STRUCTURED STATISTICAL ESTIMATION VIA OPTIMIZATION

A Dissertation
Presented to
The Academic Faculty

By

Andrew D. McRae

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2022

© Andrew D. McRae 2022

STRUCTURED STATISTICAL ESTIMATION VIA OPTIMIZATION

Thesis committee:

Dr. Mark Davenport
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Vidya Muthukumar
Electrical and Computer Engineering and
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Justin Romberg
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arkadi Nemirovski
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Vladimir Koltchinskii
Mathematics
Georgia Institute of Technology

Date approved: April 15, 2022

Oh give thanks to the LORD, for he is good,
for his steadfast love endures forever!
Let the redeemed of the LORD say so,
whom he has redeemed from trouble
and gathered in from the lands,
from the east and from the west,
from the north and from the south.

Some wandered in desert wastes,
finding no way to a city to dwell in;
hungry and thirsty,
their soul fainted within them.
Then they cried to the LORD in their trouble,
and he delivered them from their distress.
He led them by a straight way
till they reached a city to dwell in.
Let them thank the LORD for his steadfast love,
for his wondrous works to the children of man!
For he satisfies the longing soul,
and the hungry soul he fills with good things.

Psalm 107:1–9 (ESV)

This thesis, defended on Good Friday, is dedicated to the glory of God, who made this world and made me with the mathematical skill to understand and build in it, and who redeemed me from sin and death to live for him by his son, Jesus.

ACKNOWLEDGMENTS

Life is a highly nonconvex optimization landscape in which effective maximization requires traversing deep valleys. There would be no Easter without Good Friday. There is no Ph.D. and the intellectual achievement and growth it signifies without long years of difficult and often tedious and seemingly aimless toil. I want to thank and acknowledge the many people who have supported me through this process, and I thank God for providing them.

My advisor, Mark Davenport, has made this journey as intellectually fruitful and interesting and as bureaucratically straightforward as any advisor could. He provided enough direction to keep me from getting stuck while leaving me alone enough to learn to be an independent and resourceful researcher. Much of the same is true for my other close mentor and collaborator, Justin Romberg, who, in addition, is one of the best classroom teachers I have ever had (and in that capacity was a significant influence on my decision to start a Ph.D.). As I have written papers and planned my future, both of them have helped me greatly with their valuable technical suggestions, their incisive feedback, and their insightful, honest, and often witty advice.

I am also thankful for my other thesis committee members and collaborators: Vladimir Koltchinskii, Vidya Muthukumar, Arkadi Nemirovski, and Santhosh Karnik. The statistics and optimization courses I took from Vladimir and Arkadi have been foundational for much of my technical work. Vidya and Santhosh, through our collaboration, have greatly contributed to this thesis.

I am also thankful for the Children of the Norm lab group, which has been a great source of fun and camaraderie as we have all gone through the Ph.D. process together. There are too many people to list, but I am especially thankful for the friendships, advice, and collaborations of Michael Moore, Santhosh Karnik, Namrata Nadagouda, and Austin Xu.

In the personal realm, I am very thankful for my family, especially my parents, Brian and Elisabeth McRae. They have been my best cheerleaders and a reliable source of advice,

encouragement, recreation, meals, and a place to stay whenever I wanted or needed any of these things.

I am also very thankful for my church family at Westminster Presbyterian Church of Atlanta. The faithful and excellent preaching and teaching of God's word by Aaron Messner, Carlton Wynne, and others has greatly edified me. This church has been a constant encouragement and has given me the opportunity to serve the church and to do and learn many things I would never have done on my own (including singing, woodworking, leadership, and property maintenance). Again, there are too many people to list, but I am particularly grateful for the friendship of the Akins family, the Davila family, Andrew Jarrett, the Lieuwen family, the Melendez family, Angie Mercer, and Jerry Naff.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

This thesis shows how we can exploit low-dimensional structure in high-dimensional statistics and machine learning problems via optimization. We show several settings where, with an appropriate choice of optimization algorithm, we can perform useful estimation with a complexity that scales not with the original problem dimension but with a much smaller intrinsic dimension.

In the low-rank matrix completion and denoising problems, we can exploit low-rank structure to recover a large matrix from noisy observations of some or all of its entries. We prove state-of-the-art results for this problem in the case of Poisson noise and show that these results are minimax-optimal.

Next, we study the problem of recovering a sparse vector from nonlinear measurements. We present a lifted matrix framework for the sparse phase retrieval and sparse PCA problems that includes a novel atomic norm regularizer. We prove that solving certain convex optimization problems in this framework yields estimators with near-optimal performance. Although we do not know how to compute these estimators efficiently and exactly, we derive a principled heuristic algorithm for sparse phase retrieval that matches existing state-of-the-art algorithms.

Third, we show how we can exploit low-dimensional manifold structure in supervised learning. In a reproducing kernel Hilbert space framework, we show that smooth functions on a manifold can be estimated with a complexity scaling with the manifold dimension rather than a larger embedding space dimension.

Finally, we study the interaction between high ambient dimension and a lower intrinsic dimension in the harmless interpolation phenomenon (where learned functions generalize well despite interpolating noisy data). We present a general framework for this phenomenon in linear and reproducing kernel Hilbert space settings, proving that it occurs in many situations that previous work has not covered.

CHAPTER 1
INTRODUCTION

In many modern statistics and machine learning problems, we are trying to estimate an object that has a very high number of degrees of freedom. In imaging (photographic, medical, seismic, etc.), a high-resolution 2D or 3D image could have millions or even billions of pixels/voxels. In machine learning applications such as image classification or fitness tracking, we need to estimate a function that has a very high-dimensional domain (and, being a function, has infinite degrees of freedom). Classical statistics and learning theory tell us that we cannot make any meaningful estimates without a very large amount of data. However, real problems often have underlying structure that makes them more tractable.

This thesis studies how structure plays a role in two basic classes of problem: in the *recovery* problem, we want to recover a (large) set of unknown parameters via measurements that may be noisy, indirect, or incomplete; in the *learning* problem, we are given many samples of some pairing (feature, label), and we want to predict future labels from only the feature data.

For the recovery problem, one example is the way a medical CT scanner produces a 3-D image of (part of) a patient's body by sending X-rays through the body at many different angles. The more measurements we make, the more time and radiation exposure is required. To recover a 1000^3-pixel 3-D volume, we classically need at least 10^9 measurements. Another example is predicting user ratings in a recommendation system (e.g., Netflix); we want to predict all possible user/item ratings from the tiny fraction for which we have actual ratings. In both cases, hidden low-dimensional structure can help us. Real-world images are often *sparse* (i.e., mostly zero) in a suitable transformed representation (this is why image compression works). Similarly, for a recommendation system, if the number of factors that

determine user/item ratings is small, the resulting ratings matrix will be *low-rank*. Sparsity and low rank are types of structure that we can exploit to make recovery more tractable and efficient.

For the learning problem, a common modern example is learning to classify images based on many labeled example images. A much simpler example is linear regression, in which we try to fit a linear function to data. In general, this is the problem of estimating a function from samples of its values. The task difficulty depends on the complexity of the function and on the complexity of the set of possible (or likely) inputs. A completely arbitrary function is impossible to estimate, but most reasonable functions depend smoothly on their input (or are otherwise somehow "regular"); for such a function, we can make meaningful inferences about its value away from the points where we sample it. Similarly, the domain that actually matters may be far simpler than a naïve representation would suggest; for example, a 1-megapixel image in principle has 10^6 degrees of freedom, but real images are not simply 10^6 arbitrary pixel values (almost all such images would look like static on a TV screen). We want to understand how such structure (domain and smoothness) determines the difficulty of learning a function.

A common theme throughout this thesis is the role of *optimization* in estimation. A general model for both the recovery and learning problems is that we observe some data $(x_i, y_i), i = 1, \ldots, n$, where x_i are known and $y_i \approx f(x_i, \beta^*)$, where the function f encodes the model, and β^* is a vector of unknown parameters. A common method to find an estimate is to solve an optimization program of the form

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \ell_f(y_i, x_i, \beta),$$

where $\ell_f(y, x, \beta)$ is some "loss" function that measures how well the parameter β explains the data pair (x, y). Almost all maximum likelihood estimates take this form (where $\ell_f(y, x, \beta)$ is the negative log-likelihood of the observations y given x and β).

Suppose the parameter vector β^* has structure that we want to exploit. If $r(\beta)$ is some function that is large for "unstructured" candidate parameter vectors β but is small for highly structured vectors like β^*, we can solve the following *regularized* optimization program:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \ell_f(y_i, x_i, \beta) + r(\beta).$$

The additional penalty encourages solutions that are structured.

A simple example model illustrates how structure and optimization play a role in both recovery and learning. Consider the classical linear model, where, for some vector β^*, we make n noisy linear measurements of the form $y_i \approx \langle x_i, \beta^* \rangle$, $i = 1, \ldots, n$, where the x_i's are design vectors. We will suppose that the design vectors are chosen independently at random from some probability distribution. A typical estimation procedure is to solve a least-squares optimization problem of the form

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \langle x_i, \beta \rangle)^2 \ (+ \text{ regularization}).$$

In the recovery problem we care about estimating β^* itself. If $\beta^* \in \mathbf{R}^p$, then, classically, we need $n \gtrsim p$ measurements to obtain a unique solution to the above optimization problem, and our estimate $\hat{\beta}$ will have squared ℓ_2 error of order $\|\hat{\beta} - \beta^*\|_{\ell_2}^2 \approx p/n$.[1] However, if the vector β^* is sparse (say, s nonzero elements), it is well-known that, if we add an ℓ_1-norm regularization to the least-squares problem (the LASSO algorithm), we only need (within logarithmic factors) $n \gtrsim s$ measurements, and the error will scale like s/n. The sparse structure makes recovery easier in terms of sample complexity and corruption due to noise.

On the other hand, in the learning problem, we only need to predict future observations $y \approx \langle x, \beta^* \rangle$, where x has the same probability distribution as the design vectors. Note that we only need to recover the components of β^* that significantly affect the observations;

[1] This assumes that the design vectors x_i are isotropic (i.e., $\mathbf{E}\, x_i \otimes x_i = I_p$) and the measurement errors are random, independent, and zero-mean.

this is potentially a much easier problem than recovering β^* itself. The covariance of the design vectors given by $\Sigma = \mathbf{E}\, x_i \otimes x_i$ turns out to be key; the number of measurements we need and our prediction error will depend on the (effective) rank of Σ. Here, structure in the design covariance determines problem difficulty. A near-optimal predictor of y can be obtained via ridge regression, in which we solve the above least-squares problem with a simple ℓ_2-norm regularizer.

1.1 General approach

This thesis considers several different forms of low-dimensional structure. These include sparsity, low rank (both in matrix estimation/completion and linear regression), and manifold structure. Later, I will describe and motivate each of these and explain my specific methods for each problem in more detail. Here, however, I want to describe the general themes of my work, using examples from my work for illustration.

All of my work in this thesis exploits structure with estimators defined by optimization problems. For each type of structure, we must design a suitable optimization program that takes advantage of that structure; understanding the properties of solutions to such programs is a key part of this. Then, for real-world applications, we must design algorithms that can be practically implemented and run on a computer. Finally, our theoretical guarantees assume randomly-sampled data; this is often a natural assumption, and it often gives us (with high probability) very effective experiment designs even when doing so "by hand" is difficult.

1.1.1 Analyzing and encoding structure

To take advantage of structure via optimization, the first step is to choose an optimization algorithm whose output is an estimator that captures the problem's structure. For example, if we are trying to estimate a sparse vector, algorithms taking advantage of this sparsity typically have an explicit mechanism for encouraging sparse estimators, such as an ℓ_1 norm.

Another example is trying to recover a low-rank matrix. A common way to encode low

rank in an estimator is to use a convex program with a nuclear norm penalty; the nuclear norm promotes low rank because it is the ℓ_1 norm on the singular values of a matrix. This is my approach to the matrix completion problem in this thesis (Chapter 2).

Encoding even simple types of structure in an optimization problem can be quite challenging when the measurements are nonlinear. The simplest formulations of phase retrieval and PCA result in nonconvex programs; a common technique to transform these problems to be convex is to "lift" the vector parameters into a matrix space. Previous attempts to exploit sparsity in this framework gave highly suboptimal performance, since encoding the sparse-and-low-rank structure of the resulting matrix parameter is delicate. My work on sparse phase retrieval and PCA (Chapter 3) develops a novel way to encode sparse-and-low-rank structure. The key tool is a new matrix norm that we can use as regularization in a convex program.

Another quite different type of structure is *manifold* structure. This knowledge can also be exploited by an optimization program. In my work on manifold regression (Chapter 4), I develop a framework for studying and estimating functions on manifolds; the key idea is that smooth functions on a manifold lie in special intrinsic function spaces (reproducing kernel Hilbert spaces or RKHSs) defined in terms of the manifold's intrinsic spectral decomposition. We can take advantage of this structure and smoothness by performing kernel (RKHS) regression, which is an optimization program that forces solutions to be smooth functions (and can be easily solved via a kernel function associated with the function space).

My approach to RKHS regression (which also includes the work in Chapter 5) relies on the (approximate) low rank of the measurement covariance operator as in the linear regression example above. In the kernel case, the covariance is the integral operator of the kernel on the function input domain. The number of function samples we need and the error due to noise scale proportionally to the effective rank of this integral operator.

1.1.2 Properties of estimates obtained with optimization

To understand why certain optimization algorithms work well for statistical estimation and machine learning, we must carefully study the properties of estimates that these algorithms return. Some of these properties (sparsity, low rank, smoothness on a manifold, etc.) are ones we explicitly design the algorithm to exploit. Other properties seem to be present for no obvious reason. Both categories of estimator properties are important.

The previous section outlined a variety of ways in which we can design optimization algorithms to exploit problem structure. To prove error guarantees, we must understand how these algorithmic biases do, in fact, result in good estimates. For example, when estimating a sparse vector, we may set up an algorithm to encourage sparsity in the solution. However, it is usually not enough simply to say that the resulting estimate is sparse; we must carefully analyze the geometry of the problem to understand how the estimator contains the desired structure and how it compares to the true parameter vector.

In addition (and much less obviously), estimators can have very interesting properties that we did not explicitly encourage (or, in many cases, expect). A well-known example of this is the interpolation phenomenon in supervised learning. A recent empirical observation is that in many models with a very large number of parameters (e.g., deep neural networks), the learned function can interpolate *noisy* training data and still perform well when tested on new data. My contribution to this area is a new theoretical framework for understanding why this happens in linear regression and classification models (Chapter 5). We show that the fundamental reason is an implicit regularization (or smoothing) that arises when we have a very large number of individually insignificant parameters.

1.1.3 Practical optimization on large-scale problems

For any discussion of optimization algorithms to be useful in the real world, we must be able to implement something on a computer. Furthermore, in order to solve the large-scale problems that arise in modern applications, these algorithms must be efficient.

I consider this most explicitly in my work on sparse phase retrieval and PCA (Chapter 3). I derive an abstract convex program that yields near-optimal results, but it is not clear how to solve it exactly and efficiently. Much of Chapter 3 is devoted to deriving a principled implementable algorithm, and it took considerable time to implement an efficient first-order algorithm that could solve large problems.

This issue is also present in my other work. In my matrix completion work, an optimal estimator can be computed with a single step of singular value thresholding; even for very large problems, this can be tractably computed with iterative algorithms, especially when our samples of the matrix are sparse. My work on manifold regression and my work on the interpolation phenomenon both use a reproducing kernel Hilbert space (RKHS) framework; this allows us practically to solve regression and classification problems with a very large or infinite number of degrees of freedom.

