following sections.

Consequences for time-sampling

Let us begin with the time-sampling model (5.8), in which we observe the sampled functions

$$y_i = \begin{bmatrix} x_i(t_1) & x_i(t_2) & \ldots & x_i(t_m) \end{bmatrix}^T + \sigma_0 w_i, \quad \text{for } i = 1, 2, \ldots, m.$$

As noted earlier, this set-up can be modeled in our general setting (5.10) with $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$ and $\sigma_m = \sigma_0/\sqrt{m}$.

In this case, by the reproducing property of the RKHS, the matrix $K = \Phi\Phi^*$ has entries of the form $K_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \frac{\mathbb{K}(t_i,t_j)}{m}$. Letting $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \ldots \geq \widehat{\mu}_m > 0$ denote its ordered eigenvalues, we say that the kernel matrix K has polynomial-decay with parameter $\alpha > 1/2$ if there is a constant c such that $\widehat{\mu}_j \leq c\, j^{-2\alpha}$ for all $j = 1, 2, \ldots, m$. Since the kernel matrix K represents a discretized approximation of the kernel integral operator defined by $\mathbb{K}$, this type of polynomial decay is to be expected whenever the kernel operator has polynomial-α decaying eigenvalues. For example, the usual spline kernels that define Sobolev spaces have this type of polynomial decay [47]. In Appendix 5.A, we verify this property explicitly for the kernel $\mathbb{K}(s,t) = \min\{s,t\}$ that defines the Sobolev class with smoothness $\alpha = 1$.

For any such kernel, we have the following consequence of Theorem 11:

Corollary 5 (Achievable rates for time-sampling)**.** *Consider the case of a time-sampling operator Φ. In addition to conditions (A1) and (A2), suppose that the kernel matrix K has polynomial-decay with parameter $\alpha > 1/2$. Then we have*

$$\mathbb{P}\Big[\mathrm{d}^2_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0 \min\Big\{ \Big(\frac{\kappa_{r,\rho}\, \sigma_0^2}{mn}\Big)^{\frac{2\alpha}{2\alpha+1}},\, r^3\, \frac{\sigma_0^2}{n} \Big\} \Big] \geq 1 - \varphi(n,m), \tag{5.29}$$

where $\kappa_{r,\rho} := r^{3+\frac{1}{2\alpha}}\, \rho^{\frac{1}{\alpha}}$, *and* $\varphi(n,m) := c_1 \Big\{ \exp\Big(-c_2 \Big\{ \big(r^{-2}\rho^2 mn\big)^{\frac{1}{2\alpha+1}} \wedge m \Big\} \Big) + \exp(-n/64) \Big\}$.

Remarks: (a) Disregarding constant pre-factors not depending on the pair (m,n), Corollary 5 guarantees that solving the program (5.19) returns a subspace estimate $\widehat{\mathfrak{Z}}$ such that

$$\mathrm{d}^2_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \precsim \min\big\{ (mn)^{-\frac{2\alpha}{2\alpha+1}},\, n^{-1} \big\} \qquad \text{with high probability as } (m,n) \text{ increase.}$$

Depending on the scaling of the number of time samples m relative to the number of functional samples n, either term in this upper bound can be the smallest (and hence active) one. For instance, it can be verified that whenever $m \geq n^{\frac{1}{2\alpha}}$, then the first term is smallest, so that we achieve the rate $\mathrm{d}^2_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \precsim (mn)^{-\frac{2\alpha}{2\alpha+1}}$. The appearance of the term $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ is quite natural, as it corresponds to the minimax rate of a non-parameteric regression problem with smoothness α, based on m samples each of variance n^{-1}. Later, in Section 5.3.3, we provide results guaranteeing that this scaling is minimax optimal under reasonable conditions on the choice of sample points (in particular, see Theorem 13(a)).

(b) To be clear, although the bound (5.29) allows for the possibility that the error is of order *lower than* n^{-1}, we note that the probability with which the guarantee holds includes a term of the order $\exp(-n/64)$. Consequently, in terms of expected error, we cannot guarantee a rate faster than n^{-1}.

Proof. We need to bound the critical value $\epsilon_{m,n}$ defined in the theorem statement (5.27). Define the function $\mathcal{G}^2(t) := \sum_{j=1}^{m} \min\{\widehat{\mu}_j, t^2\}$, and note that $\mathcal{F}(t) = \sqrt{r}\rho\,\mathcal{G}(\frac{t}{\sqrt{r}\rho})$ by construction. Under the assumption of polynomial-α eigendecay, we have

$$\mathcal{G}^2(t) \leq \int_0^\infty \min\{cx^{-2\alpha}, t^2\}\, dx,$$

and some algebra then shows that $\mathcal{G}(t) \precsim t^{1-1/(2\alpha)}$. Disregarding constant factors, an upper bound on the critical $\epsilon_{m,n}$ can be obtained by solving the equation

$$\epsilon^2 = \frac{\sigma_m}{\sqrt{n}}\, r^{3/2}\, \sqrt{r}\rho\big(\frac{\epsilon}{\sqrt{r}\rho}\big)^{1-1/(2\alpha)}.$$

Doing so yields the upper bound $\epsilon^2 \precsim \big[\frac{\sigma_m^2}{n} r^3 (\sqrt{r}\rho)^{\frac{1}{\alpha}}\big]^{\frac{2\alpha}{2\alpha+1}}$. Otherwise, we also have the trivial upper bound $\mathcal{F}(t) \leq \sqrt{m}\, t$, which yields the alternative upper bound $\varepsilon_{m,n} \precsim \big(\frac{m\sigma_m^2}{n} r^3\big)^{1/2}$. Recalling that $\sigma_m = \sigma_0/\sqrt{m}$ and combining the pieces yields the claim. Notice that this last (trivial) bound on $\mathcal{F}(t)$ implies that condition (A3) is always satisfied for the time-sampling model. □

Consequences for basis truncation

We now turn to some consequences for the basis truncation model (5.9).

Corollary 6 (Achievable rates for basis truncation)**.** *Consider a basis truncation operator Φ in a Hilbert space with polynomial-α decay. Under conditions (A1), (A2) and $m \leq n$, we have*

$$\mathbb{P}\big[\, \mathrm{d}_{HS}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0 \, \big(\frac{\kappa_{r,\rho}\, \sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\big] \geq 1 - \varphi(n, m), \tag{5.30}$$

where $\kappa_{r,\rho} := r^{3+\frac{1}{2\alpha}}\, \rho^{\frac{1}{\alpha}}$, and $\varphi(n,m) := c_1\big\{\exp\big(-c_2\big(r^{-2}\rho^2 n\big)^{\frac{1}{2\alpha+1}}\big) + \exp(-n/64)\big\}$.

Proof. We note that as long as $m \leq n$, condition (A3) is satisfied, since $\frac{\sigma_m}{\sqrt{n}}\mathcal{F}(t) \leq \sigma_0\sqrt{\frac{m}{n}}t \leq \sigma_0 t$. The rest of the proof follows that of Corollary 5, noting that in the last step we have $\sigma_m = \sigma_0$ for the basis truncation model. □

5.3.2 Function-based estimation rates (for $\widehat{\mathfrak{F}}$)

As mentioned earlier, given the consistency of $\widehat{\mathfrak{Z}}$, the consistency of $\widehat{\mathfrak{F}}$ is closely related to approximation properties of the semi-norm $\|\cdot\|_\Phi$ induced by Φ, and in particular how closely it approximates the L^2-norm. These approximation-theoretic properties are captured in part by the nullspace width $N_m(\Phi)$ and defect $D_m(\Phi)$ defined earlier in equations (5.13)

and (5.15) respectively. In addition to these previously defined quantities, we require bounds on the following global quantity

$$R_\Phi(\epsilon;\nu) := \sup\left\{\|f\|_{L^2}^2 \mid \|f\|_{\mathcal{H}}^2 \leq \nu^2,\ \|f\|_\Phi^2 \leq \epsilon^2\right\}. \tag{5.31}$$

A general upper bound on this quantity is of the form

$$R_\Phi(\epsilon;\nu) \leq c_1\epsilon^2 + \nu^2 S_m(\Phi). \tag{5.32}$$

In fact, it is not hard to show that such a bound exists with $c_1 = 2$ and $S_m(\Phi) = 2(D_m(\Phi) + N_m(\Phi))$ using the decomposition $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$. However, this bound is not sharp. One can show that in most cases of interest, $S_m(\Phi)$ is of the order of $N_m(\Phi)$. There are different assumptions which can lead to such a scaling of the remainder term $S_m(\Phi)$. We refer to Chapter 4 for a general approach. Here, we give a simple condition, which will be verified for the first-order Sobolev RKHS, namely,

$$\textbf{(D1)} \quad \Theta \preceq c_0 K^2$$

for a positive constant c_0.

Lemma 33. *Under (D1), the bound (5.32) holds with $c_1 = 2c_0$ and $S_m(\Phi) = 2N_m(\Phi)$.*

This Lemma is proved in Appendix 5.B.4.

Theorem 12. *Suppose that condition (A1) holds, $D_m(\Phi) \leq \frac{1}{4r\rho^2} \leq 1$ and $N_m(\Phi) \leq 1$. Then there is a constant $\kappa'_{r,\rho}$ such that*

$$\mathrm{d}_{HS}^2(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho}\left\{\epsilon_{m,n}^2 + S_m(\Phi) + [D_m(\Phi)]^2\right\} \tag{5.33}$$

with the same probability as in Theorem 11.

As with Theorem 11, this is a generally applicable result, stated in abstract form. By specializing it to different sampling models, we can obtain concrete rates, as illustrated in the following sections.

For future reference, let us introduce another condition regulating the approximation property of $\|\cdot\|_\Phi$ relative to $\|\cdot\|_{L^2}$. Consider the matrix $\Psi \in \mathbb{R}^{m\times m}$ with entries $\Psi_{ij} := \langle\psi_i, \psi_j\rangle_\Phi$. Since the eigenfunctions are orthogonal in L^2, the deviation of Ψ from the identity measures how well the inner product defined by Φ approximates the L^2-inner product over the first m eigenfunctions of the kernel operator. In particular, we require an upper bound of the form

$$\textbf{(D2)} \qquad \lambda_{\max}(\Psi) \leq c_1,$$

for some universal constant $c_1 > 0$. Condition (D2) will be used when deriving minimax lower bounds.

Consequences for time-sampling

We begin by returning to the case of the time sampling model (5.8), where $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$. In this case, condition (D1) needs to be verified by some calculations. For instance, as shown in Appendix 5.A, in the case of the Sobolev kernel with smoothness $\alpha = 1$ (namely, $\mathbb{K}(s,t) = \min\{s,t\}$), we are guaranteed that (D1) holds with $c_0 = 1$, whenever the samples $\{t_j\}$ are chosen uniformly over $[0,1]$; hence, by Lemma 33, $S_m(\Phi) = 2N_m(\Phi)$. Moreover, in the case of uniform sampling, we expect that the nullspace width $N_m(\Phi)$ is upper bounded by μ_{m+1}, so will be proportional to $m^{-2\alpha}$ in the case of a kernel operator with polynomial-α decay. This is verified in Chapter 4 (up to a logarithmic factor) for the case of the first-order Sobolev kernel. In Appendix 5.A, we also show that, for this kernel, $[D_m(\Phi)]^2$ is of the order $m^{-2\alpha}$, that is, of the same order as $N_m(\Phi)$.

Corollary 7. *Consider the basis truncation model (5.9) with uniformly spaced samples, and assume condition (D1) holds and that $N_m(\Phi) + [D_m(\Phi)]^2 \precsim m^{-2\alpha}$. Then the M-estimator returns a subspace estimate $\widehat{\mathfrak{F}}$ such that*

$$\mathrm{d}^2_{HS}(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho}\Big\{ \min\Big\{\Big(\frac{\sigma_0^2}{nm}\Big)^{\frac{2\alpha}{2\alpha+1}}, \ \frac{\sigma_0^2}{n}\Big\} + \frac{1}{m^{2\alpha}}\Big\} \tag{5.34}$$

with the same probability as in Corollary 5.

In this case, there is an interesting trade-off between the bias or approximation error terms which is of order $m^{-2\alpha}$ and the estimation error. An interesting transition occurs at the point when $m \succsim n^{\frac{1}{2\alpha}}$, at which:

- the bias term $m^{-2\alpha}$ becomes of the order n^{-1}, so that it is no longer dominant, and
- for the two terms in the estimation error, we have the ordering

$$(mn)^{-\frac{2\alpha}{2\alpha+1}} \leq \big(n^{1+\frac{1}{2\alpha}}\big)^{-\frac{2\alpha}{2\alpha+1}} \ = \ n^{-1}.$$

Consequently, we conclude that the scaling $m = n^{\frac{1}{2\alpha}}$ is the minimal number of samples such that we achieve an overall bound of the order n^{-1} in the time-sampling model. In Section 5.3.3, we will see that these rates are minimax-optimal.

Consequences for basis truncation

For the basis truncation operator Φ, we have $\Theta = K^2 = \mathrm{diag}(\mu_1^2, \ldots, \mu_m^2)$ so that condition (D1) is satisfied trivially with $c_0 = 1$. Moreover, Lemma 30 implies $D_m(\Phi) = 0$. In addition, a function $f = \sum_{j=1}^{\infty} \sqrt{\mu_j} a_j \psi_j$ satisfies $\Phi f = 0$ if and only if $a_1 = a_2 = \cdots = a_m = 0$, so that

$$N_m(\Phi) = \sup\big\{ \|f\|_{L^2}^2 \ \big| \ \|f\|_{\mathcal{H}} \leq 1, \ \Phi f = 0\big\} \ = \ \mu_{m+1}.$$

Consequently, we obtain the following corollary of Theorem 12:

Corollary 8. *Consider the basis truncation model (5.9) with a kernel operator that has polynomial-α decaying eigenvalues. Then the M-estimator returns a function subspace estimate $\widehat{\mathfrak{F}}$ such that*

$$\mathrm{d}_{HS}^2(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho}\big\{\big(\frac{\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}} + \frac{1}{m^{2\alpha}}\big\} \tag{5.35}$$

with the same probability as in Corollary 6.

By comparison to Corollary 7, we see that the trade-offs between (m, n) are very different for basis truncation. In particular, there is *no interaction* between the number of functional samples m and the number of statistical samples n. Increasing m only reduces the approximation error, whereas increasing n only reduces the estimation error. Moreover, in contrast to the time sampling model of Corollary 7, it is impossible to achieve the fast rate n^{-1}, regardless of how we choose the pair (m, n). In Section 5.3.3, we will also see that the rates given in Corollary 8 are minimax optimal.

5.3.3 Lower bounds

We now turn to lower bounds on the minimax risk, demonstrating the sharpness of our achievable results in terms of their scaling with (m, n). In order to do so, it suffices to consider the simple model with a single functional component $f^* \in \mathbb{B}_{\mathcal{H}}(1)$, so that we observe $y_i = \beta_{i1}\,\Phi_m(f^*) + \sigma_m w_i$ for $i = 1, 2, \ldots, n$, where $\beta_{i1} \sim N(0,1)$ are i.i.d. standard normal variates. The minimax risk in the Φ-semi-norm is given by

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\delta^2;\ \|\cdot\|_\star) := \inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{P}_{f^*}\big[\|\widetilde{f} - f^*\|_\star^2 \geq \delta^2\big]. \tag{5.36}$$

where the function f^* ranges over the unit ball $\mathbb{B}_{\mathcal{H}}(1) = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$ of some Hilbert space, and $\widetilde{f}$ ranges over measurable functions of the data matrix $(y_1, y_2, \ldots, y_n) \in \mathbb{R}^{m \times n}$.

Theorem 13 (Lower bounds for $\|\widetilde{f} - f^*\|_\Phi^2$)**.** *Suppose that the kernel matrix K has eigenvalues with polynomial-α decay and (A1) holds.*

(a) For the time-sampling model,

$$\mathcal{M}_{m,n}^{\mathcal{H}}\Big(C \min\big\{\big(\frac{\sigma_0^2}{mn}\big)^{\frac{2\alpha}{2\alpha+1}},\ \frac{\sigma_0^2}{n}\big\};\ \|\cdot\|_\Phi\Big)\ \geq\ \frac{1}{2}. \tag{5.37}$$

(b) For the frequency-truncation model, with $m \geq (c_0 n)^{\frac{1}{2\alpha+1}}$:

$$\mathcal{M}_{m,n}^{\mathcal{H}}\Big(C\,\big(\frac{\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}};\ \|\cdot\|_\Phi\Big)\ \geq\ \frac{1}{2}. \tag{5.38}$$

Note that part (a) of Theorem 13 shows that the rates obtained in Corollary 7 for the case of time-sampling are minimax optimal. Similarly, comparing part (b) of the theorem to Corollary 8, we conclude that the rates obtained for frequency truncation model are minimax optimal for $n \in [m, c_1 m^{2\alpha+1}]$. The case of $n > c_1 m^{2\alpha+1}$ is not of practical interest as will become clear shortly as a consequence of the next theorem.