1.1.4 Randomized data acquisition (experiment design)

Randomized measurements are a fundamental part of all my work. In Chapters 3 to 5, I assume that the data pairs (x_i, y_i) are all drawn independently from the same distribution. In Chapter 2, I assume that we subsample the entries of a matrix at random (each entry being sampled with the same probability independently of the others).

This type of assumption is standard for the learning problem. In the statistical theory for this problem, we typically assume that the data samples $(x_1, y_1) \ldots, (x_n, y_n)$ and future "test" data are all drawn independently from the same distribution.

For the recovery problem, we often have control over the experiment design variables $x_1, \ldots, x_n$. However, choosing the design variables at random is often (nearly) optimal. Intuitively, random measurements work well in structured high-dimensional problems because (if properly distributed) they can, with very high probability, "discover" the hidden structure in a problem even if we don't know where to look for it. For example, in linear sparse recovery, accurate recovery requires the design matrix (the matrix with the x_i's

as rows) to satisfy a certain *restricted isometry property*; verifying that any given matrix satisfies this property is computationally intractable, but we can prove that a matrix with randomly chosen rows will work well with very high probability (see, e.g., [1, Chapter 6]).

1.2 Thesis overview

1.2.1 Low-rank matrix completion and denoising under Poisson noise

In Chapter 2, I consider the problem of recovering a low-rank matrix M from noisy observations of all or a subset of its entries. The low-dimensional structure that I consider is low matrix rank. We analyze several estimators computed via convex programs with nuclear norm regularization to promote low-rank solutions.

For the specific case of Poisson noise, we prove that these estimators achieve minimax-optimal bounds (in the Frobenius norm error metric) that depend on the matrix rank, the fraction of the elements observed, and maximal row and column sums of the true matrix. If M is a nonnegative matrix with rank r, and we sample the entries randomly (each entry being sampled with probability p independently of the others), and observe a $\mathrm{Poisson}(M_{ij})$ random variable for each sampled entry, our estimator achieves, with high probability,

$$\|M - \widehat{M}\|_F \lesssim \sqrt{\frac{r}{p}}(\max_i \sqrt{\sum_j (M_{ij} + (1-p)M_{ij}^2)} + \max_j \sqrt{\sum_i (M_{ij} + (1-p)M_{ij}^2)})$$

when p is large enough. We also extend these results to handle the case of matrix multinomial denoising. Note the key role of the rank parameter (which controls the intrinsic dimension) in the error bound.

1.2.2 Lifted sparse phase retrieval/PCA

In the phase retrieval problem (one common formulation), we want to recover a vector $\beta \in \mathbf{R}^p$ from (noisy) squared measurements of the form $y_i \approx |\langle x_i, \beta \rangle|^2$. It has lately been shown that the difficulty of this problem scales linearly in the dimension p as in the linear

measurement case. However, in the *sparse* case, the best theoretical results for practical algorithms require $n \gtrsim s^2$ measurements to recover an s-sparse vector.

In Chapter 3, I formulate a novel convex relaxation of the sparse phase retrieval problem via the lifting technique (quadratic measurements are linear when "lifted" to a matrix space as in $|\langle x, \beta \rangle|^2 = \langle x \otimes x, \beta \otimes \beta \rangle$) along with a novel matrix norm regularizer. We show that the resulting estimator does indeed achieve sample and noise complexity performance of the same order as in linear sparse recovery (in particular, complexity nearly proportional to the sparsity s). We apply the same technique with similar results to the sparse principal component analysis (sparse PCA) problem, which has had similar theoretical shortcomings. While our convex programs are abstract, and we do not know efficient algorithms for solving them, for the case of sparse phase retrieval we derive a principled heuristic and show empirically that the resulting *nonconvex* algorithm matches existing state-of-the-art sparse phase retrieval algorithms.

1.2.3 Learning on a manifold

Classically, the complexity of learning a function depends exponentially the dimension of the function's domain. If the domain is high-dimensional (e.g., we are trying to classify images with millions of pixels), this theory gives no hope of meaningful learning. However, very often, the domain of interest (e.g., the set of realistic images) has a much lower intrinsic dimension; a common model is that the domain is a low-dimensional manifold embedded in the higher-dimensional space.

In Chapter 4, I study manifold domains through a reproducing kernel Hilbert space (RKHS) framework; in particular, I study certain special RKHSs intrinsic to a manifold (e.g., corresponding to the heat kernel) and show that the difficulty of learning a function in such a space scales with the intrinsic manifold dimension rather than the embedding dimension.

The algorithm I analyze is kernel regression with the intrinsic manifold kernels; the regression estimate can be described as the solution to a convex regularized least-squares

optimization problem. With a kernel method, the solution can be computed practically, even though the function space over which we are optimizing can be infinite-dimensional.

1.2.4 Harmless interpolation in regression and classification

Classical learning theory suggests that when trying to estimate a function from noisy samples, we should not attempt to interpolate exactly the corrupted observations (we will overfit to the data, resulting in large "variance" error due to noise). However, recent research has shown that empirically, interpolating noisy samples does not seem to be a significant problem in certain highly overparametrized settings (i.e., when the number of degrees of freedom in our model is much larger than the number of samples).

Most prior theoretical results on this topic assume highly restrictive linear models. In Chapter 5, I formulate a novel analysis framework and show that this "harmless interpolation" occurs in a much more general reproducing kernel Hilbert space framework. Again, in an RKHS framework, we can write down the natural minimum-norm interpolator as a convex program.

An RKHS can be described by a sequence of numbers $\{\lambda_\ell\}_{\ell=1}^{\infty}$ (eigenvalues of an integral operator), typically arranged in decreasing order. Larger eigenvalues correspond to components of the space with more important information; the decay determines the intrinsic dimension of the function space.

To get good regression performance with finite samples, the RKHS must have a small enough intrinsic dimension. The key to getting good generalization even in the presence of noise is that the RKHS must have a very large number of small components (eigenvalues); λ_ℓ must decay quickly (to get good learning performance on the important components), but not *too* quickly. A large number of extra degrees of freedom yields an "implicit regularization" that causes our estimate to behave (except extremely close to sample points) as though we had added explicit regularization to our problem (a typical way to smooth out the noise error).

Although this problem is typically studied for *regression*, we also show how our framework can be applied to the *classification* problem. In accord with the fact that classification is fundamentally the easier problem, we demonstrate settings in which regression is not consistent but classification is (which had previously only been shown for a much more restricted class of problems).

Computationally, the kernel method again allows us to compute estimates efficiently, even though the function space is (necessarily) very high-dimensional.

CHAPTER 2

LOW-RANK MATRIX COMPLETION AND DENOISING UNDER POISSON NOISE

In this chapter,[1] we consider the problem of estimating a low-rank matrix from the observation of all or a subset of its entries in the presence of Poisson noise. When we observe all entries, this is a problem of *matrix denoising*; when we observe only a subset of the entries, this is a problem of *matrix completion*. In both cases, we exploit an assumption that the underlying matrix is *low-rank*. Specifically, we analyze several estimators, including a constrained nuclear-norm minimization program, nuclear-norm regularized least squares, and a nonconvex constrained low-rank optimization problem. We show that for all three estimators, with high probability, we have an upper error bound (in the Frobenius norm error metric) that depends on the matrix rank, the fraction of the elements observed, and maximal row and column sums of the true matrix. We furthermore show that the above results are minimax optimal (within a universal constant) in classes of matrices with low rank and bounded row and column sums. We also extend these results to handle the case of matrix multinomial denoising and completion.

2.1 Introduction

2.1.1 Low-rank models for count data

We consider the problem of estimating a non-negative matrix $M \in \mathbf{R}^{m \times n}$ given independent observations distributed according to $\mathrm{Poisson}(M_{ij})$ for $(i,j) \in \Omega$, where Ω is a (not necessarily strict) subset of $\{1,\ldots,m\} \times \{1,\ldots,n\}$. If we do not make an observation for every entry of the matrix, the recovery problem is, in general, ill-posed in the absence of

[1] This work is published in [2].

any additional assumptions on the underlying matrix. A common assumption for this type of problem is that the unknown matrix M is *low-rank*; i.e., the dimension of the spans of the columns and rows of M is much smaller than the actual numbers of columns and rows. This assumption greatly reduces the number of degrees of freedom in the model, making the recovery problem far more tractable. Note that even if we do observe every entry, we can still exploit the structure of the model to reduce the error due to noise.

While the problems of matrix completion and denoising have received a significant amount of attention in the settings of Gaussian noise and of small, bounded (in ℓ_2) perturbations (e.g., [3, 4, 5]), Poisson noise models have received comparatively less attention. In this chapter, we focus primarily on the Poisson model, but we also examine closely-related multinomial models; this includes the case in which we have a single probability distribution over matrix coordinates (for which our result is a corollary of our Poisson results) as well as the case in which we make independent multinomial observations of matrix rows. Collectively, these models are often natural in applications where the observations arise via some form of counting process. The ability to recover (or de-noise) a low-rank signal from noisy, count-based observations is useful in many situations. We briefly mention two examples.

One potential application area involves imaging systems. This includes conventional cameras (which often suffer from noise in low light or with short exposures), but also 3-D imaging methods such as X-ray computed tomography (CT) and positron emission tomography (PET), which, in medical imaging, would greatly benefit from an improved noise/radiation dose tradeoff. In these scenarios, the Poisson noise model is natural because the observations consists of counts of particle (e.g., photon) arrivals at a detector. In many of these settings, such as when observing a periodic or slowly-varying sequence of images, a low-rank assumption on the underlying data is natural (see, e.g., [6] for an overview of low-rank modeling in image applications).

Another important application is topic modeling, which is a common form of dimensionality reduction for text documents. In this case, our observations consist of counts of

word occurrences in a corpus of documents. If we suppose that these documents can be decomposed according to a small set of topics, and that within each topic documents will exhibit similar word occurrence counts, then a low-rank assumption on the word-frequency matrix is natural. For example, the popular PLSI model [7] uses a multinomial probability model parameterized by a low-rank matrix.

Low-rank models are also popular in nonnegative matrix factorization, which is commonly applied in a range of contexts where count data is common. A wide variety of such models have been developed, along with inference algorithms such as expectation maximization, variational Bayes, and Markov chain Monte Carlo [8, 9, 10, 11]. These models and algorithms have been applied to many tasks, especially recommendation systems [12, 13]. However, the algorithms used are nonconvex, and there is little in the way of theoretical guarantees for their performance in the Poisson or multinomial setting.

2.1.2 Summary of main results for Poisson noise

In our analysis, we assume a Bernoulli sampling of the matrix entries: i.e., the events $\{(i,j) \in \Omega\}$ are independent with probability $p \in (0,1]$, and the observed Poisson random variables are independent conditioned on Ω. Note that taking $p = 1$ handles the case in which we observe every entry of the matrix. For a matrix A, $\|A\|_*$, $\|A\|$, and $\|A\|_F$ denote the nuclear norm, operator norm, and Frobenius norm of A, respectively.

Let $\mathcal{A}_\Omega \colon \mathbf{R}^{m \times n} \to \mathbf{R}^\Omega$ denote the entry-wise sampling operator given by $(\mathcal{A}_\Omega(Z))_{(i,j)} = Z_{ij}$ for $(i,j) \in \Omega$. Note that its adjoint $\mathcal{A}^* \colon \mathbf{R}^\Omega \to \mathbf{R}^{m \times n}$ maps the vector $(x_{(i,j)})_{(i,j)\in\Omega} \in \mathbf{R}^\Omega$ to the $m \times n$ matrix whose (i,j)th entry is $x_{(i,j)}$ if $(i,j) \in \Omega$ and zero otherwise.

Given observations $X \sim \mathrm{Poisson}(\mathcal{A}_\Omega(M))$, we consider several different estimators with similar theoretical properties. The first can be interpreted as a matrix version of the *Dantzig selector* [14]:

$$\widehat{M}^{(1)} = \operatorname*{arg\,min}_{M' \in [0,\infty)^{m \times n}} \|M'\|_* \text{ s.t. } \|\mathcal{A}_\Omega^*(X) - pM'\| \leq \delta, \tag{2.1}$$

where $\delta > 0$ is a parameter which we will see how to set later. The second is a nuclear-norm-regularized least-squares type estimator:

$$\widehat{M}^{(2)} = \underset{M' \in [0,\infty)^{m \times n}}{\arg\min} \|\mathcal{A}_\Omega^*(X) - pM'\|_F^2 + \lambda \|M'\|_*, \tag{2.2}$$

where $\lambda > 0$ is another parameter which we will set. The third is least-squares under an exact low-rank constraint:

$$\widehat{M}^{(3)} = \underset{M' \in [0,\infty)^{m \times n}}{\arg\min} \|\mathcal{A}_\Omega^*(X) - pM'\|_F \text{ s.t. } \operatorname{rank}(M') \leq r, \tag{2.3}$$

where r is an upper bound on the rank of the true rate matrix M. This problem is not convex (and is, in general, hard to solve directly while respecting nonnegativity constraints), but we will see later how we can address this issue without affecting its theoretical properties.

Theorem 1, which is the main result of Section 2.2.1, states that, if M has rank r, and hyperparameters are properly chosen, each of the estimators $\{\widehat{M}^{(i)}\}_{i=1}^3$ satisfies, with high probability,

$$\|M - \widehat{M}^{(i)}\|_F \lesssim \sqrt{\frac{r}{p}} \widetilde{\sigma}(M) + \text{logarithmic terms}, \tag{2.4}$$

where

$$\widetilde{\sigma}(M) = \max_i \sqrt{\sum_j (M_{ij} + (1-p)M_{ij}^2)} + \max_j \sqrt{\sum_i (M_{ij} + (1-p)M_{ij}^2)}.$$

In many situations (see Section 2.3.3), the logarithmic terms are negligible, so we can approximate this result by the bound

$$\|M - \widehat{M}^{(i)}\|_F \lesssim \sqrt{\frac{r}{p}} \widetilde{\sigma}(M). \tag{2.5}$$

Section 2.3 uses two standard methods to find lower bounds on the minimax risk of any estimator in classes of matrices with bounded row and column sums. These results

(Theorems 3 and 4) can be summarized as follows: over all nonnegative matrices M such that $\mathrm{rank}(M) \leq r$, and $\widetilde{\sigma}(M) \leq \sigma$, we have

$$\inf_{\widehat{M}} \sup_{M} \mathbf{E}_M \|M - \widehat{M}\|_F \gtrsim \sqrt{\frac{r}{p}}\sigma.$$

Thus Theorem 1 is optimal (up to a multiplicative constant and an additive logarithmic factor) for this class of matrices.

To gain a more intuitive understanding of our result, it is helpful to examine the formula for $\widetilde{\sigma}$. For simplicity, assume, without loss of generality, that the row sums dominate the column sums, so that

$$\widetilde{\sigma} \approx \max_i \sqrt{\sum_j (M_{ij} + (1-p)M_{ij}^2)}.$$

The two terms inside the sum have different roles. The first term (M_{ij}) corresponds to the variance of the Poisson random variables. Indeed, if we take $p = 1$, this is the only term, so our result has the form

$$\|M - \widehat{M}^{(i)}\|_F \lesssim \sqrt{r}\left(\max_i \sqrt{\sum_j M_{ij}}\right).$$

If we do not impose any structure on the model, the maximum likelihood (and least-squares) estimate is $\widehat{M}^{\text{MLE}} = X$, which has risk

$$\mathbf{E}\|M - \widehat{M}^{\text{MLE}}\|_F^2 = \sum_{i,j} \mathrm{var}(X_{ij}) = \sum_{i,j} M_{ij}.$$

If every row of M has approximately the same sum, estimators defined above improve on $\widehat{M}^{\text{MLE}}$ (in squared Frobenius error) by a factor of approximately r/n. If the sums of the rows of M differ significantly, the improvement is smaller. However, this should not be too surprising — if the variance in the problem is already concentrated into a smaller sub-matrix, we are effectively solving a smaller problem, and hence the low-rank assumption is less

restrictive and, therefore, less beneficial.