We now turn to lower bounds on the minimax risk in the $\|\cdot\|_{L^2}$ norm—namely

$$\mathcal{M}^{\mathcal{H}}_{m,n}(\delta^2;\ \|\cdot\|_{L^2}) := \inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{P}_{f^*}\big[\|\widetilde{f} - f^*\|^2_{L^2} \geq \delta^2\big]. \tag{5.39}$$

Theorem 14 (Lower bounds for $\|\widetilde{f} - f^*\|^2_{L^2}$)**.** *Suppose that condition (D2) holds, and the operator associated with kernel function $\mathbb{K}$ of the reproducing kernel Hilbert space $\mathcal{H}$ has eigenvalues with polynomial-α-decay.*

(a) For the time-sampling model, the minimax error is lower bounded as

$$\mathcal{M}^{\mathcal{H}}_{m,n}\Big(C\big\{\min\big\{\big(\frac{\sigma_0^2}{mn}\big)^{\frac{2\alpha}{2\alpha+1}},\ \frac{\sigma_0^2}{n}\big\} + \big(\frac{1}{m}\big)^{2\alpha}\big\};\ \|\cdot\|_{L^2}\Big)\ \geq\ \frac{1}{2}. \tag{5.40}$$

(b) For the frequency-truncation model, the minimax error is lower bounded as

$$\mathcal{M}^{\mathcal{H}}_{m,n}\Big(C\big\{\big(\frac{\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}} + \big(\frac{1}{m}\big)^{2\alpha}\big\};\ \|\cdot\|_{L^2}\Big)\ \geq\ \frac{1}{2}. \tag{5.41}$$

Verifying condition (D2) requires, in general, some calculations in the case of time-sampling model. It is verified for uniform time-sampling for the first-order Sobolev RKHS in Appendix 5.A. For the frequency-truncation, (D2) always holds trivially since $\Psi = I_m$. By this theorem, the L^2 convergence rates of Corollary 7 and 8 are minimax optimal. Also note that due to the presence of the approximation term $m^{-2\alpha}$ in (5.41), the Φ-norm term $n^{\frac{2\alpha}{2\alpha+1}}$ is only dominant when $m \geq c_2 n^{\frac{1}{2\alpha+1}}$ implying that this is the interesting regime for Theorem 13(b).

5.4 Proof of subspace-based rates

We now turn to the proofs of the results involving the error $\mathrm{d}_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*)$ between the estimated $\widehat{\mathfrak{Z}}$ and true subspace $\mathfrak{Z}^*$. We begin by proving Theorem 11, and then turn to its corollaries.

5.4.1 Preliminaries

We begin with some preliminaries before proceeding to the heart of the proof. Let us first introduce some convenient notation. Consider the $n \times m$ matrices

$$Y := \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T, \quad \text{and} \quad W := \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}^T,$$

corresponding to the observation matrix Y and noise matrix W respectively. In addition, we define the matrices $B := (\beta_{ij}) \in \mathbb{R}^{n\times r}$ and $S := \mathrm{diag}(s_1, \ldots, s_r) \in \mathbb{R}^{r\times r}$. Recalling that $Z^* := (z_1^*, \ldots, z_r^*) \in \mathbb{R}^{m\times r}$, the observation model (5.10) can be written in the matrix form $Y = B(Z^*S)^T + \sigma_m W$. Moreover, let us define the matrices $\overline{B} := \frac{B^TB}{n} \in \mathbb{R}^{r\times r}$ and $\overline{W} := \frac{W^TB}{n} \in \mathbb{R}^{m\times r}$. Using this notation, some algebra shows that the associated sample covariance $\widehat{\Sigma}_n := \frac{1}{n}Y^TY$ can be written in the form

$$\widehat{\Sigma}_n = \underbrace{Z^*S\overline{B}S(Z^*)^T}_{\Gamma} + \Delta_1 + \Delta_2, \tag{5.42}$$

where $\Delta_1 := \sigma_m\left[\overline{W}S(Z^*)^T + Z^*S\overline{W}^T\right]$ and $\Delta_2 := \sigma_m^2\frac{W^TW}{n}$.

Lemma 31, proved in Appendix 5.B.3 shows the existence of a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ such that $\mathrm{Ra}(\widetilde{Z}^*) = \mathrm{Ra}(Z^*)$. As discussed earlier, due to the nature of the Steifel manifold, there are many versions of this matrix $\widetilde{Z}^*$, and also of any optimal solution matrix $\widehat{Z}$, obtained via right multiplication with an orthogonal matrix. For the subsequent arguments, we need to work with a particular version of $\widetilde{Z}^*$ (and $\widehat{Z}$) that we describe here.

Now let us fix some convenient versions of $\widetilde{Z}^*$ and $\widehat{Z}$. As a consequence of CS decomposition, as long as $r \le m/2$, there exist orthogonal matrices $U, V \in \mathbb{R}^{r\times r}$ and an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$, such that

$$Q^T\widetilde{Z}^*U = \begin{pmatrix} I_r \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad Q^T\widehat{Z}V = \begin{pmatrix} \widehat{C} \\ \widehat{S} \\ 0 \end{pmatrix}, \tag{5.43}$$

where $\widehat{C} = \mathrm{diag}(\widehat{c}_1, \cdots, \widehat{c}_r)$ and $\widehat{S} = \mathrm{diag}(\widehat{s}_1, \cdots, \widehat{s}_r)$ such that $1 \ge \widehat{s}_1 \ge \cdots \ge \widehat{s}_r \ge 0$ and $\widehat{C}^2 + \widehat{S}^2 = I_r$. (See Bhatia [15], Theorem VII.1.8, for details on this decomposition.) In the analysis to follow, we work with $\widetilde{Z}^*U$ and $\widehat{Z}V$ instead of $\widetilde{Z}^*$ and $\widehat{Z}$. To avoid extra notation, from now on, we will use $\widetilde{Z}^*$ and $\widehat{Z}$ for these new versions, which we refer to as *properly aligned*. With this choice, we may assume $U = V = I_r$ in the CS decomposition (5.43).

The following lemma isolates some useful properties of properly aligned subspaces:

Lemma 34. *Let $\widetilde{Z}^*$ and $\widehat{Z}$ be properly aligned, and define the matrices*

$$\widehat{P} := P_{\widehat{Z}} - P_{\widetilde{Z}^*} = \widehat{Z}\widehat{Z}^T - \widetilde{Z}^*(\widetilde{Z}^*)^T, \quad \text{and} \quad \widehat{E} := \widehat{Z} - \widetilde{Z}^*. \tag{5.44}$$

In terms of the CS decomposition (5.43), we have

$$\|\widehat{E}\|_{HS} \le \|\widehat{P}\|_{HS}, \tag{5.45a}$$

$$(\widetilde{Z}^*)^T(P_{\widetilde{Z}^*} - P_{\widehat{Z}})\widetilde{Z}^* = \widehat{S}^2, \quad \text{and} \tag{5.45b}$$

$$d_F^2(\widehat{Z}, \widetilde{Z}^*) = \|P_{\widetilde{Z}^*} - P_{\widehat{Z}}\|_{HS}^2 = 2\|\widehat{S}^2\|_{HS}^2 + 2\|\widehat{C}\widehat{S}\|_{HS}^2 = 2\sum_k \widehat{s}_k^2(\widehat{s}_k^2 + \widehat{c}_k^2) = 2\,\mathrm{tr}(\widehat{S}^2). \tag{5.45c}$$

Proof. From the CS decomposition (5.43), we have $\widetilde{Z}^*(\widetilde{Z}^*)^T - \widehat{Z}(\widehat{Z})^T = Q\begin{pmatrix} \widehat{S}^2 & -\widehat{C}\widehat{S} & 0 \\ -\widehat{S}\widehat{C} & -\widehat{S}^2 & 0 \\ 0 & 0 & 0 \end{pmatrix} Q^T$, from which relations (5.45b) and (5.45c) follow. From the decomposition (5.43) and the proper alignment condition $U = V = I_r$, we have

$$\begin{aligned} \|\widehat{E}\|_{HS}^2 = \|Q^T(\widehat{Z} - \widetilde{Z}^*)\|_{HS}^2 &= \|I_r - \widehat{C}\|_{HS}^2 + \|\widehat{S}\|_{HS}^2 \\ &= 2\sum_{i=1}^r (1 - \widehat{c}_i) \le 2\sum_{i=1}^r (1 - \widehat{c}_i^2) = 2\sum_{i=1}^r \widehat{s}_i^2 = \|\widehat{P}\|_{HS}^2 \end{aligned} \tag{5.46}$$

where we have used the relations $\widehat{C}^2 + \widehat{S}^2 = I_r$, $\widehat{c}_i \in [0, 1]$, and $2\operatorname{tr}(\widehat{S}^2) = \|P_{\widetilde{Z}^*} - P_{\widehat{Z}}\|_{HS}^2$. □

5.4.2 Proof of Theorem 11

Using the notation introduced in Lemma 34, our goal is to bound the Frobenius norm $\|\widehat{P}\|_{HS}$. Without loss of generality we will assume $s_1 = 1$ throughout. Recalling the definition (5.42) of the random matrix Δ, the following inequality plays a central role in the proof:

Lemma 35. *Under condition (A1) and $s_1 = 1$, we have*

$$\|\widehat{P}\|_{HS}^2 \le 128\, \langle\!\langle \widehat{P},\ \Delta_1 + \Delta_2 \rangle\!\rangle \tag{5.47}$$

with probability at least $1 - \exp(-n/32)$.

Proof. We use the shorthand notation $\Delta = \Delta_1 + \Delta_2$ for the proof. Since $\widetilde{Z}^*$ is feasible and $\widehat{Z}$ is optimal for the program (5.19), we have the basic inequality $\langle\!\langle \widehat{\Sigma}_n,\ P_{\widetilde{Z}^*} \rangle\!\rangle \le \langle\!\langle \widehat{\Sigma}_n,\ P_{\widehat{Z}} \rangle\!\rangle$. Using the decomposition $\widehat{\Sigma} = \Gamma + \Delta$ and rearranging yields the inequality

$$\langle\!\langle \Gamma,\ P_{\widetilde{Z}^*} - P_{\widehat{Z}} \rangle\!\rangle \le \langle\!\langle \Delta,\ P_{\widehat{Z}} - P_{\widetilde{Z}^*} \rangle\!\rangle. \tag{5.48}$$

From the definition (5.42) of Γ and $Z^* = \widetilde{Z}^* R$, the left-hand side of the inequality (5.48) can be lower bounded as

$$\begin{aligned} \langle\!\langle \Gamma,\ P_{\widetilde{Z}^*} - P_{\widehat{Z}} \rangle\!\rangle &= \langle\!\langle \overline{B},\ SR^T(\widetilde{Z}^*)^T(P_{\widetilde{Z}^*} - P_{\widehat{Z}})\widetilde{Z}^* RS \rangle\!\rangle \\ &= \operatorname{tr} \overline{B} S R^T \widehat{S}^2 R S \\ &\ge \lambda_{\min}(\overline{B})\lambda_{\min}(S^2)\lambda_{\min}(R^T R)\operatorname{tr}(\widehat{S}^2) \end{aligned}$$

where we have used (5.88) and (5.89) of Appendix 5.I, several times. We note that $\lambda_{\min}(S^2) = s_r^2 \ge \frac{1}{2}$ and $\lambda_{\min}(R^T R) \ge \frac{1}{2}$ provided $rC_m(f^*) \ge \frac{1}{2}$; see equation (5.68). To bound the minimum eigenvalue of $\overline{B}$, let $\gamma_{\min}(B)$ denote the minimum singular value of the $n \times r$ Gaussian matrix B. The following concentration inequality is well-known (cf. [34, 65]):

$$\mathbb{P}\big[\gamma_{\min}(B) \le \sqrt{n} - \sqrt{r} - t\big] \le \exp(-t^2/2), \quad \text{for all } t > 0.$$

Since $\lambda_{\min}(\overline{B}) = \gamma^2_{\min}(B/\sqrt{n})$, we have that $\lambda_{\min}(\overline{B}) \geq (1 - \sqrt{r/n} - t)^2$ with probability at least $1 - \exp(-nt^2/2)$. Assuming $r/n \leq \frac{1}{4}$ and setting $t = \frac{1}{4}$, we get $\lambda_{\min}(\overline{B}) \geq \frac{1}{16}$ with probability at least $1 - \exp(-n/32)$. Putting the pieces together yields the claim. □

The inequality (5.47) reduces the problem of bounding $\|\widehat{P}\|^2_{HS}$ to the sub-problem of studying the random variable $\langle\langle \widehat{P},\ \Delta_1 + \Delta_2 \rangle\rangle$. Based on Lemma 35, our next step is to establish an inequality (holding with high probability) of the form

$$\langle\langle \widehat{P},\ \Delta_1 + \Delta_2 \rangle\rangle \leq c_1 \Big\{ \frac{\sigma_m}{\sqrt{n}}\, r^{3/2}\, \mathcal{F}(\|\widehat{E}\|_{HS}) + \kappa \|\widehat{E}\|^2_{HS} + \epsilon^2_{m,n} \Big\}, \tag{5.49}$$

where c_1 is some universal constant, κ is the constant in condition (A1), and $\epsilon_{m,n}$ is the critical radius from Theorem 11. Doing so is a non-trivial task: both matrices $\widehat{P}$ and Δ are random and depend on one another, since the subspace $\widehat{Z}$ was obtained by optimizing a random function depending on Δ. Consequently, our proof of the bound (5.49) involves deriving a uniform law of large numbers for a certain matrix class.

Suppose that the bound (5.49) holds, and that the subspaces $\widetilde{Z}^*$ and $\widehat{Z}$ are properly aligned. Lemma 34 implies that $\|\widehat{E}\|_{HS} \leq \|\widehat{P}\|_{HS}$, and since $\mathcal{F}$ is a non-decreasing function, the inequality (5.49) combined with Lemma 35 implies that

$$\big(1 - 128\kappa c_1\big) \|\widehat{P}\|^2_{HS} \leq c_1 \Big\{ \frac{\sigma_m}{\sqrt{n}}\, r^{3/2}\, \mathcal{F}(\|\widehat{P}\|_{HS}) + \epsilon^2_{m,n} \Big\},$$

from which the claim follows as long as κ is suitably small (for instance, $\kappa \leq \frac{c_1}{256}$ suffices). Accordingly, in order to complete the proof of Theorem 11, it remains to prove the bound (5.49), and the remainder of our work is devoted to this goal. Given the linearity of trace, we can bound the terms $\langle\langle \widehat{P},\ \Delta_1 \rangle\rangle$ and $\langle\langle \widehat{P},\ \Delta_2 \rangle\rangle$ separately.

Bounding $\langle\langle \widehat{P},\ \Delta_1 \rangle\rangle$

Let $\{\overline{z}_j\}$, $\{\widetilde{z}^*_j\}$ and $\{\widehat{e}_j\}$ and $\{\overline{w}_j\}$ denote the columns of $\widehat{Z}$, $\widetilde{Z}^*$, $\widehat{E}$ and $\overline{W}$, respectively, where we recall the definitions of these quantities from equation (5.42) and Lemma 34. Note that $\overline{w}_j = n^{-1} \sum_{i=1}^n w_i \beta_{ij}$. In Appendix 5.C.1, we show that

$$\langle\langle \widehat{P},\ \Delta_1 \rangle\rangle \leq \sqrt{6}\, \sigma r^{3/2} \max_{j,k} |\langle \overline{w}_k, \widehat{e}_j \rangle| + \sqrt{\frac{3}{2}}\, \sigma r\, \|\widehat{E}\|^2_{HS} \max_{j,k} |\langle \overline{w}_j, \widetilde{z}^*_k \rangle|. \tag{5.50}$$

Consequently, we need to obtain bounds on quantites of the form $|\langle \overline{w}_j, v \rangle|$, where the vector v is either fixed (e.g., $v = \widetilde{z}^*_j$) or random (e.g., $v = \widehat{e}_j$). The following lemmas provide us with the requisite bounds:

Lemma 36. *We have*

$$\max_{j,k}\ \sigma r^{3/2}\ |\langle \overline{w}_k, \widehat{e}_j \rangle| \ \leq\ C \left\{ \frac{\sigma}{\sqrt{n}} r^{3/2} \mathcal{F}(\|\widehat{E}\|_{HS}) + \kappa \|\widehat{E}\|^2_{HS} + \kappa \epsilon^2_{m,n} \right\}$$

with probability at least $1 - c_1 r \exp(-\kappa^2 r^{-3} n \frac{\epsilon^2_{m,n}}{2\sigma^2}) - r \exp(-n/64)$.

Lemma 37. *We have*

$$\mathbb{P}\big[\max_{j,k} \sigma r|\overline{w}_k^T \widetilde{z}_j^*| \le \sqrt{6}\kappa\big] \ge 1 - r^2\exp(-\kappa^2 r^{-2} n/2\sigma^2).$$

See Appendix 5.C.2 and 5.C.3, respectively, for the proofs of these claims.

Bounding $\langle\langle \widehat{P},\ \Delta_2 \rangle\rangle$

Recalling the definition (5.42) of Δ_2 and using linearity of the trace, we obtain

$$\langle\langle \widehat{P},\ \Delta_2 \rangle\rangle = \frac{\sigma^2}{n}\sum_{j=1}^{r}\Big\{(\overline{z}_j)^T W^T W \overline{z}_j - (\widetilde{z}_j^*)^T W^T W \widetilde{z}_j^*\Big\}.$$

Since $\widehat{e}_j = \overline{z}_j - \widetilde{z}_j^*$, we have

$$\begin{aligned}\langle\langle \widehat{P},\ \Delta_2 \rangle\rangle &= \sigma^2\sum_{j=1}^{r}\Big\{2(\widetilde{z}_j^*)^T\Big(\frac{1}{n}W^TW - I_r\Big)\widehat{e}_j + \frac{1}{n}\|W\widehat{e}_j\|_2^2 + 2(\widetilde{z}_j^*)^T\widehat{e}_j\Big\}\\ &\le \sigma^2\sum_{j=1}^{r}\Big\{\underbrace{2\,(\widetilde{z}_j^*)^T\Big(\frac{1}{n}W^TW - I_r\Big)\widehat{e}_j}_{T_1(\widehat{e}_j;\widetilde{z}_j^*)} + \underbrace{\frac{1}{n}\|W\widehat{e}_j\|_2^2}_{T_2(\widehat{e}_j)}\Big\},\end{aligned} \tag{5.51}$$

where we have used the fact that $2\sum_j(\widetilde{z}_j^*)^T\widehat{e}_j = 2\sum_j[(\widetilde{z}_j^*)^T\overline{z}_j - 1] = 2\sum_j(\widehat{c}_j - 1) = -\|\widehat{E}\|_{HS}^2 \le 0$.