The second term (of the form $(1-p)M_{ij}^2$) in the formula for $\widetilde{\sigma}$ corresponds to the inherent difficulty in estimating the values of a matrix due to the fact that we do not observe every entry. This term in the lower bound applies regardless of the noise model, even when there is no noise. This might seem to contradict existing exact noiseless matrix completion results, but we note here that such results make stronger assumptions (incoherence of the row and column spaces) beyond what we are assuming here. In fact, the matrices used in the proof of Theorem 4 are highly coherent.

Although this second error term is necessary for general matrices, an interesting open problem is whether it could be entirely removed (leaving only the variance term) when we assume additional structure (such as incoherence) on the true rate matrix. Such a result would be a bridge between existing noisy and noiseless matrix completion literature; the existence of exact completion for the noiseless case implies that current results for the noisy case (including this work) become highly suboptimal when the signal-to-noise ratio goes to infinity. An exception is [5], but we note that this approach is not without its own drawbacks as this approach leads to error rates which are suboptimal with respect to the rank r. More recent work in this direction is [15, 16]; however, these results depend strongly on the condition number of the true matrix, and their dependence on matrix rank seems suboptimal. Thus there is still much work to do in analyzing noisy completion of incoherent matrices.

2.1.3 Summary of main results for multinomial denoising

We can also derive an interesting result on multinomial matrix denoising as a corollary of our result on Poisson denoising. If P is a non-negative $m \times n$ matrix such that $\sum_{i,j} P_{ij} = 1$, and we independently sample N objects according to the probabilities contained in P, the number of times each entry of P is sampled (which we can denote by an $m \times n$ count matrix X) has a (matrix) multinomial distribution. Our results on Poisson denoising apply by considering a multinomial distribution to be a vector of independent Poisson variables

conditioned on its sum. Corollary 1 in Section 2.2.2 shows that if P has rank r, and none of the row and column sums of P are too large, there is an estimator $\widehat{P}$ of P (which could be defined similarly to any of the estimators above) such that

$$\|\widehat{P} - P\|_F \lesssim \sqrt{\frac{r}{N(m \wedge n)}}$$

One can easily check that, for the maximum likelihood estimator $\widehat{P}^{\text{MLE}} = X/N$, $\mathbf{E}\|\widehat{P}^{\text{MLE}} - P\|_F^2 \approx N^{-1}$, so our result (approximately) reduces the squared error by a factor of $r/(m \wedge n)$, which is the effective rank deficiency.

For a more complete exploration of multinomial matrix denoising, we also consider a model in which our observations are independent multinomial samples from rows of a low-rank matrix. Concretely, our observations are now a matrix X whose *rows*, which we denote $\{X_i\}_{i=1}^m$, are independent and distributed according to $X_i \sim \text{Multinomial}(p_i, N_i)$, where $\{p_i\}_{i=1}^m$ are the rows of a rank-r $m \times n$ matrix P. Theorem 2 states that, under mild conditions on the sums of *columns* of P, there is an estimator $\widehat{P}$ (defined similarly to those above) such that, with high probability,

$$\|D^{1/2}(\widehat{P} - P)\|_F \lesssim \sqrt{r \log(m+n)},$$

where $D = \text{diag}(N_1, \ldots, N_m)$. It is easily checked that the maximum likelihood estimator $\widehat{P}_{\text{MLE}} = D^{-1}X$ has expected error $\mathbf{E}\|D^{1/2}(\widehat{P}^{\text{MLE}} - P)\|_F^2 \approx m$, so we again get a reduction in (squared) error that is (approximately, modulo a logarithmic factor) proportional to the reduction in degrees of freedom.

We do not analyze the multinomial estimation problems from a minimax risk standpoint, but, due to the similarities between the Poisson and multinomial distributions, we suspect that one could find similar matching lower bounds in a similar manner to the Poisson case.

2.1.4 Computation and implications for general noisy matrix completion

Note that all three of our estimators would be very easy to compute if we discarded the nonnegativity constraint: we could take a singular value decomposition of the matrix $\mathcal{A}_\Omega^*(X)$ and then either do singular value soft thresholding (for $\widehat{M}^{(1)}$ and $\widehat{M}^{(2)}$) or truncate it to the r largest singular values (for $\widehat{M}^{(3)}$).

We claim that ignoring the nonnegativity constraint does not change our analysis or the resulting error bounds at all; this constraint does not appear anywhere in the proof of Theorem 1. Therefore, if computational ability is a limiting factor, we could simply take the more efficient approach of solving without any nonnegativity constraints, and the error bounds we have presented will still apply. Projecting the result onto any convex constraint set that contains M can then only improve the performance.

We also note here that although we have chosen to focus on Poisson noise, our approach is fairly general and could apply to other types of noise. Indeed, if M were an arbitrary (not necessarily nonnegative) matrix, and we make observations of the form $M_{ij} + \xi_{ij}$ for $(i, j) \in \Omega$, and the ξ_{ij}'s are zero-mean noise variables with reasonably light tails, we could adapt our arguments to show that, for each of our three estimators,

$$\|M - \widehat{M}\|_F \lesssim \sqrt{\frac{r}{p}} \left(\max_i \sqrt{\sum_j (\mathrm{var}(\xi_{ij}) + (1-p)M_{ij}^2)} + \max_j \sqrt{\sum_i (\mathrm{var}(\xi_{ij}) + (1-p)M_{ij}^2)} \right).$$

The lower bound Theorem 4 is completely independent of the noise distribution; a version of Theorem 3 could be proved for many common distributions.

This has some interesting implications for general matrix completion with noise. Many of the existing algorithms, such as low-rank factorization [17], iterative imputation [18], or the many other algorithms, including (Equation 2.1), that can be expressed as semidefinite programs, are fairly complex. Our results suggest that, not only do simple SVD-based algo-

rithms have theoretical properties that are just as good as current state-of-the-art guarantees for more complex algorithms, but that, in a minimax error sense, *it is impossible to do any better*.

This realization does not, however, imply that there is no value to more sophisticated algorithms. As mentioned earlier, how well we can exploit incoherence in *noisy* matrix completion remains an important open question, and the matrices used in the proof of the minimax lower bound Theorem 4 are highly coherent. Therefore, it is likely that more sophisticated algorithms are still beneficial when trying to recover non-pathological (i.e., incoherent) matrices.

2.1.5 Comparison to prior work

There are several categories of existing literature to which we can compare our results. Some papers explicitly consider Poisson noise, using a maximum-likelihood framework. Cao and Xie [19] consider nuclear-norm penalized maximum likelihood for matrices contained in a nuclear norm ball (rather than exactly low-rank matrices). This approach uses an empirical process argument to bound the Kullback-Leibler divergence between the true and predicted distributions. This argument requires a Lipschitz condition on the log-likelihood function, which, for the Poisson distribution, requires imposing a lower bound on the rates. Soni *et al.* [20] and Soni and Haupt [21] consider a penalized maximum likelihood estimator from a carefully-chosen finite set of candidates (which is exponentially large in the size of the problem and hence computationally intractable). The matrices considered have a non-negative low-rank factorization (with a particular emphasis on the case when one factor is sparse). They use an information-theoretic argument to bound the expected error in terms of Bhattacharyya distance. The result of [20], which applies to matrix completion, requires imposing a lower bound on the rates, while that of [21], which considers only denoising, does not. All three papers find an upper bound on Frobenius error in terms of the statistical error metrics that they originally bound.

Other, more general approaches, are designed specifically with Frobenius-norm error in mind. One class of methods uses "restricted strong convexity" arguments, which were introduced by Negahban and Wainwright [3]. These methods rely on approximating the Frobenius norm in certain restricted classes of matrices (in which the error matrix must fall) using only samples of the entries. These methods lead to simple and elegant proofs, but the concentration inequalities on which they rely require imposing uniform upper bounds in magnitude on both the true matrix entries and the estimator entries. Other recent papers which use this type of argument include [22, 23, 24]. Another interesting paper which uses learning theory arguments to achieve a similar result is [25].

We note in Section 2.1.4 that we can get our minimax optimal results with a simple singular value truncation or thresholding. Other papers that analyze this algorithm include [4] (which inspired our approach) as well as [26, 27, 28]. The paper [17] also analyzes this as a first step in a more complicated factorization algorithm. These methods are very simple and lend themselves to simple proofs. Our error rate in (Equation 2.5) is better in that it applies to Poisson noise, has better dimension dependence (including eliminating multiplicative log factors in the error rate), and/or has a more refined dependence on matrix entries (e.g., row and column sums vs. absolute upper bounds on matrix entries).

An interesting blend of techniques can be seen in the papers [29, 30, 31], which combine some of the general approaches mentioned above with maximum likelihood estimation for exponential families of distributions. These methods, like those in [19] and [20], are difficult to apply to the Poisson distribution without imposing a lower bound on rates because, as the mean λ of the distribution goes to 0, the "natural parameter" $\log \lambda$ goes to $-\infty$, whereas the general methods used require parameters to be bounded. They also require (approximate) low rank in the matrix of natural parameters. In the Poisson case, this is equivalent to assuming a bound on the rank of the matrix $[\log M_{ij}]$ of elementwise logarithms of the means, which is somewhat non-standard, and certainly not the same as bounding the rank of the original matrix M.

There is some previous work on the denoising problem in terms of Poisson or exponential family principal component analysis (PCA). Papers in this area include [32], which recommends maximum likelihood approaches; [33], which uses variational Bayesian inference; and [34], which uses a singular value shrinkage algorithm on the means (much like we do). The recent preprint [35] examines a variety of models with random scaling factors and nonlinearities. These papers do not contain theoretical results applicable to our problem, however. [34] is related to [36], which contains an asymptotic analysis of a similar singular value shrinkage algorithm for more general problems (including matrix completion). Another work in this area is [37], which contains consistency results for low-dimensional subspace recovery.

Most of the papers mentioned above do not find error bounds which explicitly depend on the "true" rate matrix; rather, they find uniform upper bounds for classes of structured matrices with uniform upper (and, sometimes, lower) bounds on the entries. To compare our results directly to this literature, we consider what we obtain when we only impose a uniform upper and lower bounds (by, say $\lambda_{\max}$ and $\lambda_{\min}$) on the matrix entries. The approximate bound of (Equation 2.5) reduces to

$$\|\widehat{M} - M\|_F^2 \lesssim (\lambda_{\max} + (1-p)\lambda_{\max}^2)\frac{rm}{p},$$

where we have assumed, without loss of generality, that $m \geq n$. Previous results show similar error rates in terms of matrix dimensions for exactly low-rank matrices. For example, [20] establishes a bound of

$$\mathbf{E}\|\widehat{M} - M\|_F^2 \lesssim \frac{\lambda_{\max}^3}{\lambda_{\min}}\frac{rm}{p}\log m,$$

which provides a similar dependence on r, m, and p, but with an additional logarithmic term and a worse dependence on the minimum and maximum matrix values. In a slightly different setting, [19] shows that for matrices in the nuclear norm ball of radius $\lambda_{\max}\sqrt{rmn}$

(which is a convex relaxation of the exact low-rank constraint), we instead obtain (ignoring logarithmic terms and a complicated but severe dependence on $\lambda_{\max}$ and $\lambda_{\min}$) an error bound of

$$\|\widehat{M} - M\|_F^2 \lesssim \frac{\sqrt{rn}m}{\sqrt{p}},$$

where p is now the number of samples for entry in a uniform-at-random sampling model. The different dependence on r and p is interesting, but, if one compares it to results in linear regression over ℓ_1 balls (see, e.g., [38]), the rate given is perhaps not surprising. To compare to some of the more general methods mentioned above, we note that, if we consider the generalization of our method mentioned in Section 2.1.4, and we assume a uniform upper bound on the magnitudes of matrix elements and the noise variances, our results are comparable to [22] (albeit under less-strict assumptions).

As noted earlier, our work uses fairly general matrix completion methods. For the Frobenius norm error metric, this gives us an advantage over more distribution-specific approaches such as [19, 20, 29, 30], in part because we do not have to approximate the Frobenius-norm error by a statistical divergence measure or by a norm in a transformed parameter space. Our results also do not suffer from the fact that a Poisson distribution's likelihood function is ill-conditioned for very small rates. In addition, our results avoid a multiplicative logarithmic factor that appears in much of the previous literature (replacing it with an additive factor that is often negligible); this achievement (which also appears in [22]) is almost entirely due to the use of recent results in bounding the operator norm of a random matrix (such as [39]).

Finally, much of the previous literature in the Poisson case (from those mentioned above, [29, 19, 20]) finds lower bounds on minimax risk in certain classes of matrices. The paper [40] does the same for more general noise models (including a specialization to the Poisson case). Although these lower bounds have the same large-scale error rate (in terms of the rank and dimensions of the matrix and the number of samples) as the corresponding upper bounds, they differ from the upper bounds by factors that are logarithmic in the problem size

and that depend on the ratio of largest to smallest allowable rates. To our knowledge, the results in our work are the first for noisy low-rank matrix completion in which the minimax rate for large classes of matrices is found to within a universal constant.

We are aware of much less theoretical work for low-rank denoising. One recent work that is worth noting is [41]. This paper shows that, in the case of a matrix multinomial distribution, one can achieve a tight error bound in ℓ_1 distance (sum of absolute values of entries) for certain factorizable probability matrices using $O(mr^4)$ samples. It is difficult to compare this directly to our result, since it is in the much stronger ℓ_1 norm, but we note that, in a similar fashion as many of the results on Poisson observations, this paper also relies on lower bounding certain sums of entries in the factor matrices. Other theoretical work on topic modeling includes [42, 43, 44, 45, 46].

We add a final caveat to our results by noting that $\|\widehat{M} - M\|_F$ might not always be the most appropriate error metric; for example, there is a much larger difference qualitatively (and quantitatively, if we use an appropriate statistical divergence) between Poisson distributions of means 0 and 10 than between Poisson distributions of means 100 and 110. We see a similar disconnect between squared error and other probabilistic metrics in the case of the multinomial distribution. Further investigation of distribution-specific methods (such as maximum likelihood) that yield bounds in more statistically-motivated metrics is thus certainly warranted.

2.1.6 Outline

The remainder of this chapter is organized as follows. Section 2.2 contains the formal statements and proofs of our upper bounds on error. Section 2.3 contains the statements and proofs of two separate minimax lower bounds which, when combined, yield the matching lower bound to (Equation 2.5). Section 2.3.3 also discusses briefly some situations where the approximation of (Equation 2.5) is accurate, and thus the upper and lower bounds match within a universal constant.