The following lemmas provide high probability bounds on the terms T_1 and T_2.

Lemma 38. *We have the upper bound*

$$\sigma^2\sum_{j=1}^{r} T_1(\widehat{e}_j;\widetilde{z}_j^*) \le C\Big\{\sigma_0\frac{\sigma}{\sqrt{n}}r\mathcal{F}(\|\widehat{E}\|_{HS}) + \kappa\|\widehat{E}\|_{HS}^2 + \kappa\epsilon_{m,n}^2\Big\}$$

with probability $1 - c_2\exp(-\kappa^2 r^{-2} n\,\frac{\epsilon_{m,n}\wedge\epsilon_{m,n}^2}{16\sigma^2}) - r\exp(-n/64)$.

Lemma 39. *We have the upper bound* $\sigma^2\sum_{j=1}^{r} T_2(\widehat{e}_j) \le C\kappa\big\{\|\widehat{E}\|_{HS}^2 + \epsilon_{m,n}^2\big\}$ *with probability at least* $1 - c_3\exp(-\kappa^2 r^{-2} n\,\epsilon_{m,n}^2/2\sigma^2)$.

See Appendices 5.C.4 and 5.C.5, respectively, for the proofs of these claims.

5.5 Proof of functional rates

We now turn to the proof of Theorem 12, which provides upper bounds on the estimation error in the function domain. As in the proof of Theorem 11, let $\widehat{Z} = (\widehat{z}_1, \cdots, \widehat{z}_r) \in V_r(\mathbb{R}^m)$ and $\widetilde{Z}^* = (\widetilde{z}_1^*, \cdots, \widetilde{z}_r^*) \in V_r(\mathbb{R}^m)$ represent the subspaces $\widehat{\mathfrak{Z}}$ and $\mathfrak{Z}^*$ respectively, and assume that they are properly aligned (see Lemma 34). For $j = 1, \ldots, m$, define $\widehat{g}_j := \Phi^* K^{-1} \widehat{z}_j$ and $g_j^* := \Phi^* K^{-1} \widetilde{z}_j^*$. Let $\{\widehat{h}_j\}_{j=1}^r$ be *any* basis of $\widehat{\mathfrak{F}}$, orthonormal in L^2, and similarly, let $\{h_j^*\}_{j=1}^r$ be any orthonormal basis of $\mathfrak{F}^*$. Our goal is to bound the Hilbert-Schmidt norm $\|\!| P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*} |\!\|_{\mathrm{HS}}^2$. In order to do so, we first observe that

$$\|\!| P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*} |\!\|_{\mathrm{HS}}^2 \leq 2 \sum_{j=1}^r \|\widehat{h}_j - h_j^*\|_{L^2}^2, \tag{5.52}$$

so that it suffices to upper bound $\sum_{j=1}^r \|\widehat{h}_j - h_j^*\|_{L^2}^2$. We relate this quantity to the functions $\widehat{g}_j$ and g_j^* via the elementary inequality

$$\|\widehat{h}_j - h_j^*\|_{L^2}^2 \leq 4\big\{\|\widehat{g}_j - g_j^*\|_{L^2}^2 + \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 + \|g_j^* - h_j^*\|_{L^2}^2\big\}. \tag{5.53}$$

The remainder of our proof is focused on obtaining suitable upper bounds on each of these three terms.

We begin by bounding the first term $\|\widehat{g}_j - g_j^*\|_{L^2}^2$. Recall the definitions of $R_\Phi(\epsilon; \nu)$ and $S_m(\Phi)$ and their relation via inequality (5.32). We exploit the inequality in the following way: suppose that we can show that

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_\Phi^2 \leq A^2, \quad \text{and} \quad \sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{\mathcal{H}}^2 \leq B^2. \tag{5.54}$$

Let $S(A, B) = \{(a, b) \in \mathbb{R}^r \times \mathbb{R}^r \mid \sum_{j=1}^r a_j^2 \leq A^2, \ \sum_{j=1}^r b_j^2 \leq B^2\}$. We may then conclude that

$$\begin{aligned}
\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{L^2}^2 &\leq \sup_{(a,b) \in S(A,B)} \sum_{j=1}^r R_\Phi(a_j; b_j) \\
&\overset{(i)}{\leq} \sup_{(a,b) \in S(A,B)} \sum_{j=1}^r \big\{c_1 a_j^2 + b_j^2 S_m(\Phi)\big\} \\
&= c_1 A^2 + B^2 S_m(\Phi).
\end{aligned} \tag{5.55}$$

where inequality (i) follows by repeated application of inequality (5.32).

It remains to establish upper bounds of the form (5.54). By definition, we have $\widehat{g}_j - g_j^* \in \mathrm{Ra}(\Phi^*)$ and $\Phi(\widehat{g}_j - g_j^*) = \widehat{z}_j - \widetilde{z}_j^*$. Recalling the norm $\|a\|_K^2 := a^T K^{-1} a$, we note that the

matrices $\widehat{Z}$ and $\widetilde{Z}^*$ satisfy the trace smoothness condition $\sum_{j=1}^r \|\widehat{z}_j\|_K^2 = \langle\langle K^{-1}, ZZ^T\rangle\rangle \leq 2r\rho^2$, and hence

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{\mathcal{H}}^2 = \sum_{j=1}^r \|\widehat{z}_j - \widetilde{z}_j^*\|_K^2 \leq 2\sum_{j=1}^r (\|\widehat{z}_j\|_K^2 + \|\widetilde{z}_j^*\|_K^2) \leq \underbrace{8\,r\rho^2}_{B^2}$$

Furthermore, recalling that $\|f\|_\Phi = \|\Phi f\|_2$, we have

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_\Phi^2 = \sum_{j=1}^r \|\widehat{z}_j - \widetilde{z}_j^*\|_2^2 = \|\widehat{Z} - \widetilde{Z}^*\|_{HS}^2 \leq \underbrace{\|P_{\widehat{Z}} - P_{\widetilde{Z}^*}\|_{HS}^2}_{A^2}$$

Consequently, by the bound (5.55) with $A^2 = \|P_{\widehat{Z}} - P_{\widetilde{Z}^*}\|_{HS}^2$ and $B^2 = 8r\rho^2$, we conclude that

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{L^2}^2 \leq c_1 \|P_{\widehat{\mathfrak{Z}}} - P_{\mathfrak{Z}^*}\|_{HS}^2 + 8r\rho^2 S_m(\Phi) \tag{5.56}$$

We now need to bound the remaining two terms in the decomposition (5.53). In order to do so, we exploit the freedom in choosing the orthonormal families $\{\widehat{h}_j\}_{j=1}^r$ and $\{h_j^*\}_{j=1}^r$. By appropriate choices, we obtain the following results:

Lemma 40. *There exists an orthonormal basis $\{\widehat{h}_j\}_{j=1}^r$ of $\widehat{\mathfrak{F}}$ for which*

$$\sum_{j=1}^r \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = 2r^2\rho^4\, D_m^2(\Phi). \tag{5.57}$$

Lemma 41. *There exists an orthonormal basis $\{h_j^*\}_{j=1}^r$ of $\mathfrak{F}^*$ for which*

$$\sum_{j=1}^r \|h_j^* - g_j^*\|_{L^2}^2 \leq c_2\, r^2 C_m^2(f^*) + 6r\rho^2 S_m(\Phi). \tag{5.58}$$

As these proofs are more technical and lengthy, we defer them to Appendices 5.D.1 and 5.D.2 respectively.

Combining all of the pieces, we obtain the upper bound

$$\|P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*}\|_{HS}^2 \leq c_3\big\{\|P_{\widehat{\mathfrak{Z}}} - P_{\mathfrak{Z}^*}\|_{HS}^2 + r^2\rho^4 D_m^2(\Phi) + r^2 C_m^2(f^*) + r\rho^2 S_m(\Phi)\big\}. \tag{5.59}$$

By using polarization identity and decomposition $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$, one can show that

$$C_m(f^*) \leq \kappa_\rho''\big(D_m(\Phi) + N_m(\Phi)\big), \tag{5.60}$$

when $N_m(\Phi) \leq 1$. (See Appendix 5.B.5 for more details.) Using this inequality and noting that $S_m(\Phi) \geq N_m(\Phi) \geq [N_m(\Phi)]^2$ when $N_m(\Phi) \leq 1$, the bound (5.59) can be simplified to the form given in Theorem 12.

5.6 Proof of minimax lower bounds

We now turn to the proofs of the minimax lower bounds stated in Theorems 13 and 14. We begin with some preliminary results that apply to both proofs.

5.6.1 Preliminary results

Our proofs proceed via a standard reduction from estimation to multi-way hypothesis testing (e.g., [107, 104]). In particular, let $\{f^1, \ldots, f^M\}$ be an δ-packing set of $\mathbb{B}_{\mathcal{H}}(1)$ in a given norm $\|\cdot\|_\star$. (For our proofs, this norm will be either $\|\cdot\|_\Phi$ or $\|\cdot\|_{L^2}$.) Given such a packing set, it is known that the minimax error in the norm $\|\cdot\|_\star$ can be lower bounded by

$$\mathcal{M}^{\mathcal{H}}_{m,n}(\frac{\delta^2}{4};\ \|\cdot\|_\star) := \inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{P}_{f^*}\Big[\|\widetilde{f} - f^*\|_\star^2 \geq \frac{\delta^2}{4}\Big] \geq 1 - \frac{I(y; f) + \log 2}{\log M}. \tag{5.61}$$

where $y = (y_1, \ldots, y_n) \in \mathbb{R}^{m \times n}$ is the observation matrix, and f is a random function uniformly distributed over the packing set. The quantity $I(y; f)$ is the mutual information between y and f, and a key step in the proofs is obtaining good upper bounds on it.

Let $\mathbb{P}_f$ (respectively $\mathbb{P}_g$) be the distribution of y given that $f^* = f$ (respectively $f^* = g$). The mutual information $I(y; f)$ is intimately related to the Kullback-Leibler (KL) divergence between $\mathbb{P}_f$ and $\mathbb{P}_g$, which is given by

$$D(\mathbb{P}_f \,\|\, \mathbb{P}_g) = \int p_f(y) \log \frac{p_f(y)}{p_g(y)} dy, \tag{5.62}$$

where p_f and p_g are the densities with respect to Lebesgue measure. Our analysis requires upper bounds on this KL divergence, as provided by the following lemma:

Lemma 42. *Assume that $\|f\|_\Phi = \|g\|_\Phi$. Then the Kullback-Leibler divergence is upper bounded as $D(\mathbb{P}_f \,\|\, \mathbb{P}_g) \leq \frac{n\|f-g\|_\Phi^2}{\sigma_m^2}$.*

See Appendix 5.E.1 for the proof.

5.6.2 Proof of Theorem 13

We are now ready to begin the proof of our lower bounds on the minimax error in the (semi)-norm $\|\cdot\|_\Phi$. In order to leverage the lower bound (5.61), we need to have control on the packing and covering numbers in this norm:

Lemma 43 (Packing/covering in $\|\cdot\|_\Phi$-norm)**.** *Suppose that the kernel matrix K has polynomial-α decay.*

(a) Suppose that $m \leq (c_0 n)^{\frac{1}{2\alpha}}$ *for some constant* c_0*. Then there exists a collection of functions* $\{f^1, \ldots, f^M\}$ *contained in* $\mathbb{B}_{\mathcal{H}}(1)$ *such that* $M \geq 4^m$*, and*

$$\|f^i\|_\Phi^2 = \frac{\sigma_0^2}{16n} \quad \text{and} \quad \|f^i - f^j\|_\Phi^2 \geq \frac{\sigma_0^2}{64n}, \quad \text{for all } i \neq j \in \{1, 2, \ldots, M\}.$$

(b) The covering number of the set $\mathrm{Ra}(\Phi^*) \cap \mathbb{B}_{\mathcal{H}}(1)$ *in the* $\|\cdot\|_\Phi$*-norm is upper bounded as*

$$\log N_\Phi(\epsilon) \leq c_1 (1/\epsilon)^{\frac{1}{\alpha}}. \tag{5.63}$$

In the other direction, if $\epsilon^2 \geq \frac{\kappa_1}{m^{2\alpha}}$ *for some constant* $\kappa_1 > 0$*, then the packing number is lower bounded as*

$$\log M_\Phi(\epsilon) \geq c_2 (1/\epsilon)^{\frac{1}{\alpha}}. \tag{5.64}$$

The proof of this auxiliary result is given in Appendix 5.E.2; here we use it to prove Theorem 13.

The case of time sampling

Let us consider part (a) first. Recall that in this case $\sigma_m = \sigma_0/\sqrt{m}$. First, supposing that $m \leq (c_0 n)^{\frac{1}{2\alpha}}$, we establish a lower bound of the order $1/n$ on the minimax risk. (Note that if this upper bound on m holds, then the $1/n$ term is the minimum of the two terms in Theorem 13(a).) Let $\{f^1, \ldots, f^M\}$ be the collection of functions from part Lemma 43(a). Using the Fano bound (5.61) and the inequality $\log M \geq m \log 4$, we obtain

$$\mathcal{M}_{m,n}^{\mathcal{H}}\big(\frac{\sigma_0^2}{256n};\ \|\cdot\|_\Phi\big) \geq 1 - \frac{I(y; f) + \log 2}{m \log 4},$$

where y is the matrix of observations $(y_1, \ldots, y_n) \in \mathbb{R}^{m \times n}$, and the random variable f ranges uniformly over the packing set $\{f^1, \ldots, f^M\}$. By the convexity of the Kullback-Leibler divergence, we have

$$I(y; f) \leq \frac{1}{\binom{M}{2}} \sum_{i \neq j} D(\mathbb{P}_{f^i} \,\|\, \mathbb{P}_{f^j}) \overset{(i)}{\leq} \frac{1}{\binom{M}{2}} \sum_{i \neq j} \frac{n\|f^i - f^j\|_\Phi^2}{\sigma_m^2} \overset{(ii)}{\leq} \frac{n}{\sigma_m^2} \frac{\sigma_0^2}{4n} = \frac{m}{4},$$

where inequality (i) follows from Lemma 42, and inequality (ii) follows from the packing construction in Lemma 43(a). Consequently, we have

$$\frac{I(y; f) + \log 2}{m \log 4} \leq \frac{m/4 + \log 2}{m \log 4} \leq \frac{1}{2}$$

for all $m \geq 2$, which completes the proof.

Otherwise, we may assume that $m \geq (c_0 n)^{\frac{1}{2\alpha}}$, under which assumption we prove the lower bound involving the term of order $(mn)^{-\frac{2\alpha}{2\alpha+1}}$. (Note that this lower bound on m holds, then the $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ term is the minimum of the two terms in Theorem 13(a).) Let $\delta^2 = c_3(\frac{\sigma_0^2}{mn})^{\frac{2\alpha}{2\alpha+1}}$ for some $c_3 > 0$ to be chosen. Since $m \geq (c_0 n)^{\frac{1}{2\alpha}}$ by assumption, some algebra shows that $\delta^2 \geq \frac{\kappa_1}{m^{2\alpha}}$, so that the lower bound on the packing number from Lemma 43(b) may be applied. Combining this lower bound with the Fano inequality, we obtain

$$\mathcal{M}_{m,n}^{\mathcal{H}}\big(\frac{\delta^2}{4};\ \|\cdot\|_\Phi\big) \geq 1 - \frac{I(y;f) + \log 2}{c_2(1/\delta)^{1/\alpha}}.$$

By the upper bounding technique of Yang and Barron [104], the mutual information $I(y;f)$ is upper bounded by $\inf_{\nu>0}\{\nu^2 + \log N_{\mathrm{KL}}(\nu)\}$, where N_{KL} is the covering number in the square-root Kullback-Leibler (pseudo)-metric. By Lemma 42 and Lemma 43(b), we have $N_{\mathrm{KL}}(\nu) \leq c_1\big(\frac{\sigma_0}{\sqrt{nm}}\nu\big)^{1/\alpha}$. Re-parameterizing in terms of $\epsilon^2 = \frac{\sigma_0^2}{nm}\nu^2$, we obtain the upper bound

$$I(y;f) \leq \inf_{\epsilon>0}\big\{\frac{nm}{\sigma_0^2}\epsilon^2 + c_1(1/\epsilon)^{1/\alpha}\big\} \ \leq\ \big(\frac{1}{\epsilon_*}\big)^{1/\alpha},$$

where $\epsilon_*^2 = c_4\big(\frac{\sigma_0^2}{nm}\big)^{\frac{2\alpha}{2\alpha+1}}$ for some constant c_4. Consequently, we have

$$R := \frac{I(y;f) + \log 2}{c_2(1/\delta)^{1/\alpha}} \leq \frac{\big(\frac{1}{\epsilon_*}\big)^{1/\alpha} + \log 2}{(1/\delta)^{1/\alpha}}.$$

Note that δ and ϵ_* are of the same order. By choosing the pre-factor c_3 sufficiently small, we can thus guarantee that the ratio R is less than $1/2$, from which the claim follows.