2.2 Upper bounds

2.2.1 Poisson noise

This section is dedicated to proving our main result, which is the following theorem:

Theorem 1. *Let M be a non-negative $m \times n$ matrix with rank r. Let $\lambda_{max} = \max_{ij} M_{ij}$, and let*

$$\widetilde{\sigma}(M) = \max_i \sqrt{\sum_j (M_{ij} + (1-p)M_{ij}^2)} + \max_j \sqrt{\sum_i (M_{ij} + (1-p)M_{ij}^2)}.$$

Suppose $\Omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$ is chosen according to a Bernoulli sampling model with sampling probability p, and suppose, conditionally on Ω, $X \sim \text{Poisson}(\mathcal{A}_\Omega(M))$. Set $\epsilon \in (0, 1/2)$, and let

$$A(M, p, \epsilon) = 2\sqrt{p}\widetilde{\sigma}(M) + C \max\left\{\lambda_{\max}, 4 \log \frac{2mn}{\epsilon}\right\} \sqrt{\log \frac{m \vee n}{\epsilon}}, \tag{2.6}$$

where C is a universal constant.

Then, with probability at least $1 - 2\epsilon$, if $\delta \geq A(M, p, \epsilon)$ and $\lambda \geq 2pA(M, p, \epsilon)$, we also have

$$\|\widehat{M}^{(1)} - M\|_F \leq \frac{4\sqrt{2r}\delta}{p}$$

and

$$\|\widehat{M}^{(2)} - M\|_F \leq \frac{2\sqrt{2r}\lambda}{p^2}.$$

Moreover, we also have that with probability at least $1 - 2\epsilon$,

$$\|\widehat{M}^{(3)} - M\|_F \leq \frac{2\sqrt{2r}}{p} A(M, p, \epsilon).$$

The result follows from a series of lemmas. The first steps in upper bounding the error are the following (deterministic) results.

Lemma 1. *Suppose M is a rank-r matrix such that $\|\mathcal{A}^*_\Omega(X) - pM\| \le \delta = \frac{\lambda}{2p}$. Then, for $i \in \{1, 2\}$,*

$$\|\widehat{M}^{(i)} - M\|_F \le \frac{4\sqrt{2r}\delta}{p} = \frac{2\sqrt{2r}\lambda}{p^2}. \tag{2.7}$$

Lemma 2. *Suppose M is a rank-r matrix. Then*

$$\|\widehat{M}^{(3)} - M\|_F \le \frac{2\sqrt{2r}}{p}\|\mathcal{A}^*_\Omega(X) - pM\|.$$

Proof of Lemma 1. Let $M = U\Sigma V^*$ be the singular value decomposition of M, where $U \in \mathbf{R}^{m\times r}$ and $V \in \mathbf{R}^{n\times r}$ are such that $U^*U = V^*V = I_r$, and Σ is an $r \times r$ diagonal matrix with positive entries on the diagonal. Let T be the subspace of $\mathbf{R}^{m\times n}$ spanned by matrices of the form UA and BV^T for arbitrary matrices $A \in \mathbf{R}^{r\times n}$ and $B \in \mathbf{R}^{m\times r}$. We denote by $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$, respectively, the orthogonal projections onto T and its orthogonal complement $T^\perp$.

Denote $H^{(1)} = \widehat{M}^{(1)} - M$. Because M is feasible, and the nuclear norm is a convex function, we have

$$0 \ge \|\widehat{M}^{(1)}\|_* - \|M\|_* \ge \langle H^{(1)}, Z\rangle,$$

where $Z \in \partial\|M\|_*$ is any subgradient of the nuclear norm function at the point M. Such a subgradient must have the form

$$Z = UV^* + \mathcal{P}_{T^\perp}(W),$$

where W is an arbitrary matrix with $\|W\| \le 1$. By the duality of the nuclear norm and operator norm, we can choose W so that $\langle W, \mathcal{P}_{T^\perp}(H^{(1)})\rangle = \|\mathcal{P}_{T^\perp}(H^{(1)})\|_*$. We then have

$$\begin{aligned} 0 &\ge \langle H^{(1)}, UV^* + \mathcal{P}_{T^\perp}(W)\rangle \\ &= \langle \mathcal{P}_T(H^{(1)}), UV^*\rangle + \|\mathcal{P}_{T^\perp}(H^{(1)})\|_* \\ &\ge -\|\mathcal{P}_T(H^{(1)})\|_* + \|\mathcal{P}_{T^\perp}(H^{(1)})\|_*, \end{aligned}$$

where the last inequality follows from the fact that $\|UV^*\| = 1$. We therefore have

$$\|\mathcal{P}_{T^\perp}(H^{(1)})\|_* \leq \|\mathcal{P}_T(H^{(1)})\|_*.$$

Hence

$$\begin{aligned}\|H^{(1)}\|_* &\leq 2\|\mathcal{P}_T(H^{(1)})\|_* \\ &\leq 2\sqrt{2r}\|\mathcal{P}_T(H^{(1)})\|_F \\ &\leq 2\sqrt{2r}\|(H^{(1)})\|_F,\end{aligned}$$

where the second inequality follows from the fact that any element of T has rank at most $2r$.

By the triangle inequality and the homogeneity of the operator norm, we also have

$$\|H^{(1)}\| \leq \frac{2\delta}{p}.$$

Thus

$$\begin{aligned}\|H^{(1)}\|_F^2 &= \langle H^{(1)}, H^{(1)}\rangle \\ &\leq \|H^{(1)}\|\|H^{(1)}\|_* \\ &\leq \frac{4\sqrt{2r}\delta}{p}\|H^{(1)}\|_F,\end{aligned}$$

and the first part of the result immediately follows.

The proof of the second part is similar; letting $H^{(2)} = \widehat{M}^{(2)} - M$, we now have, by the

optimality of $\widehat{M}$,

$$\begin{aligned}
0 &\geq \|\mathcal{A}_\Omega^*(X) - p\widehat{M}^{(2)}\|_F^2 + \lambda\|\widehat{M}^{(2)}\|_* - \left(\|\mathcal{A}_\Omega^*(X) - pM\|_F^2 + \lambda\|M\|_*\right) \\
&= p^2\|\widehat{M}^{(2)}\|_F^2 - p^2\|M\|_F^2 + 2p\langle\mathcal{A}_\Omega^*(X), M - \widehat{M}^{(2)}\rangle + \lambda\left(\|\widehat{M}^{(2)}\|_* - \|M\|_*\right) \\
&= p^2\|\widehat{M}^{(2)}\|_F^2 + p^2\|M\|_F^2 - 2p^2\langle\widehat{M}^{(2)}, M\rangle + 2p^2\langle M, \widehat{M}^{(2)} - M\rangle \\
&\quad - 2p\langle\mathcal{A}_\Omega^*(X), \widehat{M}^{(2)} - M\rangle + \lambda\left(\|\widehat{M}^{(2)}\|_* - \|M\|_*\right) \\
&= p^2\|H^{(2)}\|_F^2 - 2p\langle\mathcal{A}_\Omega^*(X) - pM, H^{(2)}\rangle + \lambda\left(\|\widehat{M}^{(2)}\|_* - \|M\|_*\right).
\end{aligned}$$

Noting, as before, that $\|\widehat{M}^{(2)}\|_* - \|M\|_* \geq \|\mathcal{P}_{T^\perp}(H^{(2)})\|_* - \|\mathcal{P}_T(H^{(2)})\|_*$, and that

$$|\langle\mathcal{A}_\Omega^*(X) - pM, H^{(2)}\rangle| \leq \|\mathcal{A}_\Omega^*(X) - pM\|\|H^{(2)}\|_* \leq \frac{\lambda}{2p}\|H^{(2)}\|_*,$$

we have

$$\begin{aligned}
\|H^{(2)}\|_F^2 &\leq \frac{\lambda}{p^2}(\|H^{(2)}\|_* + \|\mathcal{P}_T(H^{(2)})\|_* - \|\mathcal{P}_{T^\perp}(H^{(2)})\|_*) \\
&\leq \frac{2\lambda}{p^2}\|\mathcal{P}_T(H^{(2)})\|_* \\
&\leq \frac{2\sqrt{2r}\lambda}{p^2}\|H^{(2)}\|_F.
\end{aligned}$$

□

Proof of Lemma 2. Because both M and $\widehat{M}^{(3)}$ have rank at most r, $H^{(3)} = \widehat{M}^{(3)} - M$ has rank at most $2r$. By a similar calculation to that in the proof of Lemma 1, the optimality of $\widehat{M}^{(3)}$ implies

$$\begin{aligned}
0 &\geq \|\mathcal{A}_\Omega^*(X) - p\widehat{M}^{(3)}\|_F^2 - \|\mathcal{A}_\Omega^*(X) - pM\|_F^2 \\
&= p^2\|H^{(3)}\|_F^2 - 2p\langle\mathcal{A}_\Omega^*(X) - pM, H^{(3)}\rangle,
\end{aligned}$$

so

$$\begin{aligned}\|H^{(3)}\|_F^2 &\leq \frac{2}{p}|\langle \mathcal{A}_\Omega^*(X) - pM, H^{(3)}\rangle| \\ &\leq \frac{2}{p}\|\mathcal{A}_\Omega^*(X) - pM\|\|H^{(3)}\|_* \\ &\leq \frac{2\sqrt{2r}}{p}\|\mathcal{A}_\Omega^*(X) - pM\|\|H^{(3)}\|_F.\end{aligned}$$

□

The remainder of the work is to show that $\|\mathcal{A}_\Omega^*(X) - pM\| \leq A(M, p, \epsilon)$ with probability at least $1 - 2\epsilon$. We will use the following fundamental lemma, which was originally proved by Bandeira and van Handel [39] and appears with a slightly improved constant in [47].

Lemma 3 (Theorem 4.9 and Remark 4.11 in [47])**.** *Let X be a random $m \times n$ matrix whose entries are independent, centered, and almost surely bounded in absolute value by a constant b. Let*

$$\sigma = \max_i \sqrt{\sum_j \mathbf{E}\, X_{ij}^2} + \max_j \sqrt{\sum_i \mathbf{E}\, X_{ij}^2}.$$

Then

$$\mathrm{P}(\|X\| \geq 2\sigma + t) \leq (m \vee n) \exp\left(-\frac{t^2}{C_0 b^2}\right),$$

where C_0 is a universal constant.

Poisson random variables are clearly unbounded, so Lemma 3 does not directly apply. The following technical lemma allows us to extend the result to the case of random variables with sub-exponential tails.

Lemma 4. *Let X be a random $m \times n$ matrix whose entries are independent and centered, and suppose that for some $v, t_0 > 0$, we have, for all $t \geq t_0$,*

$$\mathrm{P}(|X_{ij}| \geq t) \leq 2e^{-t/v}.$$

Let $\epsilon \in (0, 1/2)$, and let

$$K = \max\left\{t_0, v \log \frac{2mn}{\epsilon}\right\}.$$

Then

$$\mathrm{P}\left(\|X\| \geq 2\sigma + \frac{\epsilon v}{\sqrt{mn}} + t\right) \leq (m \vee n) \exp\left(-\frac{t^2}{C_0(2K)^2}\right) + \epsilon,$$

where σ and C_0 are the same as in Lemma 3.

Proof. First, note that, by a union bound,

$$\mathrm{P}\left(\max_{i,j}|X_{ij}| > K\right) \leq 2mne^{-K/v} \leq \epsilon.$$

Consider the truncation $X^K = [X_{ij}^K]$, where $X_{ij}^K = X_{ij}\mathbf{1}_{\{|X_{ij}|\leq K\}}$. Note that

$$\begin{aligned}
|\mathbf{E}\, X_{ij}^K| &\leq \mathbf{E}|X_{ij}^K - X_{ij}| \\
&= \mathbf{E}|X_{ij}|\mathbf{1}_{\{|X_{ij}|>K\}} \\
&= \int_K^\infty \mathrm{P}(|X_{ij}| > t)\, dt \\
&\leq \int_K^\infty 2e^{-t/v}\, dt \\
&= 2ve^{-K/v} \\
&\leq \frac{\epsilon v}{mn} \\
&\leq K.
\end{aligned}$$

Let $\widetilde{X}^K = X^K - \mathbf{E}\, X^K$ be the centered version of X^K. Clearly, $\mathbf{E}(\widetilde{X}_{ij}^K)^2 \leq \mathbf{E}\, X_{ij}^2$, and $|\widetilde{X}_{ij}^K| \leq 2K$. Then, by Lemma 3,

$$\mathrm{P}(\|\widetilde{X}^K\| \geq 2\sigma + t) \leq (m \vee n)e^{-t^2/C_0(2K)^2}.$$

Furthermore, with probability at least $1 - \epsilon$,

$$\begin{aligned} \|X\| &= \|X^K\| \\ &\leq \|\widetilde{X}^K\| + \|\mathbf{E}\, X^K\| \\ &\leq \|\widetilde{X}^K\| + \|\mathbf{E}\, X^K\|_F \\ &\leq \|\widetilde{X}^K\| + \frac{\epsilon v}{\sqrt{mn}}, \end{aligned}$$

and the result follows. □

To apply this result, we need a subexponential tail bound for the Poisson distribution.

Lemma 5. *Let* $X \sim \mathrm{Poisson}(\lambda)$. *Then*

$$\mathrm{P}(X - \lambda \geq t) \leq \exp\left(-\frac{t^2}{2(\lambda + t/3)}\right).$$

For $t \geq \lambda$,

$$\mathrm{P}(X - \lambda \geq t) \leq e^{-3t/8}.$$

The first inequality can be established by approximating the Poisson distribution with mean λ as the sum of k Bernoulli random variables with mean λ/k, applying Bernstein's inequality, and taking $k \to \infty$. The idea for this argument was suggested by an exercise in [48].

Going back to our original problem, we need to bound the operator norm of $Z = \mathcal{A}_\Omega^*(X) - pM$. Note that since we are using a Bernoulli sampling model, the entries of Z are independent. Let $\lambda_{\text{max}} = \max_{i,j} M_{ij}$. Note that for every (i, j), $\mathbf{E}\, Z_{ij} = 0$,

$$Z_{ij} \geq -pM_{ij} \geq -\lambda_{\text{max}},$$

and, for $t \geq 2\lambda_{\max}$,

$$\mathrm{P}(Z_{ij} \geq t) \leq e^{-3(t-\lambda_{\max})/8} \leq e^{-3t/16} \leq e^{-t/8}.$$

Then, by Lemma 4, we have, for $\epsilon \in (0, 1/2)$,

$$\mathrm{P}\left(\|Z\| \geq 2\sigma + \frac{8\epsilon}{\sqrt{mn}} + t\right) \leq (m \vee n) \exp\left(-\frac{t^2}{C_0(2K)^2}\right) + \epsilon,$$

where

$$K = \max\left\{2\lambda_{\max}, 8\log\frac{2mn}{\epsilon}\right\},$$

and σ is defined as before. To calculate σ in terms of p and M, we note that

$$\begin{aligned}\mathrm{var}(Z_{ij}) &= p\,\mathrm{var}(X_{ij}) + p(1-p)(\mathbf{E}\,X_{ij})^2 \\ &= pM_{ij} + p(1-p)M_{ij}^2.\end{aligned}$$

Therefore, we can calculate

$$\sigma = \max_i \sqrt{\sum_j (pM_{ij} + p(1-p)M_{ij}^2)} + \max_j \sqrt{\sum_i (pM_{ij} + p(1-p)M_{ij}^2)} = \sqrt{p}\widetilde{\sigma}.$$

Taking $t = C_1 K\sqrt{\log\frac{m\vee n}{\epsilon}}$, we have, with probability at least $1 - 2\epsilon$,

$$\|Z\| \leq 2\sqrt{p}\tilde{\sigma} + \frac{8\epsilon}{\sqrt{mn}} + C_1 K\sqrt{\log\frac{m\vee n}{\epsilon}}.$$

We take this as our $A(M, p, \epsilon)$, subsuming the $\frac{8\epsilon}{\sqrt{mn}}$ term into the last term.