The case of frequency truncation

Recall that in this case $\sigma_m = \sigma_0$. Since by assumption $m \geq (c_0 n)^{\frac{1}{2\alpha+1}}$, letting $\delta^2 = c_3\big(\frac{\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$, we have $\delta^2 \geq \frac{\kappa_1}{m^{2\alpha}}$ after some algebra. Hence, the lower bound on the packing number from Lemma 43(b) may be applied. Moreover, we have $N_{\mathrm{KL}}(\nu) \leq c_1\big(\frac{\sigma_0}{\sqrt{n}}\nu\big)^{1/\alpha}$. The rest of the proof follows that of part (a).

5.6.3 Proof of Theorem 14

On one hand, no method can estimate to an accuracy greater than $\frac{1}{2}N_m(\Phi)$. Indeed, whatever estimator $\widetilde{f}$ is used, the adversary can always choose some function f^* such that $\Phi(f^*) = 0$, and $\|\widetilde{f} - f^*\|_{L^2} \geq \frac{1}{2}N_m(\Phi)$. To see this, note that on one hand, if $\|\widetilde{f}\|_{L^2} \geq \frac{1}{2}N_m(\Phi)$, then the adversary can set $f^* = 0$. On the other hand, if $\|\widetilde{f}\|_{L^2} < \frac{1}{2}N_m(\Phi)$, then for any $\delta > 0$, adversary can choose a function $f^* \in \mathrm{Ker}(\Phi) \cap \mathbb{B}_{\mathcal{H}}(1)$ such that $\|f^*\|_{L^2} > N_m(\Phi) - \delta$, by

definition (5.13) of $N_m(\Phi)$. We then have $\|f^* - \widetilde{f}\|_{L^2} \geq \|f^*\|_{L^2} - \|\widetilde{f}\|_{L^2} > \frac{1}{2}N_m(\Phi) - \delta$ where we let $\delta \to 0$. In addition, it follows from the theory of optimal widths in Hilbert spaces [78] that $N_m(\Phi) \gtrsim \mu_{m+1}$, thereby establishing the $m^{-2\alpha}$ lower bound for a kernel operator with polynomial-α decay.

Let us now prove the lower bound involving $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ in part (a). This term is the smaller of the two terms involved in the minimum, when $m \geq n^{\frac{1}{2\alpha}}$; this is the only case we need to consider as for $m < n^{\frac{1}{2\alpha}}$, the minimum is n^{-1} which is dominated by the term $m^{-2\alpha}$. We introduce the shorthand $\Psi_1^m = \text{span}\{\psi_1, \ldots, \psi_m\} \cap \mathbb{B}_{\mathcal{H}}(1)$, corresponding to the intersection of the unit ball $\mathbb{B}_{\mathcal{H}}(1)$ with the m-dimensional subspace of $\mathcal{H}$ spanned by the first m eigenfunctions of the kernel. For this proof, our packing/covering constructions take place entirely within this set. The following lemma, proved in Appendix 5.E.3, provides bounds on these packing and covering numbers:

Lemma 44 (Packing/covering in $\|\cdot\|_{L^2}$-norm)**.** *There is a universal constant $c_1 > 0$ such that*

$$\log N_{L^2}(\epsilon; \Psi_1^m) \leq c_1(1/\epsilon)^{\frac{1}{\alpha}}. \tag{5.65}$$

In the other direction, if $\epsilon^2 \geq \frac{\kappa_1}{m^{2\alpha}}$ for some constant $\kappa_1 > 0$, there is a universal constant $c_2 > 0$ such that

$$\log M_{L^2}(\epsilon; \Psi_1^m) \geq c_2(1/\epsilon)^{\frac{1}{\alpha}}. \tag{5.66}$$

Based on this lemma, proving a $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ bound is relatively straightforward, once again using Fano's inequality (5.61). Choosing $\delta^2 = c_3(\frac{\sigma_0^2}{mn})^{\frac{2\alpha}{2\alpha+1}}$ for a constant c_3 to be specified, we construct a δ-packing in $\|\cdot\|_{L^2}$ norm, of size M such that $\log M \geq c_2(1/\delta)^{1/\alpha}$. As in the proof of Theorem 13, we upper bound the mutual information in terms of the covering number in the $\|\cdot\|_\Phi$. By condition (D2), this covering number is upper bounded (up to constant factors) by the covering number in the $\|\cdot\|_{L^2}$-norm. To see this, note that for any $f \in \Psi_1^m \cap \mathbb{B}_{L^2}(\varepsilon)$, we have $f = \sum_{j=1}^m a_j\psi_j$, with $\sum_{j=1}^m a_j^2/\mu_j \leq 1$ and $\sum_{j=1}^m a_j^2 \leq \varepsilon^2$. Then, condition (D2) implies $\|f\|_\Phi^2 = \langle a, \Psi a\rangle \leq c_1\|a\|_2^2 \leq 2\varepsilon^2$, that is $f \in \Psi_1^m \cap \mathbb{B}_\Phi(\sqrt{c_1}\varepsilon)$. Finally, by Lemma 44, the $\|\cdot\|_{L^2}$ covering number scales as $(1/\epsilon)^{1/\alpha}$, so that the same calculations as before yield the $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ rate as claimed.

The proof of part (b) is similar. We only need to consider the case $m \geq n^{\frac{1}{2\alpha+1}}$. The rest of the argument follows by taking $\delta^2 = c_3\big(\frac{\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$ and recalling that $\sigma_m = \sigma_0$ in this case.

5.7 Discussion

We studied the problem of sampling for functional PCA from a functional-theoretic viewpoint. The principal components were assumed to lie in some Hilbert subspace $\mathcal{H}$ of L^2,

usually a RKHS, and the sampling operator, a bounded linear map $\Phi : \mathcal{H} \to \mathbb{R}^m$. The observation model was taken to be the output of Φ plus some Gaussian noise. The two main examples of Φ considered were time sampling, $[\Phi f]_j = f(t_j)$, and (generalized) frequency truncation $[\Phi f]_j = \langle \psi_j, f \rangle_{L^2}$. We showed that it is possible to recover the subspace spanned by the original components, by applying a regularized version of PCA in $\mathbb{R}^m$ followed by simple linear mapping back to function space. The regularization involved the "trace-smoothness condition" (5.18) based on the matrix $K = \Phi\Phi^*$ whose eigendecay influenced the rate of convergence in $\mathbb{R}^m$.

We obtained the rates of convergence for the subspace estimators both in the discrete domain, $\mathbb{R}^m$, and the function domain, L^2. As examples, for the case of a RKHS $\mathcal{H}$ for which both the kernel integral operator and the kernel matrix K have polynomial-α eigendecay (i.e., $\mu_j \asymp \widehat{\mu}_j \asymp j^{-2\alpha}$), the following rates in HS-projection distance for subspaces in the function domain were worked out in details:

time sampling	frequency truncation
$\left(\frac{1}{mn}\right)^{\frac{2\alpha}{2\alpha+1}} + \left(\frac{1}{m}\right)^{2\alpha}$	$\left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + \left(\frac{1}{m}\right)^{2\alpha}$

The two terms in each rate can be associated, respectively, with the estimation error (due to noise) and approximation error (due to having finite samples of an infinite dimensional object). Both rates exhibit a trade-off between the number of statistical samples (n) and that of functional samples (m). The two rates are qualitatively different: the two terms in the time sampling case interact to give an overall fast rate of n^{-1} for the optimal trade-off $m \asymp n^{\frac{1}{2\alpha}}$, while there is no interaction between the two terms in the frequency truncation; the optimal trade-off gives an overall rate of $n^{-\frac{2\alpha}{2\alpha+1}}$, a characteristics of nonparametric problems. Finally, these rates were shown to be minimax optimal.

Appendix 5.A A special kernel

In this appendix, we examine a simple reproducing kernel Hilbert space, corresponding to a Sobolev or spline class with smoothness $\alpha = 1$. We provide expressions for various approximation-theoretic quantities appearing in our results, such as $D_m(\Phi)$, $N_m(\Phi)$ and Ψ. Further background on the calculations given here can be found in the paper [4].

Let us consider the time sampling model (5.8) with uniformly spaced points $t_j = j/m$ for $j = 1, \ldots, m$. Elementary calculations show that $K = \left(m^{-1}\mathbb{K}(t_i, t_j)\right) = \frac{1}{m^2} LL^T$, where $L \in \mathbb{R}^{m \times m}$ is lower triangular with all the nonzero entries equal 1. It can be shown that the eigenvalues of K are given by $\widehat{\mu}_k := \left\{4m^2 \sin^2\left(\frac{\mu_k^{-1/2}}{2m+1}\right)\right\}^{-1}$ for $k = 1, 2, \ldots, m$. Using the inequalities $\frac{2}{\pi}x \le \sin(x) \le x$, for $0 \le x \le \pi/2$, we have

$$\left(\frac{2m+1}{2m}\right)^2 \mu_k \;\le\; \widehat{\mu}_k \;\le\; \frac{\pi^2}{4}\left(\frac{2m+1}{2m}\right)^2 \mu_k,$$

showing that $\widehat{\mu}_k$ is a good approximation of μ_k, even for moderate values of m.

Recalling the definition of $\Psi \in \mathbb{R}^{m\times m}$ from Section 5.3.2, it can be shown that it takes the form $\Psi = I_m + \frac{1}{m}\mathbb{I}_s\mathbb{I}_s^T$, where $\mathbb{I}_s \in \mathbb{R}^m$ is the vector with entries $[\mathbb{I}_s]_j = (-1)^{j+1}$. Since $\lambda_{\max}(\Psi) = 2$, condition (D2) is clearly satisfied.

Now we consider the quantity $D_m(\Phi)$; by Lemma 30, it suffices to bound the operator norm of $K^{-1/2}(K^2 - \Theta)K^{-1/2}$. Some algebra shows that $K^2 - \Theta = m^{-4}(\frac{1}{2}hh^T + \frac{1}{6}m^2K)$, where $h = (1, 2, \ldots, m)$, so that

$$D_m(\Phi) = \|\!| K^{-1/2}(K^2 - \Theta)K^{-1/2} |\!\|_2 = \frac{1}{2m^4}\, h^T K^{-1} h + \frac{1}{6m^2} \;=\; \frac{1}{2m} + \frac{1}{6m^2} \;\leq\; \frac{1}{m}.$$

Finally, it was shown in Chapter 4 that $N_m(\Phi) \precsim \frac{\log m}{m^2}$.

Appendix 5.B Auxiliary lemmas

Here we collect the proofs of various auxiliary lemmas.

5.B.1 Proof of Lemma 29

The space $\mathrm{Ra}(\Phi^*)$ is finite-dimensional and hence closed, which guarantees validity of the well-known decomposition $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$. In particular, for any $f \in \mathcal{H}$, there is $a \in \mathbb{R}^m$ and $f^\perp \in \mathrm{Ker}(\Phi)$ such that $f = \Phi^* a + f^\perp$. Then, $\Phi f = Ka$, and

$$\|f\|_{\mathcal{H}}^2 \geq \|\Phi^* a\|_{\mathcal{H}}^2 = \langle \Phi^* a, \Phi^* a\rangle_{\mathcal{H}} = \langle a, \Phi\Phi^* a\rangle_{\mathbb{R}^m} = \langle Ka, Ka\rangle_K = \|\Phi f\|_K^2.$$

Equality holds iff $f^\perp = 0$ which gives the desired condition.

5.B.2 Proof of Lemma 30

By a well-known result, for a symmetric matrix, the numerical radius is equal to the operator norm. Thus, we have $\|\!| K - K^{-1/2}\Theta K^{-1/2} |\!\|_2 = \sup_{a\in\mathbb{R}^m\setminus\{0\}} \frac{|a^T\left(K - K^{-1/2}\Theta K^{-1/2}\right)a|}{\|a\|_2^2}$. Making the substitution $b = K^{-1/2}a$, or equivalently $a = K^{1/2}b$, we obtain

$$\|\!| K - K^{-1/2}\Theta K^{-1/2} |\!\|_2 = \sup_{b\in\mathbb{R}^m\setminus\{0\}} \frac{|b^T\left(K^2 - \Theta\right)b|}{b^T K b}$$

Now define the function $f = \Phi^* b \in \mathrm{Ra}(\Phi^*)$. With this definition, we have the following equivalences:

$$b^T K b = \|\Phi^* b\|_{\mathcal{H}}^2 \;=\; \|f\|_{\mathcal{H}}^2, \quad b^T K^2 b = \|\Phi f\|_2^2 \;=\; \|f\|_\Phi^2, \quad \text{and} \quad b^T\Theta b = \|\sum_{j=1}^m b_j\phi_j\|_{L^2}^2 \;=\; \|f\|_{L^2}^2,$$

from which the claim follows.

5.B.3 Proof of Lemma 31

The (truncated) QR decomposition [44] of Z^* has the form $Z^* = \widetilde{Z}^* R$, where $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$, and $R \in \mathbb{R}^{r\times r}$ is upper triangular with nonnegative diagonal entries. By construction, we have $\mathrm{Ra}(\widetilde{Z}^*) = \mathrm{Ra}(Z^*)$. Moreover, from the trace smoothness condition (5.18), we have

$$r\rho^2 \;\geq\; \sum_{j=1}^{r} \|z_j^*\|_K^2 \;=\; \mathrm{tr}\left((Z^*)^T K^{-1} Z^*\right) \;\geq\; \lambda_{\min}(R^T R)\,\mathrm{tr}\left((\widetilde{Z}^*)^T K^{-1}\widetilde{Z}^*\right) \tag{5.67}$$

where the final inequality follows from the bound (5.89) in Appendix 5.I. Recalling the definition (5.25), we have $C_m(f^*) = \|(Z^*)^T Z^* - I_r\|_\infty = \|R^T R - I_r\|_\infty$. Since $\lambda_j(R^T R) = \lambda_j(R^T R - I_r) + 1$, we have

$$\max_{j=1,\ldots,r} |\lambda_j(R^T R) - 1| \leq \|R^T R - I_r\|_2 \leq r\|R^T R - I_r\|_\infty = rC_m(f^*). \tag{5.68}$$

Since $rC_m(f^*) \leq \frac{1}{2}$, we conclude that $\lambda_{\min}(R^T R) \geq \frac{1}{2}$. Combined with our earlier bound (5.67), we conclude that $\widetilde{Z}^*$ indeed satisfies the trace-smoothness condition.

5.B.4 Proof of Lemma 33

We only need to consider the case $\nu = 1$; the general case follows by rescaling. Consider the following local one-sided version of $D_m(\Phi)$,

$$U_{\mathrm{loc}}(\epsilon;\Phi) := \sup_{\substack{f\in \mathrm{Ra}(\Phi^*),\\ \|f\|_{\mathcal{H}}\leq 1,\\ \|f\|_\Phi^2 \leq \epsilon^2}} \|f\|_{L^2}^2. \tag{5.69}$$

Using an argument similar to that of Lemma 30, (5.69) is equivalent to

$$U_{\mathrm{loc}}(\epsilon;\Phi) = \sup_{\substack{b^T b\leq 1,\\ b^T K b\leq \epsilon^2}} b^T K^{-1/2}\Theta K^{-1/2} b. \tag{5.70}$$

Using Lagrange duality, we have

$$\begin{aligned} U_{\mathrm{loc}}(\epsilon;\Phi) &\leq \inf_{t\geq 0}\; \left[\max\left(\lambda_{\max}(K^{-1/2}\Theta K^{-1/2} - tK), 0\right) + t\epsilon^2\right] \\ &\leq c_0\epsilon^2 \end{aligned} \tag{5.71}$$

since (D1) implies $\lambda_{\max}(K^{-1/2}\Theta K^{-1/2} - c_0 K) \leq 0$.

For $f \in \mathcal{H}$, let $f = g + f^\perp$ be its decomposition according to $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$. Then, $\|g\|_{\mathcal{H}}^2 + \|f^\perp\|_{\mathcal{H}}^2 = \|f\|_{\mathcal{H}}^2 \leq 1$ and $\|f\|_{L^2}^2 \leq 2\|g\|_{L^2}^2 + 2\|f^\perp\|_{L^2}^2$. Hence, we obtain

$$R_\Phi(\epsilon;1) \leq 2U_{\mathrm{loc}}(\varepsilon;\Phi) + 2N_m(\Phi). \tag{5.72}$$

Combining (5.71) and (5.72) proves the claim.

5.B.5 Proof of inequality (5.60)

By polarization identity and some algebra,

$$C_m(f^*) \le 2\rho^2 \sup_{\|f\|_{\mathcal{H}}\le 1,\ \|f\|_{L^2}=\frac{1}{\sqrt{2}\rho}} \left| \|f\|_\Phi^2 - \|f\|_{L^2}^2 \right|$$

Let $f = g + f^\perp$ be the decomposition according to $f \in \mathcal{H} = \mathrm{Ra}(\Phi^*) + \mathrm{Ker}(\Phi)$. Let $f \in \mathbb{B}_{\mathcal{H}}(1)$ and $\|f\|_{L^2} = \frac{1}{\sqrt{2}\rho}$. Then, as in Appendix 5.B.4, we have $g, f^\perp \in \mathbb{B}_{\mathcal{H}}(1)$. Hence,

$$\left| \|f\|_\Phi^2 - \|f\|_{L^2}^2 \right| \le \Big| \underbrace{\|g\|_\Phi^2 - \|g\|_{L^2}^2}_{\le D_m(\Phi)} \Big| + \Big| \underbrace{\|g + f^\perp\|_{L^2}^2}_{a^2} - \underbrace{\|g\|_{L^2}^2}_{b^2} \Big|$$

where we have define $a, b > 0$ as above for simplicity. Let $d := \|f^\perp\|_{L^2}$. By triangle inequality, $b \le a + d$ and $|a - b| \le d$. Then,

$$|a^2 - b^2| = |a - b|(a + b) \le d(2a + d) \le \big(\sqrt{\frac{2}{\rho}} + 1\big) N_m(\Phi),$$

since $a = \frac{1}{\sqrt{2}\rho}$ and $d \le N_m(\Phi) \le 1$, by assumption.

Appendix 5.C Proofs for Theorem 11

In this appendix, we collect the proofs of various auxiliary lemmas involved in the proof of Theorem 11.

5.C.1 Derivation of the bound (5.50)

From the CS-decomposition (5.43), we have $\widehat{Z}^T \widetilde{Z}^* = \widehat{C}$, and hence $\widehat{P}\widetilde{Z}^* = \widehat{Z}\widehat{C} - \widetilde{Z}^* = \widehat{E}\widehat{C} - \widetilde{Z}^*(I_r - \widehat{C})$. From the decomposition (5.42), we have

$$\begin{aligned}
\langle\!\langle \widehat{P},\ \Delta_1 \rangle\!\rangle &= \sigma \operatorname{tr}\left[\overline{W} S R^T (\widetilde{Z}^*)^T \widehat{P} + \widetilde{Z}^* R S \overline{W}^T \widehat{P}\right] \\
&= 2\sigma \operatorname{tr}\left[R S \overline{W}^T \widehat{P} \widetilde{Z}^*\right] \\
&= 2\sigma \Big\{ \operatorname{tr}\left[R S \overline{W}^T \widehat{E}\widehat{C}\right] - \operatorname{tr}\left[R S \overline{W}^T \widetilde{Z}^* (I_r - \widehat{C})\right] \Big\},
\end{aligned}$$

where we have used the standard facts $\mathrm{tr}(AB^T) = \mathrm{tr}(A^T B)$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. For the first term we have

$$\left| \operatorname{tr}\left[R S \overline{W}^T \widehat{E}\widehat{C}\right] \right| = \left| \sum_{j,k=1}^r R_{jk} s_k \langle \overline{w}_k, \widehat{e}_j \rangle\, \widehat{c}_j \right| \le \Big(\sum_{j,k=1}^r R_{j,k}^2 \Big)^{1/2} \Big(\sum_{j,k} s_k^2\, \widehat{c}_j^2\, (\langle \overline{w}_k, \widehat{e}_j \rangle)^2 \Big)^{1/2}$$

where we have used Cauchy-Schwarz. By (5.68), under the assumption $rC_m(f^*) \leq \frac{1}{2}$, we have $\mathrm{tr}(R^T R) \leq \frac{3}{2}r$. We also have $0 < s_k \leq s_1 = 1$ and $0 \leq \widehat{c}_j \leq 1$ for $j, k = 1, \ldots, r$. It follows that

$$\left| \mathrm{tr}\left[RS\overline{W}^T \widehat{E}\widehat{C}\right] \right| \leq \sqrt{\frac{3}{2}}\sqrt{r}\Big(\sum_{j,k=1}^{r} (\langle \overline{w}_k, \widehat{e}_j \rangle)^2 \Big)^{1/2} \leq \sqrt{\frac{3}{2}}\, r^{3/2} \max_{j,k} |\langle \overline{w}_k, \widehat{e}_j \rangle|.$$

For the second term, using a similar argument by applying Cauchy-Schwarz, we get

$$\begin{aligned}\left| \mathrm{tr}\left[RS\overline{W}^T \widetilde{Z}^*(I_r - \widehat{C})\right] \right| &\leq \sqrt{\frac{3}{2}}\sqrt{r}\Big(\sum_{j=1}^{r}(1-\widehat{c}_j)^2 \sum_{k=1}^{r}(\langle \overline{w}_k, \widetilde{z}_j^* \rangle)^2 \Big)^{1/2} \\ &\leq \sqrt{\frac{3}{2}}\, r\Big(\sum_{j}(1-\widehat{c}_j)^2 \Big)^{1/2} \max_{j,k} |\langle \overline{w}_k, \widetilde{z}_j^* \rangle| \leq \frac{\sqrt{3}}{2\sqrt{2}}\, r\, |||\widehat{E}|||_{HS}^2 \max_{j,k} |\langle \overline{w}_j, \widetilde{z}_k^* \rangle|.\end{aligned}$$

where the last inequality follows from the fact that $\big(\sum_j (1-\widehat{c}_j)^2\big)^{1/2} \leq \sum_j (1-\widehat{c}_j) = \frac{1}{2}|||\widehat{E}|||_{HS}^2$.

5.C.2 Proof of Lemma 36

We make use of an ellipsoid approximation (see [70]). To simplify notation, define $\widetilde{K} := (8r\rho^2)K$ and $\widetilde{\mu} := 8r\rho^2\,\widehat{\mu}$, so that we have $\mathrm{tr}(Z^T K^{-1} Z) \leq 2r\rho^2$ if and only if $\mathrm{tr}(Z^T \widetilde{K}^{-1} Z) \leq 1/4$. Since both $\widehat{Z}$ and $\widetilde{Z}^*$ satisfy this condition, it follows that $\|\overline{z}_j\|_{\widetilde{K}} \leq \frac{1}{2}$ and $\|\widetilde{z}_j^*\|_{\widetilde{K}} \leq \frac{1}{2}$ for $j = 1, \ldots, r$, where $\|a\|_{\widetilde{K}}^2 := a^T \widetilde{K}^{-1} a$. Thus, we are guaranteed that $\widehat{e}_j \in \mathcal{E}_{\widetilde{K}} := \big\{ v \in \mathbb{R}^m \mid \|v\|_{\widetilde{K}} \leq 1 \big\}$.

We first establish an upper bound on the quantity $\sup\big\{ \langle \overline{w}_k, v \rangle \;\mid\; v \in \mathcal{E}_{\widetilde{K}} \cap \mathbb{B}_2(t) \big\}$, where $\mathbb{B}_2(t) = \{v \in \mathbb{R}^m \mid \|v\|_2 \leq t\}$ is the Euclidean ball of radius t. Let $\widetilde{\mu}_1 \geq \cdots \geq \widetilde{\mu}_m$ be the eigenvalues of $\widetilde{K}$ in decreasing order and let $\widetilde{\mu} := (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_m)$. Since for $U \in V_m(\mathbb{R}^m)$, the random vectors $\overline{w}_k$ and $U\overline{w}_k$ have the same distribution, it is equivalent to bound the quantity $\sup\big\{ \langle \overline{w}_k, v \rangle \;\mid\; v \in \mathcal{E}_{\widetilde{\mu}} \cap \mathbb{B}_2(t) \big\}$. Now for $v \in \mathcal{E}_{\widetilde{\mu}} \cap \mathbb{B}_2(t)$, we have $\sum_{i=1}^m \widetilde{\mu}_i^{-1} v_i^2 \leq 1$ and $\sum_{i=1}^m t^{-2} v_i^2 \leq 1$ implying $\sum_{i=1}^m \max\{\widetilde{\mu}_i^{-1}, t^{-2}\} v_i^2 \leq 2$. Consequently, if we define the modified ellipse $\mathcal{E}_\gamma := \big\{ v \in \mathbb{R}^m \mid \sum_{i=1}^m \frac{v_i^2}{\gamma_i} \leq 1 \big\}$ where $\gamma_i := 2\min\{t^2, \widetilde{\mu}_i\}$, then we are guaranteed that $v \in \mathcal{E}_\gamma$, so that it suffices to upper bound $\sup_{v \in \mathcal{E}_\gamma} \langle \overline{w}_k, v \rangle$. For future reference, we note that

$$\|\gamma\|_1 = 16\,\mathcal{F}^2(t/\sqrt{8}), \quad \text{and} \quad \|\gamma\|_\infty \leq 2t^2 \tag{5.73}$$

where $\mathcal{F}$ was defined previously (5.24). Define the random variables $\overline{w}_k := \frac{1}{n}\sum_{i=1}^n w_i \beta_{ik}$ and $\overline{B}_{kk} := \frac{1}{n}\sum_{i=1}^n \beta_{ik}^2$. For each index k, Lemma 45 (see Appendix 5.F), combined with the relations (5.73), yields

$$\sigma \sup_{v \in \mathcal{E}_\gamma} |\langle \overline{w}_k, v \rangle| \leq \sigma \overline{B}_{kk}^{1/2} \Big\{ \sqrt{\frac{\|\gamma\|_1}{n}} + \delta \sqrt{\frac{\|\gamma\|_\infty}{n}} \Big\} \;\leq\; C_1\, \overline{B}_{kk}^{1/2} \frac{\sigma}{\sqrt{n}} \big\{ \mathcal{F}(t/\sqrt{8}) + \delta t \big\}$$

with probability at least $1-\exp(-\delta^2/2)$. Taking $\delta = A_r\sqrt{n}\,t/\sigma$, where $A_r := \kappa r^{-3/2}$ for some small enough constant $\kappa > 0$, we obtain

$$\sigma \sup_{v\in\mathcal{E}_\gamma} |\langle \overline{w}_k, v\rangle| \le C_1\, \overline{B}_{kk}^{1/2}\Big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(t/\sqrt{8}) + A_r t^2\Big\}$$

with probability at least $1-\exp(-A_r^2\, n\, t^2/2\sigma^2)$.

As was mentioned earlier, the same bound with the same probability holds for $\sup\big\{\sigma|\langle \overline{w}_k, v\rangle \;|\; v \in \mathcal{E}_{\widetilde{K}} \cap B_2(t)\big\}$. Since $\widehat{e}_j \in \mathcal{E}_{\widetilde{K}}$, $j = 1,\dots,r$ we can apply the technical Lemma 48 of Appendix 5.H with $\nu = (n,m)$, $\theta_\nu = A_r n/\sigma^2$ and $t_\nu = \epsilon_{m,n}$ to obtain

$$\sigma|\langle \overline{w}_k, \widehat{e}_j\rangle| \le C_1\, \overline{B}_{kk}^{1/2}\big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(2\|\widehat{e}_j\|_2/\sqrt{8}) + A_r(2\|\widehat{e}_j\|_2)^2 + \frac{\sigma}{\sqrt{n}}\mathcal{F}(2\epsilon_{m,n}/\sqrt{8}) + A_r(2\epsilon_{m,n})^2\big\},$$

for all $j \in \{1,\dots,r\}$, with probability at least $1-c_1\exp(-A_r^2\, n\,\epsilon_{m,n}^2/2\sigma^2)$. Note that $\|\widehat{e}_j\|_2 \le \|\!|\widehat{E}|\!\|_{HS}$, $j = 1,\dots,r$. Since the bound obtained above is nondecreasing in $\|\widehat{e}_j\|$, we can replace $\|\widehat{e}_j\|$ everywhere with $\|\!|\widehat{E}|\!\|_{HS}$. We also note that by χ^2_n concentration [55, 63], we have $\overline{B}_{kk} \le 3/2$ with probability at least $1-\exp(-n/64)$. Finally, by definition of $\epsilon_{m,n}$ and monotonicity of $\mathcal{F}$ we have $\frac{\sigma}{\sqrt{n}}\mathcal{F}(2\epsilon_{m,n}/\sqrt{8}) \le \frac{\sigma}{\sqrt{n}}\mathcal{F}(\epsilon_{m,n}) \le A_r\epsilon_{m,n}^2$. Putting together the pieces, we conclude that

$$\max_{j,k}\, \sigma|\langle \overline{w}_k, \widehat{e}_j\rangle| \le C_2\Big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(\|\!|\widehat{E}|\!\|_{HS}) + A_r\|\!|\widehat{E}|\!\|_{HS}^2 + A_r\,\epsilon_{m,n}^2\Big\},$$

with probability at least $1 - c_1 r\exp(-A_r^2 n\,\epsilon_{m,n}^2/2\sigma^2) - r\exp(-n/64)$, where we have used union bound to obtain a uniform result over k.

5.C.3 Proof of Lemma 37

We control terms of the form $\langle \overline{w}_k, \widetilde{z}_j^*\rangle$ using Lemma 45 in Appendix 5.F, this time with w_i replaced with $\langle w_i, \widetilde{z}_j^*\rangle$ and $\gamma = 1$ (i.e., we are looking at sums of products of univariate Gaussians). Thus, for any fixed j and k, we have $\sigma|\langle \overline{w}_k, \widetilde{z}_j^*\rangle| \le \sigma\,\overline{B}_{kk}^{1/2}\big\{\frac{1}{\sqrt{n}} + \delta\frac{1}{\sqrt{n}}\big\}$, with probability at least $1-\exp(-\delta^2/2)$. Taking $\delta = \kappa r^{-1}\sqrt{n}/\sigma$, then the event $\max_k \overline{B}_{kk} \le 3/2$, which we have already accounted for, we have by union bound

$$\max_{j,k}\, \sigma r|\langle \overline{w}_k, \widetilde{z}_j^*\rangle| \le \sqrt{\frac{3}{2}}\Big\{\frac{\sigma}{\sqrt{n}}r + \kappa\Big\} \le \sqrt{6}\kappa$$

with probability at least $1 - r^2\exp(-\kappa^2 r^{-2} n/2\sigma^2)$. The second inequality follows by our assumption $r \le \kappa\sqrt{n}/\sigma$.

5.C.4 Proof of Lemma 38

For each $j \in \{1, \dots, r\}$, we define the vector $\zeta^j := W\widetilde{z}_j^* \in \mathbb{R}^n$ so that $\zeta^j = (\zeta_i^j)$ where $\zeta_i^j = w_i^T \widetilde{z}_j^*$. We can use the same ellipsoid approximation as Appendix 5.C.2— that is, we first look at $\sup\{T_1(v; \widetilde{z}_j^*) \mid v \in \mathcal{E}_{\widetilde{K}} \cap \mathbb{B}_2(t)\}$ and then argue that it is enough to bound $\sup_{v \in \mathcal{E}_\gamma} T_1(v; \widetilde{z}_j^*) = \sup_{v \in \mathcal{E}_\gamma} \langle v, \frac{1}{n}\sum_{i=1}^n \zeta_i^j w_i - \widetilde{z}_j^* \rangle$, due to the invariance of the underlying distribution under orthogonal transformations of v. Now applying Lemma 46 from Appendix 5.F yields

$$\sigma^2 \sup_{v \in \mathcal{E}_\gamma} T_1(v; \widetilde{z}_j^*) \;\le\; \Big(\frac{\|\zeta^j\|_2}{\sqrt{n}} + 1\Big)\Big\{\frac{\sigma^2}{\sqrt{n}}\sqrt{\|\gamma\|_1} + \delta\sigma^2\sqrt{\|\gamma\|_\infty}\Big\} \tag{5.74}$$

with probability at least $1 - 2\exp(-n\frac{\delta \wedge \delta^2}{16})$. Recalling that by assumption $\sigma \le \sigma_0$, let $\widetilde{A}_r = \kappa r^{-1}$. For $t \le \sigma\sigma_0/\widetilde{A}_r$, take $\delta = \widetilde{A}_r t/(\sigma\sigma_0) \le 1$. Then, using (5.73), the left-hand side of (5.74) is bounded above by

$$C_1\Big(\frac{\|\zeta^j\|_2}{\sqrt{n}} + 1\Big)\Big\{\sigma_0\frac{\sigma}{\sqrt{n}}\mathcal{F}(t/\sqrt{8}) + \widetilde{A}_r t^2\Big\} \tag{5.75}$$

with probability at least $1 - 2\exp\big(-\widetilde{A}_r^2 n\, t^2/(16\,\sigma^2\sigma_0^2)\big)$. For $t > \sigma\sigma_0/\widetilde{A}_r$, take $\delta = \widetilde{A}_r t/\sigma^2$. In this case, $\widetilde{A}_r t > \sigma\sigma_0 \ge \sigma^2$ implying $\delta > 1$. Then, the left-hand side of (5.74) is again bounded above by (5.75), this time with probability at least $1 - 2\exp(-\widetilde{A}_r n\, t/(16\,\sigma^2))$. Assuming $\kappa \le 1$, which is going to be the case, we have $\widetilde{A}_r^2 \le \widetilde{A}_r$. Combining the two cases, we have the upper bound (5.75) with probability at least

$$1 - 2\underbrace{\exp\big\{-\widetilde{A}_r^2\, n\,(\sigma_0^{-2} \wedge 1)\,(t \wedge t^2)\,/\,(16\sigma^2)\big\}}_{p_1(t)} \tag{5.76}$$

for all $t > 0$. (Note the break-up into two cases was to obtain a dependence of σ^{-2} in the probability exponent for all $t > 0$.)