2.2.2 Corollary on multinomial estimation

Here we prove a corollary of Theorem 1, showing how it implies a result for estimating the low-rank-matrix parameter of a matrix multinomial distribution. For the sake of brevity, we

only consider the analogue of $\widehat{M}^{(1)}$.

Corollary 1. *Let P be a nonnegative $m \times n$ matrix with rank r such that $\sum_{i,j} P_{ij} = 1$. Suppose, furthermore, that we have*

$$\max_i \sum_j P_{ij} \leq \frac{a}{m}, \max_j \sum_i P_{ij} \leq \frac{b}{n}$$

for some constants $a, b \geq 1$, and that $\max_{i,j} P_{ij} \leq c$. Let N be a positive integer, and suppose that $X \sim \text{Multinomial}(P, N)$. Let $\epsilon \in (0, 1)$, choose $\delta > 0$ such that

$$\delta \geq \frac{1}{N}\left(2\sqrt{N\left(\frac{a}{m}+\frac{b}{n}\right)} + \frac{4\epsilon}{e\sqrt{mnN}} + C\max\left\{Nc, 4\log\frac{4emn\sqrt{N}}{\epsilon}\right\}\sqrt{\log\frac{2e\sqrt{N}(m\vee n)}{\epsilon}}\right),$$

and let

$$\widehat{P}^{\delta}(X) = \underset{\substack{P' \in [0,1]^{m\times n} \\ \sum_{i,j} P_{ij}=1}}{\arg\min} \|P'\|_* \text{ s.t. } \|X - NP'\| \leq N\delta.$$

Then, with probability at least $1 - \epsilon$,

$$\|\widehat{P}^{\delta} - P\|_F \leq 4\sqrt{2r}\delta.$$

As in the Poisson case, there are many situations where the additive logarithmic term in the definition of δ is negligible, and we have

$$\|\widehat{P}^{\delta} - P\|_F \lesssim \sqrt{\frac{r}{N}\left(\frac{a}{m}+\frac{b}{n}\right)}.$$

Proof of Corollary 1. Suppose $Y \sim \text{Poisson}(NP)$. Let $\epsilon' = \epsilon/2e\sqrt{N}$. Define the event $A = \{\|\widehat{P}^{\delta}(Y) - P\|_F \leq 4\sqrt{2r}\delta\}$. Theorem 1 (with $p = 1$) implies that $\text{P}(A) \geq 1 - 2\epsilon' = 1 - \epsilon/e\sqrt{N}$. Note that the additional constraints in the optimization problem do not affect

this fact, since P, by definition, meets these constraints.

X has the same distribution as Y conditioned on $B = \{\sum_{i,j} Y_{ij} = N\}$, so it suffices to show that the probability of A conditioned on this event is at least $1 - \epsilon$. Indeed, note that $\sum_{i,j} Y_{ij} \sim \text{Poisson}(N)$, so

$$\mathrm{P}(B) = \frac{e^{-N} N^N}{N!} \geq \frac{1}{e\sqrt{N}}$$

by Stirling's approximation. Then, by Bayes' rule,

$$\begin{aligned}
\mathrm{P}(\|\widehat{P}^{\delta}(X) - P\|_F \geq 4\sqrt{2r}\delta) &= \mathrm{P}(A^c \mid B) \\
&= \frac{\mathrm{P}(A^c \cap B)}{\mathrm{P}(B)} \\
&\leq \frac{\mathrm{P}(A^c)}{\mathrm{P}(B)} \\
&\leq 2\epsilon' e\sqrt{N} \\
&= \epsilon
\end{aligned}$$

□

2.2.3 Multinomial denoising with independent rows

We now consider a slightly different setting for multinomial estimation. Here, we consider a model in which we observe a collection of independent multinomial random variables, each of which has distribution parametrized by a row in a low-rank matrix.

Theorem 2. *Let $X_1, \ldots, X_m$ be independent multinomial random vectors, with*

$$X_i \sim \text{Multinomial}(p_i, N_i),$$

where, for each i, $N_i \geq 1$ is an integer, and $p_i = (p_{i1}, \ldots, p_{in})$ is a vector of probabilities.

Denote the matrices

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix} \in \mathbf{R}^{m \times n},\ P = \begin{bmatrix} p_1 \\ \vdots \\ p_m, \end{bmatrix} \in \mathbf{R}^{m \times n},\ D = \begin{bmatrix} N_1 & & \\ & \ddots & \\ & & N_m \end{bmatrix} \in \mathbf{R}^{m \times m},$$

where each X_i and p_i are considered row vectors. Define the estimator

$$\widehat{P}^{\delta} = \underset{P' \in [0,1]^{m \times n}}{\arg\min} \|D^{1/2} P'\|_* \text{ s.t. } \|D^{-1/2}(X - DP')\| \leq \delta,\ \sum_j P_{ij} = 1\ \forall i \in \{1, \ldots, m\}.$$

Let $D_{\min} = \min_i D_i$, and choose δ such that

$$\delta \geq \max\left\{ 2\sqrt{\max\left\{1, \max_j \sum_i p_{ij}\right\} \log \frac{m+n}{\epsilon}}, \frac{4}{3\sqrt{D_{\min}}} \log \frac{m+n}{\epsilon} \right\}.$$

Then, with probability at least $1 - \epsilon$,

$$\|D^{1/2}(\widehat{P}^{\delta} - P)\|_F \leq 4\sqrt{2r}\delta,$$

where r is the rank of P.

Our approach uses the following matrix Bernstein inequality.

Lemma 6 ([49, Theorem 6.1.1])**.** *Suppose $Z = \sum_k S_k \in \mathbf{R}^{m \times n}$, where $\{S_k\}_k$ is a finite sequence of independent, zero-mean random matrices. Suppose that for some constant $L > 0$, for each k, $\|S_k\| \leq L$ almost surely. Let*

$$v = \max\{\|\mathbf{E}\, ZZ^T\|, \|\mathbf{E}\, Z^T Z\|\} = \max\left\{ \left\| \sum_k \mathbf{E}\, S_k S_k^T \right\|, \left\| \sum_k \mathbf{E}\, S_k^T S_k \right\| \right\}.$$

Then, for $t \geq 0$,

$$\mathrm{P}(\|Z\| \geq t) \leq (m+n) \exp\left(\frac{-t^2}{2(v + Lt/3)} \right).$$

Proof of Theorem 2. On the event that $\|D^{-1/2}(X - DP)\| \leq \delta$, we have $\|D^{1/2}(\widehat{P}^{\delta} - P)\| \leq 2\delta$, and the result follows by the same steps as in the proof of Theorem 1.

To find a bound on the operator norm of $Z = D^{-1/2}(X - DP)$, we apply Lemma 6, noting that Z is the sum of independent, zero-mean matrices:

$$Z = \sum_{i=1}^{m} \sum_{k=1}^{D_i} \frac{1}{\sqrt{D_i}} e_i (u_{ik} - p_i)^T,$$

where e_i is the i^{th} standard basis element in $\mathbf{R}^m$, u_{ik} is a random vector in $\mathbf{R}^n$ that is equal to e_j with probability p_{ij} (and all of the u_{ik}'s are independent), and $p_i = (p_{i1}, \ldots, p_{in})$ is the i^{th} row of P.

One can verify that

$$\mathbf{E}\, ZZ^T = \operatorname{diag}\left(\sum_{j=1}^{n} p_{1j}(1 - p_{1j}), \ldots, \sum_{j=1}^{n} p_{mj}(1 - p_{mj}) \right) \preceq I_m,$$

where I_m denotes the $m \times m$ identity matrix, and that

$$\mathbf{E}\, Z^T Z = \sum_{i=1}^{m} \operatorname{diag}(p_i) - p_i p_i^T \preceq \operatorname{diag}\left(\sum_{i=1}^{m} p_{i1}(1 - p_{i1}), \ldots, \sum_{i=1}^{m} p_{in}(1 - p_{in}) \right).$$

We can then take $v \leq \max\{1, \max_j \sum_i p_{ij}\}$. Clearly, each term in the sum has operator norm bounded above by $L = 1/\sqrt{D_{\min}}$.

Some algebraic manipulation of the result of Lemma 6 implies that, for $\epsilon \in (0, 1)$, we have, with probability at least $1 - \epsilon$,

$$\begin{aligned} \|Z\| &\leq \max\left\{ 2\sqrt{v \log \frac{m+n}{\epsilon}}, \frac{4L}{3} \log \frac{m+n}{\epsilon} \right\} \\ &\leq \max\left\{ 2\sqrt{\max\left\{1, \max_j \sum_i p_{ij}\right\} \log \frac{m+n}{\epsilon}}, \frac{4}{3\sqrt{D_{\min}}} \log \frac{m+n}{\epsilon} \right\}, \end{aligned}$$

which establishes the result.

□

2.3 Minimax lower bounds

In this section, we show that the rate in (Equation 2.5) is optimal (within a multiplicative constant) in the sense of minimax risk. We do this in two parts; the two minimax lower bounds derived in the next two sections, when combined, match the rate in (Equation 2.5) is optimal.

2.3.1 First lower bound

In the Poisson upper error bound, $\widetilde{\sigma}$ is partially determined by the maximal row and column sums of the rate matrix M, which we can think of as the maximal variance of any row or column (without sampling a subset of the entries). Our first lower bound shows that we cannot improve on this term:

Theorem 3. *Let r, k, and ℓ, be positive integers, and take $m = rk$, $n = r\ell$. Let $\lambda_{\max} \geq 1/8\ell p$, set $\sigma_1^2 = k\lambda_{\max}$, and let*

$$S = \left\{ M \in [0, \lambda_{\max}]^{m \times n} : \operatorname{rank}(M) \leq r, \sqrt{\max_i \sum_j M_{ij}} + \sqrt{\max_j \sum_i M_{ij}} \leq 2\sigma_1 \right\}.$$

Then, under a Bernoulli sampling model with sampling probability p,

$$\inf_{\widehat{M}} \sup_{M \in S_1} \mathrm{P}_M \left(\|\widehat{M} - M\|_F \geq \frac{\sqrt{r}\sigma_1}{8\sqrt{2p}} \right) \geq \frac{1}{2} - \frac{8 \log 2}{m \vee n}.$$

Proof. Assume, without loss of generality, that $k \geq \ell$. We use a variant of Fano's method. We first find a large hypercube of matrices, and then we use the fact that we can find a large subset that is well-separated.

For $i \in \{1, \ldots, m\}$, let

$$A_i = \{\ell(q-1)+1, \ldots \ell q\},$$

where $q \in \{1, \ldots, r\}$ is the unique integer such that $i \in \{k(q-1)+1, \ldots, kq\}$. For λ_0, λ_1 to be chosen later, we define, for $\theta \in \{0,1\}^m$, the block-diagonal matrix $M_\theta \in S$ by

$$(M_\theta)_{ij} = \begin{cases} \lambda_{\theta_i}, & j \in A_i, \\ 0, & \text{otherwise.} \end{cases}$$

The nonzero elements of the i^{th} row of M_θ are all either λ_0 or λ_1, depending on the value of θ_i.

By a combinatorial argument (see, e.g., [50]), one can show that there exists $\Theta \subset \{0,1\}^m$ such that $\mathrm{card}(\Theta) \geq e^{m/8}$ and, for all distinct $\theta, \theta' \in \Theta$, the Hamming distance $d_H(\theta, \theta') \geq m/4$.

Take $\lambda_0 = \lambda_{\max}/2 - \delta$ and $\lambda_1 = \lambda_{\max}/2 + \delta$, where $\delta \leq \lambda_{\max}/2$ is a constant to be chosen later. Note that for all distinct $\theta, \theta' \in \Theta$, we have

$$\|M_\theta - M_{\theta'}\|_F \geq \sqrt{m\ell}\delta.$$

We denote by P_θ the distribution of $\mathcal{P}_\Omega(X)$ when $X \sim \mathrm{Poisson}(M_\theta)$, and Ω is independently chosen from the Bernoulli sampling model with parameter p. From Fano's inequality from information theory, one can derive (see, e.g., [50]) a lower bound on the probability of an

estimator's error exceeding half the distance between points indexed by Θ:

$$\begin{aligned}\inf_{\widehat{M}} \sup_{M \in S} \mathrm{P}_M \left(\|\widehat{M} - M\|_F \geq \frac{\sqrt{m\ell}\delta}{2} \right) &\geq p_e \\ &\coloneqq \inf_{\phi} \sup_{\theta \in \Theta} \mathrm{P}(\phi(\mathcal{P}_\Omega(X)) \neq \theta) \\ &\geq 1 - \frac{\sup_{\theta \in \Theta} \mathrm{D}_{\mathrm{KL}}(P_\theta \parallel Q) + \log 2}{\log \operatorname{card}(\Theta)},\end{aligned}$$

where ϕ denotes a test taking values in Θ, and Q is any probability distribution on $\mathbf{R}^{m \times n}$. We take Q to be the distribution generated in the same manner as each P_θ, simply with λ_1 and λ_2 replaced by $\lambda_{\max}/2$.

Note that the Kullback-Leibler divergence between two Poisson distributions with rates λ and λ' is

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\lambda \parallel \lambda') &= \lambda' - \lambda + \lambda \log \frac{\lambda}{\lambda'} \\ &\leq \lambda' - \lambda + \lambda \left(\frac{\lambda}{\lambda'} - 1 \right) \\ &= \frac{(\lambda - \lambda')^2}{\lambda'}.\end{aligned}$$

Therefore, for any $\theta \in \Theta$,

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(P_\theta \parallel Q) &\leq rk\ell p \frac{\delta^2}{\lambda_{\max}/2} \\ &= m\ell p \frac{2\delta^2}{\lambda_{\max}}.\end{aligned}$$

Take

$$\delta = \sqrt{\frac{\lambda_{\max}}{32\ell p}} \leq \frac{\lambda_{\max}}{2}.$$

Then

$$p_e \geq \frac{1}{2} - \frac{8 \log 2}{m},$$

and half the separation between points indexed by Θ is at least

$$\frac{\sqrt{m\ell}\delta}{2} = \frac{1}{8\sqrt{2}}\sqrt{\frac{m\lambda_{\max}}{p}} = \frac{1}{8\sqrt{2}}\sqrt{\frac{r}{p}}\sigma_1.$$

□

2.3.2 Second lower bound

The previous theorem relies on the fact that the observations are conditionally Poisson. The next result, which provides the second part of a matching lower bound to (Equation 2.5), does not depend on the conditional distribution of the observations, and instead shows a fundamental limit in inferring missing matrix entries.

Theorem 4. *Take again $m = rk$, $n = r\ell$. Set $\sigma_2^2 = k\lambda_{\max}^2$. Let*

$$S = \left\{ M \in [0, \lambda_{\max}]^{m\times n} : \operatorname{rank}(M) \le r, \sqrt{\max_i \sum_j M_{ij}^2} + \sqrt{\max_j \sum_i M_{ij}^2} \le 2\sigma_2 \right\}.$$

Suppose $p \ge \frac{1}{2(k\wedge\ell)} = \frac{r}{2(m\wedge n)}$. Then, under a Bernoulli sampling model with probability p (with any conditional distribution on the observations),

$$\begin{aligned}\inf_{\widehat{M}} \sup_{M\in S_2} \mathbf{E}\|\widehat{M} - M\|_F^2 &\ge \frac{r\sigma_2^2}{8}\max\left\{\frac{1}{2}\left\lfloor\frac{1}{2p}\right\rfloor, 1-p\right\} \\ &\ge \frac{1}{64}\frac{1-p}{p}r\sigma_2^2.\end{aligned}$$

Proof. Again, we assume that $k \ge \ell$. We first prove the lower bound with the first item in the maximum. For notational simplicity, we can assume that $1/2p$ is an integer. Furthermore, we can assume that $\ell = 1/2p$, since decreasing the number of columns does not increase risk.

We consider a set of matrices $\{M_\theta\}_{\theta\in\{0,1\}^m}$ with the same structure as in the proof of Theorem 3, but now, we take $\lambda_0 = 0$, $\lambda_1 = \lambda_{\max}$.

Assouad's lemma (see, e.g., [51] or [50]) gives a lower bound on the Bayes risk of an estimator $\widehat{M}$ for a uniform prior on $\{0,1\}^m$:

$$\begin{aligned} R(\widehat{M}) &:= \frac{1}{2^m} \sum_{\theta \in \{0,1\}^m} \mathbf{E}_\theta \|\widehat{M} - M_\theta\|_F^2 \\ &\geq \frac{1}{2} \sum_{i=1}^{m} \frac{\ell \lambda_{\max}^2}{4} \inf_\phi \left(\mathrm{P}_\theta(\phi \neq \theta_i) + \mathrm{P}_{\theta^i}(\phi \neq 1 - \theta_i) \right), \end{aligned}$$

where $\phi \colon \mathbf{R}^{m \times n} \to \{0,1\}$ is a test, and, for $\theta \in \{0,1\}^m$, θ^i denotes the element of $\{0,1\}^m$ that is equal to θ except in the i^{th} position.

Denote by P_θ^i the marginal distribution of the i^{th} row of a matrix with distribution P_θ. The minimal testing risk which appears in the sum is equal to the L_1 norm of the minimum of the densities of P_θ^i and $P_{\theta^i}^i$, which measures how much the distributions overlap. Thus we have

$$\begin{aligned} \inf_\phi \left(\mathrm{P}_\theta(\phi \neq \theta_i) + \mathrm{P}_{\theta^i}(\phi \neq 1 - \theta_i) \right) &= \|P_\theta \wedge P_{\theta^i}\|_1 \\ &\geq (1-p)^\ell \\ &\geq 1 - \ell p \\ &= \frac{1}{2}, \end{aligned}$$

where the first inequality is due to the fact that, with probability $(1-p)^\ell$, no entry from the i^{th} row of M is observed.

Then

$$R(\widehat{M}) \geq \frac{m \ell \lambda_{\max}^2}{16} = \frac{r \sigma_2^2}{32 p}.$$

The result follows from the fact that minimax risk always exceeds Bayes risk.

A simple modification with $\ell = 1$ yields the result for the second term in the maximum.

□

We note here that we could also use a similar argument as in the proof of Theorem 3

to get a high-probability lower bound on error (with a somewhat worse constant). For example, setting $Q = \frac{P_{(0,\ldots,0)}+P_{(1,\ldots,1)}}{2}$, it is easily verified that the resulting Kullback-Leibler divergences $\mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, Q)$ can be upper bounded by a sum of coordinate-wise total variation distance.

2.3.3 When do the upper and lower bounds match?

Within multiplicative constants, the lower bounds of Theorems 3 and 4 match the approximate upper bound of (Equation 2.5). We must therefore consider when the approximation in (Equation 2.5) is accurate.

Finding technical conditions that guarantee matching rates is not something we think likely to be very instructive at this point, especially since a different proof technique could potentially change the logarithmic term in $A(M, p, \epsilon))$. However, we think it is helpful to look at the matrices involved in the proofs of Theorems 3 and 4. Note that when the bounds match for those particular matrices, the minimax error rate bounds are tight for the matrix classes considered in those proofs.

For these matrices (assuming that $m \geq n$),

$$\widetilde{\sigma} \approx \sqrt{\frac{m}{r}}(\sqrt{\lambda_{\max}} + \sqrt{1-p}\lambda_{\max}).$$

For this term to dominate (Equation 2.6), we must have

$$\widetilde{\sigma} \gtrsim \frac{\lambda_{\max} \vee c \log m}{\sqrt{p}} \sqrt{\log m}.$$

For example, if $\lambda_{\max} \geq \log m$, it would suffice to take

$$p \gtrsim \frac{r \log m}{m},$$

which is a standard condition in noiseless matrix completion. If $\lambda_{\max} \leq \log m$, it would

suffice to take

$$p \gtrsim \frac{r \log^3 m}{m \lambda_{\max}^2}.$$

2.4 Conclusion and future work

In this chapter, we have derived an upper bound in Frobenius norm error for an estimator for Poisson matrix completion, and we have derived a minimax lower bound that matches this upper bound (within a universal constant) for many classes of nonnegative rate matrices. We have also derived similar upper bounds for error in two types of multinomial matrix denoising problems. The estimators we use are computationally tractable, and require significantly fewer assumptions on the underlying matrix than previous results in the literature. Significantly, we impose no lower bounds on the entries of the underlying matrix. This is crucial in many applications (such as topic modelling) where zero or very small means can be relatively common.

Because we have found upper and lower error bounds in Frobenius norm, the only theoretical improvement remaining for this model and error metric in general classes of matrices is to try to relax the conditions under which the bounds match (although, as we have seen, they are not too restrictive now). This could potentially come about by reducing the logarithmic term in (Equation 2.6) and/or by finding a logarithmic term to add to the minimax lower bounds. One could also further examine how to obtain better error rates for more restrictive classes of matrices, such as incoherent matrices.

It would also be interesting to extend the results presented here to matrices that are not exactly low-rank, but are instead "approximately low-rank"; for example, we could consider matrices which are contained in Schatten balls (which, for $q \in [0, 1]$, are sets of matrices for which $\sum_i \sigma_i^q \leq R$, where $\{\sigma_i\}$ is the set of singular values). As mentioned previously, Cao and Xie [19] used the Schatten 1-norm ($q = 1$, or nuclear norm ball); Negahban and Wainwright [3] also examined these classes of matrices.

Another avenue of research would be to examine structured Poisson or multinomial

estimation under different, more statistically motivated error metrics. Maximum likelihood methods seem more suitable here than least-squares, but analysis of maximum likelihood estimators has proved difficult for the reasons outlined in Section 2.1.5. It is not clear what kind of structure would be relevant in a different error metric. Low-rank structure seems to work well with a least-squares error framework, but there is *a priori* not much reason to think that it would work similarly well for another metric; for example, the Bhattacharyya distance between Poisson distributions, is proportional to the (squared) ℓ_2 distance between the *square roots* of the rates, but the element-wise square root of a low-rank matrix is not, in general, low rank. Thus, this approach may not immediately bear much fruit. However, an analysis of matrix estimation under alternative error metrics remains an important area for future research.

CHAPTER 3

OPTIMAL CONVEX LIFTED SPARSE PHASE RETRIEVAL AND PCA WITH AN ATOMIC MATRIX NORM REGULARIZER

In this chapter,[1] we present novel analysis and algorithms for solving sparse phase retrieval and sparse principal component analysis (PCA) with convex lifted matrix formulations. The key innovation is a new mixed atomic matrix norm that, when used as regularization, promotes low-rank matrices with sparse factors. We show that convex programs with this atomic norm as a regularizer provide near-optimal sample complexity and error rate guarantees for sparse phase retrieval and sparse PCA. While we do not know how to solve the convex programs exactly with an efficient algorithm, for the phase retrieval case we carefully analyze the program and its dual and thereby derive a practical heuristic algorithm. We show empirically that this practical algorithm performs similarly to existing state-of-the-art algorithms.

3.1 Introduction

3.1.1 Sparsity, phase retrieval, and PCA

Consider the standard linear regression problem in which we make observations of the form $y_i = \langle x_i, \beta^* \rangle + \xi_i$, $i = 1, \ldots, n$, where $x_1, \ldots, x_n$ are measurement vectors and $\xi_1, \ldots, \xi_n$ represent noise or other error. If the x_i's are chosen randomly and independently (e.g., i.i.d. Gaussian), and the noise is zero-mean and independent with $\mathrm{var}(\xi_i) \leq \sigma^2$, it is well-known that in general, we need[2] $n \gtrsim p$ measurements to estimate β^* meaningfully, and the best possible error we can obtain is $\|\hat{\beta} - \beta^*\|_2 \lesssim \sigma\sqrt{p/n}$.

We can potentially do much better if we exploit *sparsity* in the vector β^*. If β^* has

[1] This work is available as a preprint in [52],

[2] Here and throughout the chapter, $\lesssim$ and $\gtrsim$ denote, respectively, $\leq$ and $\geq$ within absolute constants.

(at most) s nonzero entries, the standard LASSO algorithm, which requires solving an ℓ_1-regularized least-squares optimization problem, yields an estimator $\hat{\beta}$ satisfying $\|\hat{\beta} - \beta^*\|_2 \lesssim \sigma\sqrt{(s/n)\log(p/s)}$ as long as the number of measurements satisfies $n \gtrsim s\log(p/s)$ (see, e.g., [53, Chapter 10]). Thus by using a convex regularized optimization problem we can exploit sparsity to reduce the number of measurements n and the estimation error proportionally to sparsity level (i.e., the number of nonzero entries in β^*). In this work, we seek to extend this phenomenon to two problems: *phase retrieval* and *principal component analysis* (PCA). To introduce our main results, we briefly describe phase retrieval and PCA and their sparse variants. We focus on the formulations most relevant to our results. More complete background and related literature can be found in Sections 3.1.2 and 3.1.3.

In phase retrieval, we seek to estimate a vector β^* from n noisy *quadratic* observations of the form $y_i = |\langle x_i, \beta^* \rangle|^2 + \xi_i$. The nonlinearity in the measurement model makes estimation and analysis more complicated than if our measurements are linear. To get around this, a common approach is to note that for any $x, \beta \in \mathbf{R}^p$, $|\langle x, \beta \rangle|^2 = \langle X, B \rangle_{\mathrm{HS}}$, where $X = x \otimes x$ and $B = \beta \otimes \beta$ are rank-1 positive semidefinite (PSD) matrices, and $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt (Frobenius) matrix inner product. We can then write our observations as the *linear* measurements $y_i = \langle X_i, B^* \rangle_{\mathrm{HS}} + \xi_i$, where $B^* = \beta^* \otimes \beta^*$ and $X_i = x_i \otimes x_i$. This is often called a "lifted" formulation, since we are mapping the parameter of interest from $\mathbf{R}^p$ to the larger space of $p \times p$ PSD matrices. If the x_i's are randomly chosen (say, Gaussian), and we solve the semidefinite program

$$\widehat{B} = \operatorname*{arg\,min}_{B \succeq 0} \ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle_{\mathrm{HS}})^2,$$

we can bound $\|\widehat{B} - B^*\|_{\mathrm{HS}} \lesssim \sigma\sqrt{p/n}$ as long as $n \gtrsim p$, where σ is the standard deviation of the ξ_i's. (As shown in [54], this implies that the leading eigenvector of $\widehat{B}$ is close to β^* up to its sign.) Both the sample complexity and the error rate are comparable to those in ordinary linear regression.

In PCA, we observe n i.i.d. random vectors $\{x_i\}_{i=1}^n$, and we want to estimate the leading eigenvector v_1 of the covariance matrix $\Sigma = \mathbf{E}(x_1 \otimes x_1)$. Again, this can be solved in a lifted manner with a semidefinite program, noting that

$$P_1 \coloneqq v_1 \otimes v_1 = \arg\max_{P \in \mathbf{R}^{p \times p}} \langle \Sigma, P \rangle_{\mathrm{HS}} \text{ s.t. } \|P\|_* \leq 1.$$

An estimator $\widehat{P}$ of P_1 is obtained[3] by replacing Σ with the empirical covariance $\widehat{\Sigma}$. Again, if $n \gtrsim p$, we can recover P_1 within error proportional to $\sqrt{p/n}$ (where the constants depend on the gap between the first and second leading eigenvalues of Σ).

Sparse phase retrieval seeks to combine phase retrieval with sparse recovery. If β^* is s-sparse, and we observe $y_i = |\langle x_i, \beta^* \rangle|^2 + \xi_i$ for $i \in \{1, \ldots, n\}$, can we recover β^* with a similar sample complexity and error as in linear sparse recovery? Similarly, the question we consider in *sparse PCA* is whether, if the leading eigenvector v_1 is s-sparse, we can recover it with a similar sample complexity and error as in linear recovery.

Our main contributions are the following:

- We present novel convex relaxations of the sparse phase retrieval and sparse PCA problems that use both a lifted formulation and a sparsity-inducing regularization, and we prove that for both problems, an estimator computed via a convex program achieves $O(s \log(p/s))$ sample complexity and $O(\sigma\sqrt{(s/n)\log(p/s)})$ error rate as in linear sparse recovery.

- Although we do not know how to compute the convex programs exactly (we suspect they may, in fact, be computationally intractable), we present a heuristic motivated by a careful analysis of the dual problem and the problem's optimality conditions, and we show that in the case of sparse phase retrieval, the resulting algorithm achieves nearly identical empirical performance to existing state-of-the-art sparse phase retrieval algorithms.

[3]It would be computationally suboptimal in practice to compute the leading eigenvector of $\widehat{\Sigma}$ with a semidefinite program, but this formulation helps motivate our approach to the sparse case.

In the following sections, we describe the sparse phase retrieval and sparse PCA problems in more detail, and we review the related literature.

3.1.2 Sparse phase retrieval

Phase retrieval in p dimensions with (sub-)Gaussian measurements is by now well-studied. If we have n observations of the form $y_i \approx |\langle x_i, \beta^* \rangle|^2$, we can solve the optimization problem

$$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{\arg\min} \sum_{i=1}^{n} (y_i - |\langle x_i, \beta^* \rangle|^2)^2. \tag{3.1}$$

Unfortunately, this is a nonconvex problem, so there is no immediately obvious way to solve it efficiently. (A similar optimization problem and similar nonconvexity appear if we instead write our measurements without the square, i.e., our observations are $\approx |\langle x_i, \beta^* \rangle|$.)

Most approaches to this algorithmic difficulty fall into one of two categories. One method is to optimize a nonconvex loss function such as (Equation 3.1) directly (and iteratively) with a suitable initialization (e.g., [55]). The other is the lifted semidefinite approach outlined in Section 3.1.1. For example, Candès and Li [56] show that if the design vectors x_i are Gaussian, $y_i = |\langle x_i, \beta^* \rangle|^2 + \xi_i$, and we have $n \gtrsim p$ measurements, solving

$$\widehat{B} = \underset{B \succeq 0}{\arg\min} \sum_{i=1}^{n} |y_i - \langle X_i, B \rangle_{\mathrm{HS}}|$$

achieves $\|\widehat{B} - B^*\|_F \lesssim \frac{1}{n} \sum_{i=1}^{n} |\xi_i|$ with high probability. In the case of zero-mean random noise with standard deviation σ, we can, by using a squared loss, improve this to $\|\widehat{B} - B^*\|_F \lesssim \sigma \sqrt{p/n}$ (see [57]). Thus we can solve the phase retrieval problem with a sample complexity and susceptibility to noise proportional to the dimension p; this is the same complexity as ordinary linear regression.