By an argument similar to Appendix 5.C.2—that is, using technical Lemma 48—we have $\|\widehat{e}_j\|_2 \le |\!|\!|\widehat{E}|\!|\!|_{HS}$ and $\frac{\sigma}{\sqrt{n}}\mathcal{F}(2\epsilon_{m,n}/\sqrt{8}) \le \widetilde{A}_r\epsilon_{m,n}^2$ from the definition; we obtain

$$\sigma^2 T_1(\widehat{e}_j; \widetilde{z}_j^*) \le C_2\Big(\frac{\|\zeta^j\|_2}{\sqrt{n}} + 1\Big)\big\{\sigma_0\frac{\sigma}{\sqrt{n}}\mathcal{F}(|\!|\!|\widehat{E}|\!|\!|_{HS}) + \widetilde{A}_r|\!|\!|\widehat{E}|\!|\!|_{HS}^2 + \widetilde{A}_r\epsilon_{m,n}^2\big\}$$

for all $j \in \{1, \dots, r\}$, with probability at least that of (5.76) with $t = \epsilon_{m,n}$ and 2 replaced with some constant $c_2 > 2$, i.e. $1 - c_2\, p_1(\epsilon_{m,n})$. By concentration of χ_n^2 variables and union bound, we have $\max_j n^{-1}\|\zeta^j\|_2^2 \le 3/2$ with probability at least $1 - r\exp(-n/64)$. Putting together the pieces, we conclude that

$$\sigma^2 \sum_{j=1}^r T_1(\widehat{e}_j; \widetilde{z}_j^*) \le C_3\Big\{\sigma_0\frac{\sigma}{\sqrt{n}} r\mathcal{F}(|\!|\!|\widehat{E}|\!|\!|_{HS}) + \kappa|\!|\!|\widehat{E}|\!|\!|_{HS}^2 + \kappa\epsilon_{m,n}^2\Big\}$$

with probability at least $1 - c_2\, p_1(\epsilon_{m,n}) - r\exp(-n/64)$, as claimed.

5.C.5 Proof of Lemma 39

As before, the problem of bounding $T_2(\widehat{e}_j)$ can be reduced to controlling $\sup_{v\in\mathcal{E}_\gamma} T_2(v)$, by invariance under orthogonal transformation. Applying Lemma 47 of Appendix 5.G with with $\delta = \kappa\sqrt{n}\,t/\sigma$ yields

$$\begin{aligned}\sigma \sup_{v\in\mathcal{E}_\gamma} \sqrt{T_2(v)} = \sup_{v\,\in\,\mathcal{E}_\gamma} \frac{\sigma}{\sqrt{n}}\|Wv\|_2 &\leq \sigma\Big\{\sqrt{\frac{\|\gamma\|_1}{n}} + \big(1+\kappa\frac{t}{\sigma}\big)\sqrt{\|\gamma\|_\infty}\Big\} \\ &\leq C_1\big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(t/\sqrt{8}) + \sigma t + \kappa t^2\big\} \\ &\leq C_1\big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(t) + \sigma t + \sqrt{2}\kappa t\big\}\end{aligned}$$

with probability at least $1-\exp(-\kappa^2 n\,t^2/2\sigma^2)$, valid for all $t \leq \sqrt{2}$, Note that since $\|\widehat{e}_j\|_2 \leq \sqrt{2}$ (by proper alignment), it is enough to only have a bound for $t \leq \sqrt{2}$. Recall the assumption (A1), $\sigma \leq \sqrt{\kappa}$ and by (A3), $\frac{\sigma}{\sqrt{n}}\,\mathcal{F}(t) \leq \sqrt{\kappa}t$. Assuming $\kappa < 1$, we obtain

$$\sigma^2 \sup_{v\,\in\,\mathcal{E}_\gamma} T_2(v) \leq C_1^2\big(2\sqrt{\kappa}\,t + \sqrt{2}\kappa t\big)^2 \leq C_2\kappa t^2$$

with the same probability. As before, applying technical Lemma 48, this time with $t_\nu = r^{-1/2}\epsilon_{m,n}$, we obtain

$$\sigma^2 T_2(\widehat{e}_j) \;\leq\; C_2\kappa\Big\{(2\|\widehat{e}_j\|_2)^2 + \Big(2\frac{\epsilon_{m,n}}{\sqrt{r}}\Big)^2\Big\}$$

for all $j \in \{1,\ldots,m\}$ with probability at least $1 - c_3\exp(-\kappa^2 r^{-2} n\epsilon_{m,n}^2/2\sigma^2)$. Thus, we have

$$\sigma^2\sum_{j=1}^{r} T_2(\widehat{e}_j) \leq C_3\kappa\Big\{\|\!|\widehat{E}|\!\|_{HS}^2 + \epsilon_{m,n}^2\Big\}$$

with probability the same probability. Note that we have used $\|\!|\widehat{E}|\!\|_{HS}^2 = \sum_j \|\widehat{e}_j\|_2^2$.

Appendix 5.D Proofs for Theorem 12

In this appendix, we collect the proofs of various auxiliary lemmas involved in the proof of Theorem 12.

5.D.1 Proof of Lemma 40

By definition, each $\widehat{h}_j$ lies in $\widehat{\mathfrak{F}}$, so that we have $\widehat{h}_j = \Phi^*\big(\sum_i B_{ij}K^{-1}\widehat{z}_i\big)$ for some $B \in \mathbb{R}^{r\times r}$. Recalling that $[K^{-1}\widehat{Z}B]_j$ denotes the j-th column of $K^{-1}\widehat{Z}B$, we can write $\widehat{h}_j =$

$\Phi^*[K^{-1}\widehat{Z}B]_j$. Recalling the formula (5.5) for the adjoint, observe that for any $a, b \in \mathbb{R}^m$, we have

$$\langle \Phi^* a, \Phi^* b \rangle_{L^2} = \langle \sum_i a_i \varphi_i, \sum_j b_j \varphi_j \rangle_{L^2} = \sum_{i,j} a_i b_j \langle \varphi_i, \varphi_j \rangle_{L^2} = a^T \Theta\, b \tag{5.77}$$

where $\Theta = (\langle \varphi_i, \varphi_j \rangle_{L^2}) \in \mathbb{S}_+^m$, as previously defined in Lemma 30. Since the functions $\{\widehat{h}_j\}_{j=1}^r$ are orthonormal in L^2, we must have $\langle \widehat{h}_j, \widehat{h}_k \rangle_{L^2} = [K^{-1}\widehat{Z}B]_j^T \Theta [K^{-1}\widehat{Z}B]_k = \delta_{jk}$, or in matrix form $(K^{-1}\widehat{Z}B)^T \Theta (K^{-1}\widehat{Z}B) = I_{r\times r}$. This condition can be re-written as

$$B^T \widehat{Q} B = I_{r\times r} = I, \quad \text{where} \quad \widehat{Q} := \widehat{Z}^T K^{-1} \Theta K^{-1} \widehat{Z}.$$

Since $\widehat{h}_j - \widehat{g}_j = \Phi^*[K^{-1}\widehat{Z}(B - I_r)]_j$, we have $\|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = [K^{-1}\widehat{Z}(B-I_r)]_j^T \Theta [K^{-1}\widehat{Z}(B-I_r)]_j$, using the definition of Θ. Consequently, we obtain

$$\sum_{j=1}^r \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = \operatorname{tr}\left\{ \left(K^{-1}\widehat{Z}(B-I)\right)^T \Theta \left(K^{-1}\widehat{Z}(B-I)\right) \right\} = \operatorname{tr}\left\{ I + \widehat{Q} - 2\widehat{Q}B \right\},$$

using the symmetry of $\widehat{Q}$ and the constraint $B^T\widehat{Q}B = I$. Subject to this constraint, we are free to choose B as we please; setting $B = \widehat{Q}^{-1/2}$ yields

$$\sum_{j=1}^r \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = \operatorname{tr}\left\{ \left(I - \widehat{Q}^{1/2}\right)^2 \right\} = |\!|\!| I - \widehat{Q}^{1/2} |\!|\!|_{HS}^2.$$

In order to upper bound $|\!|\!| I - \widehat{Q}^{1/2} |\!|\!|_{HS}$, we first control the closely related quantity $|\!|\!| I - \widehat{Q} |\!|\!|_{HS}$. We have

$$\begin{aligned} |\!|\!| I_r - \widehat{Q} |\!|\!|_{HS} &= |\!|\!| \widehat{Z}^T K^{-1/2} (K - K^{-1/2}\Theta K^{-1/2}) K^{-1/2} \widehat{Z} |\!|\!|_{HS} \\ &\leq |\!|\!| K - K^{-1/2}\Theta K^{-1/2} |\!|\!|_2 \, |\!|\!| \widehat{Z}^T K^{-1} \widehat{Z} |\!|\!|_{HS} \\ &\leq 2r\rho^2 \, D_m(\Phi), \end{aligned} \tag{5.78}$$

where we have used inequality (5.92), Lemma 30, the trace-smoothness condition $\operatorname{tr}(\widehat{Z}^T K^{-1} \widehat{Z}) \leq 2r\rho^2$, and the inequality $|\!|\!| M |\!|\!|_{HS} \leq \operatorname{tr}(M)$, valid for any $M \succeq 0$.

In order to bound $|\!|\!| I - \widehat{Q}^{1/2} |\!|\!|_{HS}$, we apply the inequality

$$|\!|\!| A^q - D^q |\!|\!| \leq q a^{q-1} |\!|\!| A - D |\!|\!|, \quad 0 < q < 1, \tag{5.79}$$

valid for any operators A, D such that $A \succeq aI$ and $D \succeq aI$ for some positive number a, where $|\!|\!| \cdot |\!|\!|$ is any unitarily invariant norm. (See Bhatia [15], equation (X.46) on p. 305). As long as $2r\rho^2 D_m(\Phi) \leq 1/2$ so that the bound (5.78) implies that $\widehat{Q} \succeq 1/2 I$, we may apply the inequality (5.79) with $A = I_r$, $D = \widehat{Q}$, $a = q = 1/2$ and $|\!|\!| \cdot |\!|\!| = |\!|\!| \cdot |\!|\!|_{HS}$ so as to obtain the inequality $|\!|\!| I_r - \widehat{Q}^{1/2} |\!|\!|_{HS} \leq \frac{1}{\sqrt{2}} |\!|\!| I_r - \widehat{Q} |\!|\!|_{HS}$, which completes the proof.

5.D.2 Proof of Lemma 41

By definition, we have $h_j^* = \sum_{i=1}^r E_{ij} f_i^*$ for some $E \in \mathbb{R}^{r\times r}$. Since both $\{h_j^*\}$ and $\{f_j^*\}$ are assumed orthonormal in L^2, the matrix E must be orthonormal. In addition, we have $\sum_{j=1}^r \|h_j^*\|_{\mathcal{H}}^2 = \sum_{j=1}^r \|f_j^*\|_{\mathcal{H}}^2 \le r\rho^2$, implying that

$$\sum_{j=1}^r \|h_j^* - g_j^*\|_{\mathcal{H}}^2 \le 2\sum_{j=1}^r \left(\|h_j^*\|_{\mathcal{H}}^2 + \|g_j^*\|_{\mathcal{H}}^2\right) \le 2(r\rho^2 + 2r\rho^2) \le 6r\rho^2$$

where we have used the fact that $\sum_{j=1}^r \|g_j^*\|_{\mathcal{H}}^2 = \sum_{j=1}^r \|\widetilde{z}_j^*\|_K^2 \le 2r\rho^2$.

Recall the argument leading to the bound (5.55); applying this same reasoning to the pair (h_j^*, g_j^*) with the choices $A^2 = \sum_{j=1}^r \|h_j^* - g_j^*\|_\Phi^2$ and $B^2 = 6r\rho^2$ leads to

$$\sum_{j=1}^r \|h_j^* - g_j^*\|_{L^2}^2 \le c_1 \sum_{j=1}^r \|h_j^* - g_j^*\|_\Phi^2 + 6r\rho^2 S_m(\Phi).$$

It remains to bound the term $\sum_{j=1}^r \|h_j^* - g_j^*\|_\Phi^2$. Recalling that $\Phi f_i^* = z_i^*$, we note that $\Phi h_j^* = \sum_i E_{ij} z_i^* = [Z^* E]_j$. It follows that $\sum_{j=1}^r \|h_j^* - g_j^*\|_\Phi^2 = \|Z^*E - \widetilde{Z}^*\|_{HS}^2$. Since $\mathrm{Ra}(Z^*) = \mathrm{Ra}(\widetilde{Z}^*)$, there exists a matrix $R \in \mathbb{R}^{r\times r}$ such that $Z^* = \widetilde{Z}^* R$. Letting $V_1 \Upsilon V_2^T$ denote the SVD of R, we have

$$\|Z^*E - \widetilde{Z}^*\|_{HS} = \|RE - I_r\|_{HS} = \|R - E^T\|_{HS} = \|V_1 \Upsilon V_2^T - E^T\|_{HS},$$

where we have used the unitary invariance of the Hilbert-Schmidt norm. Take $E^T = V_1 V_2^T$ which is orthogonal, hence a valid choice. By unitary invariance, we have $\|Z^*E - \widetilde{Z}^*\|_{HS} = \|\Upsilon - I_r\|_{HS}$. We now apply inequality (5.79) with $a = q = 1/2$, $A = \Upsilon^2$, $D = I_r$, $|||\cdot||| = |||\cdot|||_{HS}$. The condition $rC_m(f^*) \le \frac{1}{2}$ implies $\Upsilon^2 \succeq \frac{1}{2} I_r$. (See Appendix 5.B.3, in particular the argument following (5.68).) Consequently, we have $\|\Upsilon - I_r\|_{HS} \le \frac{1}{\sqrt{2}} \|\Upsilon^2 - I_r\|_{HS} \le \frac{1}{\sqrt{2}} |||Z^{*T} Z^* - I_r|||_{HS}$, where we have used $V_2 \Upsilon^2 V_2^T = R^T R = Z^{*T} Z^*$. Recalling that $|||Z^{*T} Z^* - I_r|||_{HS} \le rC_m(f^*)$ and putting together the pieces, we obtain the stated inequality (5.58).

Appendix 5.E Proofs for Theorems 13 and 14

In this appendix, we prove various lemmas that are involved in the proofs of the lower bounds given in Theorems 13 and 14.

5.E.1 Proof of Lemma 42

Let us introduce the shorthand notation $u = \Phi(f)$ and $v = \Phi(g)$. Under the model $\mathbb{P}_f$, for each $i = 1, 2, \ldots, n$, the vector $y_i \in \mathbb{R}^m$ has a zero-mean Gaussian distribution with

covariance matrix $\Sigma_f := uu^T + \sigma_m^2 I$. Similarly, under the model $\mathbb{P}_g$, it is zero-mean Gaussian with covariance $\Sigma_g := vv^T + \sigma_m^2 I$. Since the data is i.i.d. and using standard formula for the Kullback-Leibler divergence between multivariate Gaussian distributions, we have $\frac{2}{n} D(\mathbb{P}_f \,\|\, \mathbb{P}_g) = \log \frac{\det \Sigma_g}{\det \Sigma_f} + \operatorname{tr}(\Sigma_g^{-1}\Sigma_f) - m$. Since $\|u\|_2 = \|v\|_2$ by construction, the matrices Σ_f and Σ_g have the same eigenvalues, and so the first term vanishes. Using the matrix inversion formula, we have

$$\frac{2}{n} D(\mathbb{P}_f \,\|\, \mathbb{P}_g) + m \;=\; \langle\!\langle (\sigma_m^2 I + vv^T)^{-1},\, \sigma_m^2 I + uu^T \rangle\!\rangle = \langle\!\langle \sigma_m^{-2} I - \sigma_m^{-4} \frac{vv^T}{1 + \|v\|_2^2 \sigma_m^{-2}},\, \sigma_m^2 I + uu^T \rangle\!\rangle.$$

Using the fact that $\|u\|_2 = \|v\|_2 =: a$ implies

$$\frac{2}{n} D(\mathbb{P}_f \,\|\, \mathbb{P}_g) = \sigma_m^{-2} a^2 - \frac{\sigma_m^{-2} a^2}{1 + a^2 \sigma_m^{-2}} - \frac{\sigma_m^{-4}}{1 + a^2 \sigma_m^{-2}} \langle u, v \rangle^2 \;=\; \frac{\sigma_m^{-4}}{1 + a^2 \sigma_m^{-2}} \big(a^2 + \langle u, v \rangle\big)\, \frac{\|u - v\|_2^2}{2}.$$

Since $|\langle u, v \rangle| \leq a^2$ by Cauchy-Schwarz, we have $a^2 + \langle u, v \rangle \leq 2a^2$, and hence

$$\frac{2}{n} D(\mathbb{P}_f \,\|\, \mathbb{P}_g) \leq \frac{a^2 \sigma_m^{-2}}{1 + a^2 \sigma_m^{-2}} \frac{\|u - v\|_2^2}{\sigma_m^2} \;\leq\; \frac{\|u - v\|_2^2}{\sigma_m^2},$$

as claimed.

5.E.2 Proof of Lemma 43

As previously observed, any function $f \in \operatorname{Ra}(\Phi^*) \cap \mathbb{B}_{\mathcal{H}}(1)$ can be represented by a vector in the ellipse $\mathcal{E} := \big\{\theta \in \mathbb{R}^m \;\big|\; \sum_{j=1}^m \theta_j^2 / \widehat{\mu}_j \leq 1\big\}$ such that $\|f\|_\Phi = \|\theta\|_2$. The proofs of both parts (a) and (b) exploit this representation.

(a) Note that the ellipse $\mathcal{E}$ contains the ℓ_2^m-ball of radius $\sqrt{\widehat{\mu}_m}$. It is known [68] that there exists a $1/2$ packing of the ℓ_2^m-ball which has at least $M = 4^m$ elements, all of which have unit norm. By rescaling this packing by $\frac{\sigma_0}{\sqrt{n}}$, we obtain a collection of M vectors $\{\theta^1, \ldots, \theta^M\}$ such that

$$\|\theta^i\|_2^2 = \frac{\sigma_0^2}{n} \quad \text{and} \quad \|\theta^i - \theta^j\|_2^2 \geq \frac{\sigma_0^2}{4n}, \quad \text{for all } i \neq j \in [M].$$

The condition $m \leq (c_0 n)^{\frac{1}{2\alpha}}$ implies that $\|\theta^i\|_2^2 \leq (c_0 \sigma_0^2) m^{-2\alpha} \leq \widehat{\mu}_m$, where the second inequality follows since by assumption (A1) we can take σ_0^2 sufficiently small. Thus, these vectors are also contained within the ellipse $\mathcal{E}$, even after we rescale them further by $1/4$, which establishes the claim.