Several results have been published on how to adapt iterative nonconvex phase retrieval algorithms to the sparse setting [58, 59, 60, 61, 62, 63]. Some [59, 62] do indeed achieve $O(\sigma \sqrt{(s/n) \log p})$ error bounds with zero-mean noise—this is very close to the optimal

rate in linear sparse recovery (the rest do not analyze theoretically the noisy case). However, the theory in this literature requires $n \gtrsim s^2 \log p$, which, unless s is very small, is much larger than what is required in linear sparse recovery. As Soltanolkotabi [64] points out, the key difficulty is finding a good initialization for the algorithms—once we are close enough to β^*, we only need[4] $n \gtrsim_{\log} s$ measurements to converge to a correct estimate. In practice, the first initialization step is often to estimate the support of β^*; the best known methods require $n \gtrsim_{\log} s^2$ measurements. We compare several of these algorithms (in addition to that of the purely algorithmic/empirical work [65]) to ours empirically in Section 3.4.3, and we see that all of them appear empirically to have *linear* sample complexity in s.

More related to our results are methods to adapt the lifted convex phase retrieval approach to the sparse setting. The foundational theoretical work in this area is by Li and Voroninski [66], although the method was previously studied empirically in [67]. The key idea is that if $\beta^* \in \mathbf{R}^p$ is s-sparse, the lifted version $B^* = \beta^* \otimes \beta^*$ is both rank-1 and at most s^2-sparse. In the noiseless case, they solve the optimization problem

$$\widehat{B} = \underset{B \succeq 0}{\arg\min}\ \lambda_1 \operatorname{tr}(B) + \lambda_2 \|B\|_{1,1} \text{ s.t. } \langle X_i, B \rangle_{\mathrm{HS}} = y_i,\ i = 1, \ldots, n, \tag{3.2}$$

where $\|\cdot\|_{1,1}$ denotes the elementwise ℓ_1 norm of a matrix. The trace regularization term promotes low rank, while the ℓ_1 norm promotes sparsity. As with the nonconvex methods, their theory requires $n \gtrsim s^2 \log p$ measurements to get exact recovery. The result of [57], when specialized to sparse phase retrieval, extends this approach to the noisy case, getting, within log factors, the same $O(s^2)$ sample and noise complexity.

3.1.3 Sparse PCA

PCA is a well-established technique with which, given points $x_1, \ldots, x_n \in \mathbf{R}^p$, we try to find a low-dimensional linear (or affine) subspace that contains most of the energy in the data. If $x_1, \ldots, x_n$ have zero empirical mean (e.g., after centering), the closest r-dimensional

[4] Here and hereafter, $\gtrsim_{\log}$ ($\lesssim_{\log}$) will denote "greater (less) than within a logarithmic factor."

subspace to the points (in mean square ℓ_2 distance) is the space spanned by the top r eigenvectors of the empirical covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i \otimes x_i$.

For simplicity, take $r = 1$. Suppose the x_i's are i.i.d. copies of a random variable x with true covariance Σ with eigenvalue decomposition $\Sigma = \sum_\ell \sigma_\ell v_\ell \otimes v_\ell$, where $\sigma_1 > \sigma_2 \geq \cdots \geq \sigma_p$. If x is Gaussian, and $\sigma_2 \gtrsim \frac{\sigma_1}{p-1}$, then, with high probability [68],

$$\|\widehat{\Sigma} - \Sigma\|_2 \lesssim \sqrt{\sigma_1 \frac{\sigma_1 + (p-1)\sigma_2}{n}} \lesssim \sqrt{\sigma_1 \sigma_2 \frac{p}{n}}.$$

Then, if $\hat{v}_1$ is the leading eigenvector of $\widehat{\Sigma}$, the Davis-Kahan $\sin\Theta$ theorem gives

$$\|\hat{v}_1 \otimes \hat{v}_1 - v_1 \otimes v_1\|_2 \lesssim \frac{\sqrt{\sigma_1\sigma_2}}{\sigma_1 - \sigma_2}\sqrt{\frac{p}{n}}.$$

This rate is minimax-optimal over general covariance matrices with the given σ_1, σ_2 (see [69]).

When p is large compared to n, we need to impose more structure on Σ to recover the leading eigenvector(s) accurately. In sparse PCA, we consider the case in which the eigenvector(s) of interest are *sparse*. This problem has been extensively studied in the past decade: see [70] for a recent review.

In the single-eigenvector recovery case ($r = 1$), Cai *et al.* [71] show that if the leading eigenvector v_1 is s-sparse, the minimax rate for all estimators $\hat{v}_1$ of v_1 over the simple class $\{\Sigma = \sigma_2 I_p + (\sigma_1 - \sigma_2) v_1 \otimes v_1 \colon v_1 \text{ } s\text{-sparse}, \|v_1\|_2 = 1\}$ is

$$\|\hat{v}_1 \otimes \hat{v}_1 - v_1 \otimes v_1\|_2 \approx \frac{\sqrt{\sigma_1\sigma_2}}{\sigma_1 - \sigma_2}\sqrt{\frac{s\log(p/s)}{n}}.$$

While this theoretical result is clean and achieves our desire to bring sparse-recovery sample complexity and error to the PCA problem, one practical problem remains: how do we *compute* an estimator $\hat{v}_1$ that achieves these theoretical properties? As with sparse phase retrieval, the best theoretical results for computationally efficient algorithms require $n \gtrsim_{\log}$

s^2 to guarantee accurate recovery (see, e.g., [71, 72]). Once again, proper initialization (often by estimating the support of v_1) is the key difficulty.

There is strong evidence to suggest that this s^2 barrier may be intrinsic for computationally efficient algorithms. Recent results suggest that any statistically optimal estimator that requires fewer measurements must be NP-hard to compute. Berthet and Rigollet [73] showed that if a certain testing problem in random graph theory (the *planted clique problem*) is NP-hard to compute in certain regimes (which is widely believed although so-far unproved in standard computational models), then accurately *testing for the existence of* a sparse leading eigenvector when $n \lesssim_{\log} s^2$ is NP-hard. Wang *et al.* [74] and Gao *et al.* [75] further refine this by showing that, under a similar assumption, there is no efficiently computable consistent estimator of v_1 when $n \lesssim_{\log} s^2$.

3.2 Key tool: A sparsity-and-low-rank–inducing atomic norm

To motivate our approach, consider the optimization problem (Equation 3.2) from [66] for sparse phase retrieval or its least-squares version

$$\widehat{B} = \underset{B \succeq 0}{\arg\min} \ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle_{\mathrm{HS}})^2 + \lambda_1 \operatorname{tr}(B) + \lambda_2 \|B\|_{1,1}. \tag{3.3}$$

It turns out that quadratic (in sparsity) $O(s^2)$ complexity is a fundamental performance bound for this class of methods. Our target matrix B^* has two kinds of structure: it is rank-1 and s^2-sparse. The trace regularization in our estimator encourages low rank, while the ℓ_1 regularization encourages sparsity. However, recent work [76, 77] has shown it is impossible to take advantage of both kinds of structure simultaneously with a regularizer that is merely a convex combination of the two structure-inducing regularizers; the best we can do is exploit either the low rank as in non-sparse phase retrieval, in which case we get $O(p)$ complexity, or the s^2-sparsity, in which case we get $O(s^2)$ complexity.

To see intuitively why we have this problem, note that the nuclear norm and elementwise

ℓ_1 norm are both examples of *projective tensor norms* [78]. For matrix A of any size,

$$\begin{aligned}\|A\|_* &= \inf\Big\{\sum \|u_i\|_2\|v_i\|_2 : A = \sum u_i \otimes v_i\Big\} \\ &= \inf\Big\{\sum \frac{\|u_i\|_2^2 + \|v_i\|_2^2}{2} : A = \sum u_i \otimes v_i\Big\},\end{aligned}$$

and

$$\begin{aligned}\|A\|_{1,1} &= \inf\Big\{\sum \|u_i\|_1\|v_i\|_1 : A = \sum u_i \otimes v_i\Big\} \\ &= \inf\Big\{\sum \frac{\|u_i\|_1^2 + \|v_i\|_1^2}{2} : A = \sum u_i \otimes v_i\Big\}.\end{aligned}$$

Equivalently, these norms are atomic norms [79] where the atoms are rank-1 matrices with unit ℓ_2 or ℓ_1 norms. For a PSD matrix, the trace is the nuclear norm, so the regularizer in (Equation 3.3) can be expressed as

$$\begin{aligned}\lambda_1 \operatorname{tr}(B) + \lambda_2\|B\|_{1,1} = \lambda_1 \inf\Big\{\sum \frac{\|u_i\|_2^2 + \|v_i\|_2^2}{2} : B = \sum u_i \otimes v_i\Big\} \\ + \lambda_2 \inf\Big\{\sum \frac{\|w_i\|_1^2 + \|z_i\|_1^2}{2} : B = \sum w_i \otimes z_i\Big\}.\end{aligned} \tag{3.4}$$

A key feature of $B^* = \beta^* \otimes \beta^*$ is that the factors of its rank-1 decomposition have a certain ℓ_2 norm *and* are sparse. Because the two infima in (Equation 3.4) are separate, the regularizer promotes matrices with two *separate* atomic decompositions of low ℓ_2 and ℓ_1 norm respectively. It does not encourage a decomposition into low-rank matrices with factors that have *simultaneously* low ℓ_2 norm and low ℓ_1 norm.

Inspired by the framework of Haeffele and Vidal [80], we propose the following regularizer:

$$\|B\|_{*,s} \coloneqq \inf\Big\{\sum \theta_s(u_i, v_i) : B = \sum u_i \otimes v_i\Big\}, \tag{3.5}$$

where

$$\theta_s(u, v) = \frac{1}{2}\left(\|u\|_2^2 + \|v\|_2^2 + \frac{\|u\|_1^2 + \|v\|_1^2}{s}\right),$$

and $s > 0$ is a parameter that represents the sparsity (or an approximation thereof) of the vector we are interested in recovering. Similar notions of atomic norms that promote simultaneous low rank and sparsity have appeared in [81, 77].

We will show in the next section that using $\|\cdot\|_{*,s}$ as a regularizer in lifted formulations of sparse phase retrieval and PCA gives sample complexity and error bounds nearly identical to the linear regression case.

3.3 Theoretical guarantees for atomic-norm regularized estimators

In this section, we state precisely our main problems, assumptions, abstract convex optimization algorithm, and theoretical guarantees.

3.3.1 Sparse phase retrieval

Suppose $\beta^* \in \mathbf{R}^p$ is an s-sparse vector. Let x be a random vector in $\mathbf{R}^p$. We observe n i.i.d. copies $(x_1, y_1), \ldots, (x_n, y_n)$ of the random couple (x, y), where y is a real random variable whose distribution conditioned on x depends only on $\langle x, \beta^* \rangle^2$ (i.e., $y \sim p_y(y \mid \langle x, \beta^* \rangle^2)$). Let $\xi \coloneqq y - \langle x, \beta^* \rangle^2$ denote the "noise." We make the following assumptions:

Assumption 1 (Sub-Gaussian measurements)**.** The entries $(x^{(1)}, \ldots, x^{(p)})$ of x are i.i.d. real random variables with $\mathbf{E}\, x^{(i)} = 0$, $\mathbf{E}(x^{(i)})^2 = 1$, $\mathbf{E}(x^{(i)})^4 > 1$, and sub-Gaussian norm $\|x^{(i)}\|_{\psi_2} \leq K$ for some $K > 0$.

Note that the fourth-moment assumption excludes Rademacher random variables. In what follows, for simplicity of presentation, all dependence on K and the difference $\mathbf{E}(x^{(i)})^4 - 1$ will be subsumed into unspecified constants.

Assumption 2 (Zero-mean, bounded-moment noise)**.** $\mathbf{E}[\xi \mid x] = 0$ almost surely, and, for all $u \in \mathbf{R}^p$ such that $\|u\|_2 \leq 1$,

$$\mathbf{E}\, \xi^2 \langle x, u \rangle^4 \leq \sigma^2(\beta^*),$$

where $\sigma^2(\beta^*)$ is a quantity that possibly depends on the vector β^*, the distribution of x, and the conditional distribution of y. Furthermore, there are $M, \eta \geq 0$ such that

$$\|\xi\langle x, u\rangle^2\|_\alpha \leq M\alpha^{\eta+1}$$

for $\alpha \geq 3$ and all $u \in \mathbf{R}^p$ such that $\|u\|_2 \leq 1$ (where $\|Z\|_\alpha := (\mathbf{E}|Z|^\alpha)^{1/\alpha}$ for any random variable Z).

Our two working examples are the following:

- Independent additive noise: ξ is independent of all other quantities, in which case we can take $\sigma^2(\beta) \approx \mathrm{var}(\xi)$, and M and η depend on the moments of ξ.
- Poisson noise: $y \sim \mathrm{Poisson}(\langle x, \beta^*\rangle^2)$ conditioned on x. In this case, under Assumption 1, we can take $\sigma^2(\beta^*) \approx \|\beta^*\|_2^2$, $M \approx \|\beta^*\|_2 + 1$, and $\eta = 1$ (we prove this in Section A.5).

As before, we lift the problem into the space of PSD matrices by setting $B^* = \beta^* \otimes \beta^*$ and $X = x \otimes x$. We then choose a regularization parameter $\lambda \geq 0$ and compute our estimate by the following optimization problem:

$$\widehat{B} = \underset{B \in \mathbf{R}^{p \times p}}{\arg\min} \ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle X_i, B\rangle_{\mathrm{HS}})^2 + \lambda \|B\|_{*,s}. \tag{3.6}$$

We then have the following guarantee for sample complexity and error, proved in Section A.3:

Theorem 5. *Suppose Assumptions 1 and 2 hold. Suppose β^* is s-sparse and that the number of measurements n satisfies $n \gtrsim s\log(ep/s)$. If the regularization parameter satisfies*

$$\lambda \gtrsim \sqrt{\frac{s\log(ep/s)}{n}\sigma^2(\beta^*)} + \frac{M}{n^{1-c}}\Big(s\log\frac{ep}{s}\Big)^{\eta+1},$$

where $c \approx (s\log(ep/s))^{-1}$*, then, with probability at least* $1 - e^{-n} - e^{-s}(s/p)^s$*, the estimator* $\widehat{B}$ *from* (Equation 3.6) *satisfies*

$$\|\widehat{B} - B^*\|_* \lesssim \lambda.$$

Remark 1. For simplicity of presentation, we assume that the sparsity level s used in the regularizer is in fact (an upper bound on) the sparsity of β^*. We could easily extend our results to the "misspecified" case $\|\beta^*\|_0 = s_0 > s$.

Remark 2. By a standard argument (found, e.g., in [54]), if $\hat{\beta} \otimes \hat{\beta}$ is the closest rank-1 approximation to $\widehat{B}$, then $\hat{\beta}$ satisfies

$$\min\{\|\hat{\beta} - \beta^*\|_2, \|\hat{\beta} + \beta^*\|_2\} \lesssim \frac{\lambda}{\|\beta^*\|_2}.$$

Remark 3. The required sample complexity $s\log(ep/s)$ is precisely the optimal sample complexity from traditional linear sparse recovery. For large n, the noise error rate (with appropriately chosen λ) is also the optimal $\sqrt{(s/n)\log(ep/s)}$, but, if $\eta > 0$, this only holds when n is significantly larger than the required $s\log(ep/s)$. In our proof, this is due to the fact that we need concentration inequalities for sums depending on $\langle x, u\rangle^2$ for arbitrary vectors u; these terms have higher moments than the $\langle x, u\rangle$ terms we would typically see in linear settings.

If $\log n \gtrsim s\log(ep/s)$, we could have the result hold with probability $\geq 1 - 2e^{-n}$ if we replaced $\sqrt{(s/n)\log(ep/s)}$ by $(\log n)/n$ in the lower bound on λ. The factor of n^c in the second term in the lower bound on λ is negligible except in the case of very badly-behaved noise and very large n.