(b) This part makes use of the elementary inequality

$$k \log k - k \overset{(\ell)}{\leq} \sum_{j=1}^k \log j \overset{(u)}{\leq} (k+1)\log(k+1) - (k+1). \tag{5.80}$$

We use known results on the entropy numbers of diagonal operators, in particular for the operator mapping the ℓ_2-ball to the ellipse $\mathcal{E}$. By assumption, we have $\widehat{\mu}_j j^{2\alpha} \in [c_\ell, c_u]$ for all $j = 1, 2, \ldots, m$. By Proposition 1.3.2 of [28] with $p = 2$, we have

$$\begin{aligned}
\log N_\Phi(\epsilon; \mathcal{E}) &\leq \max_{k=1,2,\ldots,m} \Big\{ \frac{1}{2} \sum_{j=1}^{k} \log \widehat{\mu}_j + k \log(1/\epsilon) \Big\} + \log 6 \\
&\leq \max_{k=1,2,\ldots,m} \Big\{ -\alpha \sum_{j=1}^{k} \log j + k \log(1/\epsilon) \Big\} + \log(6 c_u) \\
&\leq \max_{1 \leq k \leq m} f(k) + \log(6 c_u),
\end{aligned}$$

where $f(k) = \alpha(k - k \log k) + k \log(1/\epsilon)$. Since $f'(k) = -\alpha \log k + \log(1/\epsilon)$, the optimum is achieved for $k^* = (1/\epsilon)^{1/\alpha}$, and has value $f(k^*) = \alpha(1/\epsilon)^{1/\alpha}$, which establishes the claim.

In the other direction, for all $k \in \{1, 2, \ldots, m\}$, we have

$$\log M_\Phi(\epsilon; \mathcal{E}) \geq \frac{1}{2} \sum_{j=1}^{k} \log \widehat{\mu}_j + k \log(1/\epsilon) \geq -\alpha \sum_{j=1}^{k} \log j + k \log(1/\epsilon) + \log c_\ell.$$

Using the lower bound (5.80)(u), we obtain

$$\log M_\Phi(\epsilon; \mathcal{E}) \geq \alpha((k+1) - (k+1)\log(k+1)) + k \log(1/\epsilon) + \log c_\ell.$$

The choice $k+1 = (1/\epsilon)^{1/\alpha}$, which is valid under the given condition $(1/\epsilon)^{1/\alpha} \leq m-1$, yields the claim.

5.E.3 Proof of Lemma 44

Any function f in the set Ψ_1^m has the form $f = \sum_{j=1}^m a_j \psi_j$ for a vector of coefficients $a \in \mathbb{R}^m$ such that $\sum_{j=1}^m a_j^2/\mu_j \leq 1$. If $g = \sum_{j=1}^m b_j \psi_j$ is a second function, then we have $\|f - g\|_{L^2} = \|a - b\|_2$ by construction. Thus, the problem is equivalent to bounding the covering/packing numbers of the m-dimensional ellipse specified by the eigenvalues $\{\mu_1, \ldots, \mu_m\}$. The claim thus follows from the proof of Lemma 43(b).

Appendix 5.F Suprema involving Gaussian products

Given a diagonal matrix $Q := \operatorname{diag}(\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^{m \times m}$, this appendix provides bounds on $\|Q^{1/2}\xi\|_2$ where $\xi \in \mathbb{R}^m$ is some random vector (product of Gaussians in particular). The following bound, which follows from Jensen's inequality, is useful:

$$\mathbb{E}\, \|Q^{1/2}\xi\|_2 \leq \sqrt{\mathbb{E}\, \|Q^{1/2}\xi\|_2^2} = \sqrt{\operatorname{tr}(Q\Sigma_\xi)}, \quad \text{where } \Sigma_\xi := \mathbb{E}\, \xi\xi^T. \tag{5.81}$$

We prove a bound for the random vector $\xi := n^{-1}\sum_{i=1}^n \beta_i w_i \in \mathbb{R}^m$, where $\beta_i \sim N(0,1)$, independent of $w_i \sim N(0, I_m)$, and the pairs (β_i, w_i) i.i.d. for $i = 1, \dots, n$.

Lemma 45. *For all $t \geq 0$, we have*

$$\mathbb{P}\left[\frac{\|Q^{1/2}\sum_{i=1}^n \beta_i w_i\|_2}{\|\beta\|_2} > \sqrt{\operatorname{tr}(Q)} + t\sqrt{|\!|\!|Q|\!|\!|}\right] \leq \exp(-t^2/2), \tag{5.82}$$

where $\beta = (\beta_1, \dots, \beta_n)$.

Proof. Define $\theta := \beta/\|\beta\|_2$, and observe that θ is uniformly distributed on the sphere S^{n-1}, independent of (w_i); we use σ^{n-1} to denote this uniform distribution. The claim is a deviation bound for $\|Q^{1/2}\sum_{i=1}^n \theta_i w_i\|_2$. With θ held fixed, we have $\widetilde{w} := \sum_{i=1}^n \theta_i w_i \sim N(0, I_m)$. The map $\widetilde{w} \mapsto \|Q^{1/2}\widetilde{w}\|_2$ is Lipschitz, from ℓ_2^m to $\mathbb{R}$, with Lipschitz constant bounded by $|\!|\!|Q^{1/2}|\!|\!| = \sqrt{|\!|\!|Q|\!|\!|}$. Hence, by concentration of the canonical Gaussian measure in $\mathbb{R}^m$, with θ held fixed, we have

$$\mathbb{P}\big[\|Q^{1/2}\widetilde{w}\|_2 - \mathbb{E}\,\|Q^{1/2}\widetilde{w}\|_2 \geq t\sqrt{|\!|\!|Q|\!|\!|}\big] \leq \exp(-t^2/2).$$

Since this bound holds for all realizations of θ, the tower property implies that the same bound holds unconditionally. Finally, from the bound (5.81), we have $\mathbb{E}\,\|Q^{1/2}\widetilde{w}\|_2 \leq \sqrt{\operatorname{tr}(Q)}$, from which the claim follows. □

We now turn to bounding $\|Q^{1/2}(n^{-1}\sum_i \eta_i w_i - u)\|_2$, where $u \in \mathbb{R}^m$ is some fixed vector. Let us patch u with $u_2, \dots, u_m$ so that $\{u, u_2, \dots, u_m\}$ is an orthonormal basis for ℓ_2^m. Let us define the function $\zeta : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ as $\zeta(x) := \frac{n^{-1}\|x\|_2^2 - 1}{n^{-1}\|x\|_2}$. With this notation, we have the following:

Lemma 46. *Let $u \in S^{m-1}$ and assume that $U := (u\ U_2) = (u\ u_2\ \cdots\ u_m) \in \mathbb{R}^{m\times m}$ is orthogonal. Let $(w_i, \eta_i) \in \mathbb{R}^{m+1}$ be i.i.d. Gaussian random vectors for $i = 1, \dots, n$ with distribution*

$$\begin{bmatrix} w_i \\ \eta_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_m & u \\ u^T & 1 \end{bmatrix}\right).$$

Then for all $t \geq 0$,

$$\mathbb{P}\left[\left\|Q^{1/2}\,(n^{-1}\sum_{i=1}^n \eta_i w_i - u)\right\|_2 \geq \big(1 + \frac{\|\eta\|_2}{\sqrt{n}}\big)\,\Big(\sqrt{\frac{\operatorname{tr}(Q)}{n}} + t\sqrt{|\!|\!|Q|\!|\!|}\Big)\right] \leq 2\exp(-n\frac{t \wedge t^2}{16}),$$

where $\eta = (\eta_1, \dots, \eta_n)$.

Proof. Since the pair (w_i, η) is jointly Gaussian, vectors $\{w_i\}$ conditioned on $\eta = (\eta_i)$ are i.i.d. Gaussian with $\mathbb{E}[w_i \,|\, \eta_i] = \eta_i u$ and $\operatorname{cov}(w_i \,|\, \eta_i) = I_m - uu^T$. Consequently, conditioned on η, the variable $\widehat{w}_\eta := n^{-1}\sum_i \eta_i w_i - u$ is Gaussian with mean $u(n^{-1}\|\eta\|_2^2 - 1)$ and covariance $n^{-2}\|\eta\|_2^2(I_m - uu^T)$. Consequently, for $\widetilde{w}_\eta := \widehat{w}_\eta/(n^{-1}\|\eta\|_2)$, we have

$$U^T\widetilde{w}_\eta \sim N\Big(\begin{bmatrix}1\\0\end{bmatrix}\zeta(\eta), \begin{bmatrix}0 & 0\\ 0 & I_{m-1}\end{bmatrix}\Big),$$

where we have used $U^T u = \binom{1}{0}$. Note that $U^T\widetilde{w}_\eta$ is actually a degenerate Gaussian vector, so that we can write $U^T\widetilde{w}_\eta = \big(\zeta(\eta),\ w'\big)$, for some $w' \sim N(0, I_{m-1})$.

Defining $\widetilde{Q} := U^T Q U$, we have

$$\|Q^{1/2}\widetilde{w}_\eta\|_2 = \|U^T Q^{1/2} U U^T \widetilde{w}_\eta\|_2 = \|\widetilde{Q}^{1/2}\, U^T\widetilde{w}_\eta\|_2 = \Big\|\widetilde{Q}^{1/2}\begin{bmatrix}\zeta(\eta)\\ w'\end{bmatrix}\Big\|_2.$$

The map $w' \mapsto \big\|\widetilde{Q}^{1/2}\big[\begin{smallmatrix}\zeta(\eta)\\ w'\end{smallmatrix}\big]\big\|_2$ is Lipschitz, from ℓ_2^{m-1} to $\mathbb{R}$, with Lipschitz constant bounded by $|||\widetilde{Q}^{1/2}||| = |||Q^{1/2}||| = \sqrt{|||Q|||}$. By concentration of canonical Gaussian measure in $\mathbb{R}^{m-1}$, we have

$$\mathbb{P}\big[\|Q^{1/2}\widetilde{w}_\eta\|_2 - \mathbb{E}\,\|Q^{1/2}\widetilde{w}_\eta\|_2 > t\sqrt{|||Q|||} \;\big|\; \eta\big] \le \exp(-t^2/2).$$

Define the function $\kappa(\eta) := \langle\!\langle Q,\ I_m + (\zeta^2(\eta) - 1)uu^T\rangle\!\rangle$. Applying the inequality (5.81) with $\xi = \big[\begin{smallmatrix}\zeta(\eta)\\ w'\end{smallmatrix}\big]$ and $\widetilde{Q}$ instead of Q, we obtain

$$\mathbb{E}\,\|Q^{1/2}\widetilde{w}_\eta\|_2 = \mathbb{E}\,\Big\|\widetilde{Q}^{1/2}\begin{bmatrix}\zeta(\eta)\\ w'\end{bmatrix}\Big\|_2 \le \Big\{\operatorname{tr}\Big(\widetilde{Q}\begin{bmatrix}\zeta^2(\eta) & 0\\ 0 & I_{m-1}\end{bmatrix}\Big)\Big\}^{1/2} = \sqrt{\kappa(\eta)}.$$

Since $Q \succeq 0$, we have

$$\kappa(\eta) = \operatorname{tr} Q + [\zeta^2(\eta) - 1]u^T Q u \ \le\ \operatorname{tr} Q + \zeta^2(\eta)\, u^T Q u \ \le \operatorname{tr}(Q) + \zeta^2(\eta)|||Q|||.$$

Applying the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ yields $\sqrt{\kappa(\eta)} \le \sqrt{\operatorname{tr}(Q)} + |\zeta(\eta)|\sqrt{|||Q|||}$. Consequently, we have shown the conditional bound,

$$\mathbb{P}\Big\{\frac{\|Q^{1/2}\,(n^{-1}\sum_{i=1}^n \eta_i w_i - u)\|_2}{n^{-1}\|\eta\|_2} > \sqrt{\operatorname{tr}(Q)} + (\sqrt{n}t + |\zeta(\eta)|)\sqrt{|||Q|||} \;\Big|\; \eta\Big\} \le \exp(-nt^2/2). \tag{5.83}$$

By χ^2-tail bounds, we have $\mathbb{P}\big[|\frac{\|\eta\|_2^2}{n} - 1| \ge t\big] \le \exp(-n\frac{t\wedge t^2}{16})$. Conditioned on the complement of this event, we have $|\zeta(\eta)| \le \frac{t}{n^{-1}\|\eta\|_2}$, and hence conditioning also on the complement of the event in bound (5.83), we are guaranteed that

$$\begin{aligned}\|Q^{1/2}\,(n^{-1}\sum_{i=1}^n \eta_i w_i - u)\|_2 &\le n^{-1}\|\eta\|_2\Big\{\sqrt{\operatorname{tr}(Q)} + (\sqrt{n}t + \frac{t}{n^{-1}\|\eta\|_2})\sqrt{|||Q|||}\Big\}\\ &\le \big(1 + \frac{\|\eta\|_2}{\sqrt{n}}\big)\,\big(\sqrt{\frac{\operatorname{tr}(Q)}{n}} + t\sqrt{|||Q|||}\big),\end{aligned}$$

with probability at least $1 - 2\exp(-n\frac{t\wedge t^2}{16})$. □

Appendix 5.G Bounding an operator norm of a Gaussian matrix

Given a sequence positive numbers $\{\gamma_i\}_{i=1}^m$, consider the $\mathcal{E}_\gamma := \{v \in \mathbb{R}^m : \sum_{i=1}^m \gamma_i^{-1} v_i^2 \leq 1\}$. In this appendix, we derive an upper bound on the operator norm of a standard Gaussian random matrix $W \in \mathbb{R}^{n \times m}$, viewed as an operator from $\mathbb{R}^m$ equipped with the norm induced by $\mathcal{E}_\gamma$, to $\mathbb{R}^n$ equipped with the standard Euclidean norm $\|\cdot\|_2$.

Lemma 47. *Let $W \in \mathbb{R}^{n \times m}$ be a standard Gaussian matrix. Then for all $t \geq 0$,*

$$\mathbb{P}\big[\sup_{v \in \mathcal{E}_\gamma} \|Wv\|_2 > \sqrt{\|\gamma\|_1} + (\sqrt{n} + t)\sqrt{\|\gamma\|_\infty}\big] \leq \exp\big(-\frac{t^2}{2}\big). \tag{5.84}$$

Proof. Let $S^{n-1} := \{u \in \mathbb{R}^n \mid \|u\|_2 = 1\}$ denote the Euclidean unit sphere in $\mathbb{R}^n$. Defining $\mathcal{S} = \{s = (u, v) \mid u \in S^{n-1}, v \in \mathcal{E}_\gamma\}$, consider the Gaussian process $\{Z_s\}_{s \in \mathcal{S}}$ where $Z_s = \langle\!\langle W, uv^T \rangle\!\rangle$. By construction, we have $\sup_{v \in \mathcal{E}_\gamma} \|Wv\|_2 = \sup_{s \in \mathcal{S}} Z_s$. Our approach is to use Slepian's comparison for Gaussian processes [66] in order to bound $\mathbb{E}[\sup_{s \in \mathcal{S}} Z_s]$ by $\mathbb{E}[\sup_{s \in \mathcal{S}} X_s]$, where X_s is a second Gaussian process. Concretely, we define $X_s := \sqrt{\|\gamma\|_\infty}\langle u, g\rangle + \langle v, h\rangle$, where g and h are independent canonical Gaussian vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let $s = (u, v)$ and $s' = (u', v')$ belong to $\mathcal{S}$; by an elementary calculation, we have

$$\mathbb{E}\left[(Z_s - Z_{s'})\right]^2 = \|\!|uv^T - u'v'^T|\!\|_{\mathrm{HS}}^2 \;\leq\; \|\gamma\|_\infty \|u - u'\|_2^2 + \|v - v'\|_2^2 = \mathbb{E}\left[(X_s - X_{s'})^2\right],$$

Consequently, we may apply Slepian's lemma to conclude

$$\begin{aligned}
\mathbb{E}\left[\sup_{s \in \mathcal{S}} Z_s\right] \leq \mathbb{E}\left[\sup_{s \in \mathcal{S}} X_s\right] &= \sqrt{\|\gamma\|_\infty}\, \mathbb{E}\big[\sup_{u \in S^{n-1}} \langle u, g\rangle\big] + \mathbb{E} \sup_{v \in \mathcal{E}_\gamma} \langle v, h\rangle \\
&= \sqrt{\|\gamma\|_\infty}\, \big(\mathbb{E}\, \|g\|_2\big) + \mathbb{E}\, \|Q^{1/2} h\|_2 \\
&\leq \sqrt{\|\gamma\|_\infty}\, \sqrt{n} + \sqrt{\|\gamma\|_1},
\end{aligned}$$

where the final inequality follows by Jensen's inequality, and the relation $\mathrm{tr}(Q) = \|\gamma\|_1$.

Finally, we note that $\|W\|_{\mathcal{E}_\gamma, B_2} = \sup_{v \in \mathcal{E}_\gamma} \|Wv\|_2$ is a Lipschitz function of the Gaussian matrix W, viewed as a vector in ℓ_2^{mn} with Lipschitz constant $\sqrt{\|\gamma\|_\infty}$. Indeed, it is straighforward to verify that $\sup_{v \in \mathcal{E}_\gamma} \|Wv\|_2 - \sup_{v' \in \mathcal{E}_\gamma} \|W'v'\|_2 \leq |\!|\!|W - W'|\!|\!|_{\mathrm{HS}} \sqrt{\|\gamma\|_\infty}$ so that the claim follows by concentration of the canonical Gaussian measure in ℓ_2^{mn} (e.g., see Ledoux [65]). □

Appendix 5.H A uniform law

In this appendix, we state and prove a technical lemma used in parts of our analysis. Consider some subset $\mathcal{D}$ of $\mathbb{R}^m$. Let ν be an index taking values in some index set $\mathcal{I}$. We assume that

ν is indexing a collection of random (noise) matrices Δ_ν. Suppose that there is a collection of nonnegative nondecreasing (possibly random) functions $\mathfrak{r}_\nu : [0,\infty) \to [0,\infty)$ such that for all $t \geq 0$ and $\nu \in \mathcal{I}$

$$\mathbb{P}\Big\{ \sup_{v \in \mathcal{D},\, \|v\|_2 \leq t} G(v;\Delta_\nu) > \mathfrak{r}_\nu(t) \Big\} \leq c_1 \exp[-c_2\, \theta_\nu (t \wedge t^2)], \tag{5.85}$$

where $\theta_\nu, \nu \in \mathcal{I}$ are some positive numbers and G is some function.

Lemma 48. *Under (5.85) and for any collection $\{t_\nu\}_{\nu\in\mathcal{I}}$ such that $\inf_{\nu\in\mathcal{I}} \theta_\nu(t_\nu \wedge t_\nu^2) > 0$, we have for any $\nu \in \mathcal{I}$,*

$$\sup_{v\in\mathcal{D}} \big[G(v;\Delta_\nu) - \mathfrak{r}_\nu(2\|v\|_2)\big] \leq \mathfrak{r}_\nu(2t_\nu). \tag{5.86}$$

with probability at least $1 - \tilde{c}_1 \exp[-c_2\, \theta_\nu(t_\nu \wedge t_\nu^2)]$.

Proof. The proof is based on a peeling argument (e.g., [93]). Define $c := \inf_{\nu\in\mathcal{I}} \theta_\nu t_\nu^2$, and fix some $\nu \in \mathcal{I}$. First, note that as v varies over $\mathcal{D}$, the function $v \mapsto \|v\|_2 \vee t_\nu$ varies over $[t_\nu, \infty)$. Define, for $s \in \{1, 2, \dots\}$,

$$\mathcal{D}_s := \Big\{ v \in \mathcal{D} \,:\, 2^{s-1} t_\nu \leq \big(\|v\|_2 \vee t_\nu\big) < 2^s t_\nu \Big\}.$$

We have $\mathcal{D} = \bigcup_{s=1}^\infty \mathcal{D}_s$. If there exists $v \in \mathcal{D}$ such that

$$G(v,\Delta_\nu) > \mathfrak{r}_\nu\big(2\|v\|_2 \vee 2t_\nu\big), \tag{5.87}$$

then there exist $s \in \{1,2,\dots\}$ and $\mathcal{D}_s \ni v$ such that (5.87) holds for v. Using union bound,

$$\mathbb{P}\Big(\exists v \in \mathcal{D} \,:\, G(v,\Delta_\nu) > \mathfrak{r}_\nu\big(2\|v\|_2 \vee 2t_\nu\big)\Big) \;\leq\; \sum_{s=1}^\infty \mathbb{P}\Big(\exists v \in \mathcal{D}_s \,:\, G(v,\Delta_\nu) > \mathfrak{r}_\nu\big(2\|v\|_2 \vee 2t_\nu\big)\Big).$$

For $v \in \mathcal{D}_s$, (5.87) implies

$$G(v,\Delta_\nu) > \mathfrak{r}_\nu\big(2\|v\|_2 \vee 2t_\nu\big) \geq \mathfrak{r}_\nu(2\, 2^{s-1} t_\nu) = \mathfrak{r}_\nu(\, 2^s t_\nu)$$

where we have used $\mathfrak{r}_\nu$ being increasing. Since $\mathcal{D}_s \subset \{v : \|v\|_2 < 2^s t_\nu\}$, we conclude that

$$\begin{aligned}\mathbb{P}\Big(\exists v \in \mathcal{D} \,:\, G(v,\Delta_\nu) > \mathfrak{r}_\nu\big(2\|v\|_2 \vee 2t_\nu\big)\Big) \;&\leq\; \sum_{s=1}^\infty \mathbb{P}\Big(\sup_{\substack{v\in\mathcal{D},\\ \|v\|_2 < 2^s t_\nu}} G(v,\Delta_\nu) > \mathfrak{r}_\nu(2^s t_\nu)\Big) \\ &\leq\; \sum_{s=1}^\infty \exp[-\theta_\nu\, 2^s (t_\nu \wedge\, t_\nu^2)]\end{aligned}$$

from assumption (5.85). The last summation is bounded above by

$$\sum_{k=1}^\infty \exp[-\theta_\nu\, k\, (t_\nu \wedge t_\nu^2)] = \frac{e^{-\theta_\nu(t_\nu \wedge t_\nu^2)}}{1 - e^{-\theta_\nu(t_\nu\wedge t_\nu^2)}} \leq \frac{e^{-\theta_\nu(t_\nu\wedge t_\nu^2)}}{1-e^{-c}} = C\, e^{-\theta_\nu(t_\nu \wedge t_\nu^2)}.$$

We get the assertion by noting that for $a, b \geq 0$, $\mathfrak{r}_\nu(a \vee b) = \mathfrak{r}_\nu(a) \vee \mathfrak{r}_\nu(b) \leq \mathfrak{r}_\nu(a) + \mathfrak{r}_\nu(b)$ because $\mathfrak{r}_\nu$ is assumed to be nondecreasing and nonnegative. □

Appendix 5.I Some useful matrix-theoretic inequalities

Fan's inequality states that for symmetric matrices A and B and eigenvalues ordered as $\lambda_1(A) \geq \ldots \geq \lambda_m(A)$ (and similarly for B), we have $\operatorname{tr}(AB) \leq \sum_{i=1}^{m} \lambda_i(A)\lambda_i(B)$. As a consequence, for a symmetric matrix B and symmetric matrix $A \succeq 0$, we have

$$\lambda_{\min}(B)\operatorname{tr}(A) \;\leq\; \operatorname{tr}(AB) \;\leq\; \lambda_{\max}(B)\operatorname{tr}(A). \tag{5.88}$$

It follows that for a symmetric matrix $D \succeq 0$ and $R \in \mathbb{R}^{r\times r}$, we have

$$\lambda_{\min}(R^TR)\operatorname{tr}(D) \;\leq\; \operatorname{tr}(DRR^T) \;=\; \operatorname{tr}(R^TDR) \;\leq\; \lambda_{\max}(R^TR)\operatorname{tr}(D), \tag{5.89}$$

where we have used the fact that R^TR and RR^T have the same eigenvalues.

For $B \succeq 0$, we have

$$\lambda_{\min}(B)\,\lambda_j^{\downarrow}(R^TR) \;\leq\; \lambda_j^{\downarrow}(R^TBR) \;\leq\; \lambda_{\max}(B)\,\lambda_j^{\downarrow}(R^TR), \tag{5.90}$$

which can be established using the classical min-max formulation of the j^{th} eigenvalue—namely

$$\lambda_j^{\downarrow}(C) = \max_{\mathcal{M}:\,\dim(\mathcal{M})=j} \;\min_{x\,\in\,\mathcal{M}\,\cap\,S^{r-1}} z^TCz \tag{5.91}$$

where the maximum is taken over all j-dimensional subspaces of $\mathbb{R}^k$. Finally, the inequality (5.90) implies that

$$\|\!|R^TBR|\!\|_{HS} \;\leq\; \|\!|B|\!\|_2\,\|\!|R^TR|\!\|_{HS}. \tag{5.92}$$

Bibliography

[1] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Info. Theory*, 48(3):569–579, March 2002.

[2] J. Ahn, J. S. Marron, K. M. Muller, and Y.-Y. Chi. The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766, 2007.

[3] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, 37(5B):2877–2921, October 2009.

[4] A. A. Amini and M. J. Wainwright. Approximation properties of certain operator-induced norms on Hilbert spaces. Technical report, UC Berkeley, 2010.

[5] A. A. Amini and M. J. Wainwright. Sampled forms of functional PCA in reproducing kernel Hilbert spaces. Technical report, UC Berkeley, 2011.

[6] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2009.

[7] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1984.

[8] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[9] Z. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2nd edition, 2009.

[10] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of large dimensional covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[11] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked populations models. *J. Multivariate Analysis*, 97(6):1382–1408, July 2006.

[12] A. Barrat, M. Barthlemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge Univ. Press, 2008.

[13] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Norwell, MA, 2004.

[14] P. Besse and J. O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311, June 1986.

[15] R. Bhatia. *Matrix Analysis*. Springer, 1996.

[16] P. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume I. Prentice Hall, 2nd edition, 2000.

[17] P. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008.

[18] P. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

[19] L. Birgé. An alternative point of view on lepski's method. In *State of the Art in Probability and Statistics*, number 37 in IMS Lecture Notes, pages 113–133. Institute of Mathematical Statistics, 2001.

[20] G. Boente and R. Fraiman. Kernel-based functional principal components. *Statistics & Probability Letters*, 48(4):335–345, 2000.

[21] B. Bollobás. *Linear Analysis: An Introductory Course*. Cambridge University Press, 2nd edition, 1999.

[22] D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, 2000.

[23] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[24] S. M. Buchman, A. B. Lee, and C. M. Schafer. High-dimensional density estimation via SCA: An example in the modelling of hurricane tracks. *Statistical Methodology*, 8(1):18–30, 2011.

[25] V. V. Buldygin and Y. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society, 2000.

[26] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[27] H. Cardot. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparametric Statist.*, 12(4):503–538, 2000.

[28] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, 1990.

[29] J. B. Conway. *A Course in Functional Analysis*. Springer, 2nd edition, 1990.

[30] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[31] A. d'Aspremont, O. Bannerjee, and L. El Ghaoui. First order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 2007.

[32] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, July 2007.

[33] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.*, 12(1):136–154, March 1982.

[34] K. R. Davidson and S. J. Szarek. *Handbook of Banach Spaces*, volume 1, chapter Local operator theory, random matrices and Banach spaces, pages 317–336. Elsevier, Amsterdan, NL, 2001.

[35] R. DeVore. Approximation of functions. In *Proc. Symp. Applied Mathematics*, volume 36, pages 1–20, 1986.

[36] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, USA, 2002.

[37] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[38] M. Draief and L. Massouli. *Epidemics and Rumours in Complex Networks*. Cambridge Univ. Press, 2009.

[39] N. El Karoui. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics*, 36(6):2717–2756, 2008.

[40] D. Freedman. *Statistical Models: Theory and Practice*. Cambridge Univ. Press, 2009.

[41] R. V. Gamkrelidze, D. Newton, and V. M. Tikhomirov. *Analysis: Convex analysis and approximation theory*. Birkhäuser, 1990. ISBN 0387181792.

[42] D. J. H. Garling. *Inequalities: A Journey into Linear Analysis*. Cambridge Univ. Press, 2007.

[43] S. Geman. A limit theorem for the norm of random matrices. *Annals of Probability*, 8(2):252–261, 1980.

[44] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

[45] I. S. Gradshteyn and I. M. Ryzhik. *Tables of integrals, series, and products*. Academic Press, New York, NY, 2000.

[46] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford Univ. Press, 2001.

[47] C. Gu. *Smoothing spline ANOVA models*. Springer Series in Statistics. Springer, New York, NY, 2002.

[48] A. Guntuboyina. Lower bounds for the minimax risk using f -divergences, and applications. *IEEE Tran*, 57(4):2386–2399, April 2011.

[49] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society*, 68(1):109–126, February 2006.

[50] P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc. B*, 67(3):427–444, 2005.

[51] P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, 34(3):1493–1517, June 2006.

[52] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:94–798, 1978.

[53] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

[54] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, 1990.

[55] I. M. Johnson. On the distribution of the largest eigenvalue in principal components analysis. *Annals. Stat.*, 29(2):295–327, 2001.

[56] I. M. Johnstone. Chi-square oracle inequalities. In M. de Gunst, C. Klaassen, and A. van der Vaart, editors, *State of the Art in Probability and Statistics*, number 37 in IMS Lecture Notes, pages 399–418. Institute of Mathematical Statistics, 2001.

[57] I. M. Johnstone. High dimensional statistical inference and random matrices. In *Proceedings of International Congress of Mathematicians*. European Mathematical Society, 2007.

[58] I. M. Johnstone and A. Lu. Sparse principal components. Technical report, Stanford University, July 2004.

[59] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12:531–547, 2003.

[60] O. Kallenberg. *Foundation of Modern Probability*. Springer, 2004.

[61] N. J. Kalton, N. T. Peck, and J. W. Roberts. *An F-space Sampler*. Cambridge Univ. Press, 1984.

[62] R. W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010.

[63] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1303–1338, 1998.

[64] P. D. Lax. *Functional Analysis*. Wiley-Interscience, 2002.

[65] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[66] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[67] V. A. Marchenko and L. A. Pastur. Distribution for some sets of random of random matrices. *Math. USSR-Sb*, 1:457483, 1967.

[68] J. Matousek. *Lectures on discrete geometry*. Springer-Verlag, New York, 2002.

[69] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[70] S. Mendelson. Geometric Parameters of Kernel Machines. *Lecture Notes In Computer Science*, 2375:29—-43, 2002.

[71] B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.

[72] M. E. J. Newman. *Networks: An Introduction*. Oxford Univ. Press, 2010.

[73] D. Paul. *Nonparametric estimation of principal components*. PhD thesis, Stanford University, 2005.

[74] D. Paul. Asymptotics of sample eigenstructure for a large-dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

[75] D. Paul and I. Johnstone. Augmented sparse principal component analysis for high-dimensional data. Technical report, UC Davis, January 2008.

[76] V. Paulsen. An introduction to the theory of reproducing kernel Hilbert spaces. Technical report, Department of Math., Univ. of Houston, Texas, 2009. URL www.math.uh.edu/~vern/rkhs.pdf.

[77] S. Pezzulli and B.W. Silverman. Some properties of smoothed principal components analysis for functional data. *Comput Stat.*, 8(1):1–16, 1993.

[78] A. Pinkus. *N-Widths in Approximation Theory (Ergebnisse Der Mathematik Und Ihrer Grenzgebiete 3 Folge)*. Springer, 1985. ISBN 038713638X.

[79] A. Pinkus. N-widths and optimal recovery. In *Proc. Symp. Applied Mathematics*, volume 36, pages 51–66, 1986.

[80] J. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2005.

[81] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer, 2002.

[82] J. A. Rice and B. W. Silverman. Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves. *Journal of the Royal Statistical Society*, 53(1):233 – 243, 1991.

[83] F. Riesz and B. Sz.-Nagy. *Functional Analysis*. Dover Publications, 1990. ISBN 0486662896.

[84] G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[85] M. Rudelson. Random Vectors in the Isotropic Position,. *Journal of Functional Analysis*, 164(1):60–72, May 1999. doi: 10.1006/jfan.1998.3384.

[86] B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1):1–24, February 1996.

[87] T. Speed. *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall/CRC, 2003.

[88] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[89] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[90] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info Theory*, 52(3):1030–1051, March 2006.

[91] J. A. Tropp. User-friendly tail bounds for sums of random matrices. April 2010. URL: http://arxiv.org/abs/1004.4389.

[92] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge Univ. Press, 2005.

[93] S. A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[94] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ.Press, 2000.

[95] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, H.L. Mao, M.and Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002.

[96] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, March 1996.

[97] R. Vershynin. Lectures in functional analysis. Technical report, Dept.of Math, Univ. of Michigan, 2010. URL http://www-personal.umich.edu/~romanv/teaching/2010-11/602/functional-analysis.pdf.

[98] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok., editors, *Compressed Sensing*. Cambridge University Press, to appear. URL http://www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf.

[99] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

[100] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using the lasso. Technical Report Technical Report 709, Department of Statistics, UC Berkeley, 2006.

[101] M. J. Wainwright. Information-theoretic bounds for sparsity recovery in the high-dimensional and noisy setting. Technical Report Technical Report 725, Department of Statistics, UC Berkeley, 2007.

[102] A. Wigderson and D. Xiao. Derandomizing the Ahlswede-Winter matrix-valued Chernoff bound using pessimistic estimators, and applications. *Theory of Computing*, 4(1): 53–76, 2008. doi: 10.4086/toc.2008.v004a003.

[103] E. P. Wigner. On the distributions of the roots of certain symmetric matrices. *Ann. Math.*, 67:325327, 1958.

[104] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[105] F. Yao, H.-G. Müller, and J.-L. Wang. Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 100(470):577–590, June 2005.

[106] Y. Q. Yin. Limiting spectral distribution for a class of random matrices. *J. Multivariate Anal.*, 20:5068, 1986.

[107] B. Yu. Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, Berlin, 1997.

[108] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[109] Z. Zhang, H. Zha, and H. Simon. Low-rank approximations with sparse factors I: Basic algorithms and error analysis. *SIAM Journal on Matrix Analysis and Applications*, 23(3):706–727, 2002.

[110] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):262–286, 2006.