Remark 4. In the independent additive noise case, a simple modification of our proof would show that the result holds uniformly over β^*. If $\mathrm{var}(\xi) = \sigma^2$, we get, for appropriately chosen λ,

$$\|\widehat{B} - B^*\|_* \lesssim \sqrt{\frac{s\log(ep/s)}{n}}\sigma + \frac{M}{n^{1-c}}\Big(s\log\frac{ep}{s}\Big)^{\eta+1}.$$

Remark 5. In the Poisson observation case, we obtain, for appropriately chosen λ,

$$\|\widehat{B} - B^*\|_* \lesssim \sqrt{\frac{s\log(ep/s)}{n}}\|\beta^*\|_2 + \frac{\|\beta^*\|_2 + 1}{n^{1-c}}\Big(s\log\frac{ep}{s}\Big)^2.$$

When $\beta^* \neq 0$, and n is large enough that the first error term dominates, we have, up to a sign, that

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \sqrt{\frac{s\log(ep/s)}{n}},$$

where $\hat{\beta}$ is the appropriately-scaled leading eigenvector of $\widehat{B}$. Thus we get an error bound does that not depend on $\|\beta^*\|_2$.

3.3.2 Sparse PCA

We can apply the atomic regularizer to the sparse PCA problem via another standard lifted formulation:

Theorem 6. *Suppose we observe n i.i.d. copies of the p-dimensional vector $x \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma = \sigma_1 v_1 \otimes v_1 + \Sigma_2$, v_1 is s-sparse and unit-norm, $\sigma_1 > \|\Sigma_2\| =: \sigma_2$, and $\Sigma_2 v_1 = 0$. Choose*

$$\lambda \gtrsim \sqrt{\sigma_1\sigma_2}\sqrt{\frac{s\log(ep/s)}{n}}$$

and let

$$\widehat{P} = \operatorname*{arg\,min}_{P\in\mathbf{R}^{p\times p}} \; -\langle\widehat{\Sigma}, P\rangle_{\mathrm{HS}} + \lambda\|P\|_{*,s} \text{ s.t. } \|P\|_* \leq 1, \tag{3.7}$$

where

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\otimes(x_i - \bar{x}) = \left(\frac{1}{n}\sum_{i=1}^n x_i\otimes x_i\right) - \bar{x}\otimes\bar{x}$$

is the empirical covariance of $x_1, \ldots, x_n$ ($\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$).

For $t > 0$, if $n \gtrsim \max\left\{s\log\frac{ep}{s}, \left(\frac{\sigma_1}{\sigma_1-\sigma_2}\right)^2 t\right\}$, then, with probability at least $1 - e^{-t} - e^{-s}(s/p)^s$,

$$\|\widehat{P} - P_1\|_F \lesssim \frac{\lambda}{\sigma_1 - \sigma_2},$$

where $P_1 = v_1 \otimes v_1$.

We prove this result in Section A.4.

Remark 6. The assumption that x is Gaussian could easily be relaxed to $x = \Sigma^{1/2} z$, where z is a sub-Gaussian random vector, as in, for example, [69].

Remark 7. For properly chosen λ and large enough n, the resulting error rate

$$\|\widehat{P} - P_1\|_F \lesssim \frac{\sqrt{\sigma_1 \sigma_2}}{\sigma_1 - \sigma_2} \sqrt{\frac{s \log(ep/s)}{n}}$$

matches the minimax lower bounds in [69, 71].

3.3.3 PSD constraints and another regularizer

For phase retrieval and PCA, it is natural to restrict our estimators to be PSD. All of our theoretical results hold if we add a $B \succeq 0$ constraint to (Equation 3.6) or a $P \succeq 0$ constraint to (Equation 3.7).

Unlike the nuclear norm case (where the optimal decomposition is the singular value decoposition, which is identical to the eigenvalue decomposition for a PSD matrix), it is not clear whether every PSD matrix B admits a symmetric (i.e., $u_i = v_i$) optimal decomposition with regard the definition of $\|B\|_{*,s}$ in (Equation 3.5). Therefore, it is natural to define as a new regularizer the following gauge function/asymmetric norm on the space of PSD matrices: for $B \succeq 0$,

$$\Theta_s(B) = \inf\Big\{ \sum \theta_s(u_i, u_i) : B = \sum u_i \otimes u_i \Big\}.$$

All of our theoretical and computational results in Sections 3.3 and 3.4 can be easily extended to this choice of regularizer. This choice of regularizer is computationally convenient because if we optimize over a matrix B by optimizing over factors u_i, v_i such that $B = \sum_i u_i \otimes v_i$ as in Section 3.4.2, we can enforce a PSD constraint simply by forcing $u_i = v_i$.

3.4 Computational limitations and a practical algorithm for phase retrieval

Although the mixed atomic norm $\|\cdot\|_{*,s}$ is a powerful theoretical tool, it is not clear how to calculate (let alone optimize) it for a general matrix in practice, since it is defined as an infimum over infinite sets of possible factorizations.

A warning that computations with these atomic regularizers may be difficult in general is that they can be used to get $O_{\log}(s)$ sample complexity for sparse PCA, which, as discussed in Section 3.1.3, is widely believed to be impossible with efficient algorithms.

In this section, we will analyze the convex programs more carefully, with a particular focus on phase retrieval.[5] We will analyze the optimality conditions via a dual problem and thereby develop a heuristic algorithm.

This problem was studied in greater generality in [80]. Their Corollary 1 is similar to our Corollary 2. However, our analysis of the dual problem is quite different from their perturbation argument, and we can much more easily apply our method to the sparse PCA optimization problem (Equation 3.7) with its inequality constraint. Furthermore, we think the reader will benefit from our deriving the optimality conditions from more elementary principles for the particular problem we are trying to solve.

3.4.1 Factorization, duality, and optimality conditions

To move toward a practical algorithm, we consider optimizing (Equation 3.6) in factored form; rather than optimizing over B directly, we optimize over the factors $\{u_k, v_k\}$ of a factorization $B = \sum_k u_k \otimes v_k$. Then (Equation 3.6) is equivalent to

$$\min_{\{u_k, v_k\} \subset \mathbf{R}^p} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \left\langle X_i, \sum_k u_k \otimes v_k \right\rangle_{\mathrm{HS}} \right)^2 + \lambda \sum_k \theta_s(u_k, v_k). \tag{3.8}$$

[5] While our algorithmic approach led to strong empirical performance for sparse phase retrieval, the approach was less effective for sparse PCA. We leave a more thorough investigation of this phenomenon for future work.

The obvious drawback to this form is that the optimization problem is no longer convex; therefore, it is not clear whether finding a global minimum is computationally feasible.

To determine how well a factored algorithm works (e.g., to certify optimality), we examine a dual problem to (Equation 3.6). We formulate the dual via a trick found in [82]: note that $b^2/2 = \max_a \ ab - a^2/2$ (achieved if and only if $a = b$), and therefore

$$\begin{aligned}
&\min_{B \in \mathbf{R}^{p \times p}} \ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle_{\mathrm{HS}})^2 + \lambda \|B\|_{*,s} \\
&\quad = \min_{B \in \mathbf{R}^{p \times p}} \ \frac{1}{2n} \sum_{i=1}^{n} \max_{\alpha_i} \ \left(2\alpha_i (y_i - \langle X_i, B \rangle_{\mathrm{HS}}) - \alpha_i^2 \right) + \lambda \|B\|_{*,s} \\
&\quad \geq \max_{\alpha \in \mathbf{R}^n} \ \left[\frac{1}{n} \sum_{i=1}^{n} \left(\alpha_i y_i - \frac{\alpha_i^2}{2} \right) + \min_{B \in \mathbf{R}^{p \times p}} \ \left(\lambda \|B\|_{*,s} - \frac{1}{n} \sum_{i=1}^{n} \alpha_i \langle X_i, B \rangle_{\mathrm{HS}} \right) \right],
\end{aligned}$$

where the inequality comes from swapping the maximum over $\alpha = (\alpha_1, \ldots, \alpha_n)$ and the minimum over B.

Define the dual norm $\|\cdot\|_{*,s}^*$ by

$$\|Z\|_{*,s}^* = \max_{\substack{B \in \mathbf{R}^{p \times p} \\ \|B\|_{*,s} \leq 1}} \ \langle Z, B \rangle_{\mathrm{HS}}.$$

Because $\|\cdot\|_{*,s}^*$ is nonnegatively homogeneous,

$$\min_{B \in \mathbf{R}^{p \times p}} \ \left(\lambda \|B\|_{*,s} - \left\langle \frac{1}{n} \sum_{i=1}^{n} \alpha_i X_i, B \right\rangle_{\mathrm{HS}} \right) = \begin{cases} 0 & \text{if } \left\| \frac{1}{n} \sum_{i=1}^{n} \alpha_i X_i \right\|_{*,s}^* \leq \lambda \\ -\infty & \text{otherwise.} \end{cases}$$

Therefore, a dual formulation of (Equation 3.6) is

$$\max_{\alpha \in \mathbf{R}^n} \ \left(\frac{1}{n} \sum_{i=1}^{n} \alpha_i y_i - \frac{\alpha_i^2}{2} \right) \text{ s.t. } \left\| \frac{1}{n} \sum_{i=1}^{n} \alpha_i X_i \right\|_{*,s}^* \leq \lambda. \tag{3.9}$$

This is a convex problem, since the dual norm is convex (as a maximum of linear functions).

Before we go further, note that,

$$\|Z\|_{*,s}^* = \max_{\substack{u,v\in\mathbf{R}^p \\ \theta_s(u,v)\le 1}} \langle Zu, v\rangle.$$

To see this, note that

$$\begin{aligned}
\|Z\|_{*,s}^* &= \sup\left\{ \langle Z, B\rangle_{\mathrm{HS}} : B = \sum_k u_k \otimes v_k, \{u_k, v_k\} \subset \mathbf{R}^p, \sum_k \theta_s(u_k, v_k) \le 1 \right\} \\
&= \sup\left\{ \sum_k \langle Zu_k, v_k\rangle : \{u_k, v_k\} \subset \mathbf{R}^p, \sum_k \theta_s(u_k, v_k) \le 1 \right\} \\
&= \sup\left\{ \sum_{k=1}^K \langle Zu_k, v_k\rangle : K \ge 1, \{u_k, v_k\}_{k=1}^K \subset \mathbf{R}^p, \sum_{k=1}^K \theta_s(u_k, v_k) \le 1 \right\}.
\end{aligned}$$

For any finite sequence $\{u_k, v_k\}_{k=1}^K$ with $\sum_{k=1}^K \theta_s(u_k, v_k) \le 1$, set $k^* = \arg\max_k \frac{\langle Zu_k, v_k\rangle}{\theta_s(u_k, v_k)}$, $\tilde{u} = \frac{u_{k^*}}{\sqrt{\theta_s(u_{k^*}, v_{k^*})}}$, and $\tilde{v} = \frac{v_{k^*}}{\sqrt{\theta_s(u_{k^*}, v_{k^*})}}$; then $\theta_s(\tilde{u}, \tilde{v}) = 1$, and $\langle Z\tilde{u}, \tilde{v}\rangle \ge \sum_{k=1}^K \langle Zu_k, u_k\rangle$. Therefore,

$$\|Z\|_{*,s}^* = \sup\{\langle Zu, v\rangle : \theta_s(u, v) \le 1\}.$$

We can replace the supremum by a maximum because the objective function is continuous and the constraint set is compact.

Returning to the optimization problem, note that a feasible point α for the dual problem gives us a *lower* bound on the primal optimal value. If there exist $B \in \mathbf{R}^{p\times p}$, $\alpha \in \mathbf{R}^n$ such that α is feasible and the two objective functions are *equal*, then we know B is optimal for the primal problem. More precisely, (B, α) is an optimal primal-dual pair if and only if

(a) the primal objective function at B equals the dual objective functions at α, and

(b) α is feasible, i.e., $\left\| \frac{1}{n} \sum_{i=1}^n \alpha_i X_i \right\|_{*,s}^* \le \lambda$.

From the derivation of the dual problem above, (a) requires $\alpha_i = y_i - \langle X_i, B\rangle_{\mathrm{HS}}$. Making this substitution, setting the objective functions equal, and simplifying gives one direction of the following result:

Lemma 7. *B solves* (Equation 3.6) *if and only if both of the following hold:*

(a) $\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle X_i, B\rangle_{\mathrm{HS}})\langle X_i, B\rangle_{\mathrm{HS}} = \lambda\|B\|_{*,s}$.

(b) $\left\|\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle X_i, B\rangle_{\mathrm{HS}})X_i\right\|_{*,s}^{*} \leq \lambda$.

Proof. We have already shown that these conditions are *sufficient* for optimality. To see the other direction (that these conditions are *necessary* for optimality), note that $Z \coloneqq \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle X_i, B\rangle_{\mathrm{HS}})X_i$ is the negative gradient of the empirical loss at B. Because condition (b) is equivalent to

$$\langle Zu, v\rangle \leq \lambda\theta_s(u, v)\ \forall u \in \mathbf{R}^p,$$

if (b) does not hold, there exists some $\bar{u}, \bar{v} \in \mathbf{R}^p$ such that $\langle Z\bar{u}, \bar{v}\rangle > \lambda\theta_s(\bar{u}, \bar{v})$, and then we can decrease the objective function by moving to $B + \epsilon\bar{u} \otimes \bar{v}$ for some sufficiently small $\epsilon > 0$. Thus (b) is a necessary condition for the optimality of B.

Now suppose (b) holds, but (a) does not. Condition (b) implies that $\langle Z, B\rangle_{\mathrm{HS}} \leq \lambda\|B\|_{*,s}$, so we must have $\langle Z, B\rangle_{\mathrm{HS}} < \lambda\|B\|_{*,s}$.

Let $B = \sum_k u_k \otimes v_k$ be an optimal factorization with respect to the definition of $\|B\|_{*,s}$, that is, such that $\|B\|_{*,s} = \sum_k \theta_s(u_k, v_k)$ (we assume, for clarity, that an optimal factorization exists—if not, we could use an approximation argument). There must be some u_k, v_k such that $\langle Zu_k, v_k\rangle < \lambda\theta_s(u_k, v_k)$. Then, modifying B by replacing (u_k, v_k) with $((1-\epsilon)u_k, (1-\epsilon)v_k)$ for some sufficiently small $\epsilon > 0$ will decrease the objective function. □

Note that the proof of Lemma 7 gives us an explicit way to improve the objective function whenever one of the optimality conditions is not satisfied.

Applying our derivation to the factored optimization problem, we get the following result:

Corollary 2. *B solves* (Equation 3.6) *and $B = \sum_k u_k \otimes v_k$ is an optimal factorization with respect to $\|\cdot\|_{*,s}$ (equivalently, $\{u_k, v_k\}$ solve* (Equation 3.8)*) if and only if the following hold:*

(a) For all k, $\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})\langle X_i u_k, v_k\rangle = \lambda\theta_s(u_k, v_k)$.

(b) $\left\|\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})X_i\right\|_{*,s}^* \leq \lambda$; equivalently, for all $u, v \in \mathbf{R}^p$,

$$\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})\langle X_i u, v\rangle \leq \lambda\theta_s(u, v).$$

Note that we have broken out condition (a) into individual equalities (rather than equating the sums of each side); condition (b) allows us to do this. It is even easier to find a descent direction when one of these conditions fails to hold, since the objective function of (Equation 3.8) already depends explicitly on the vectors u_k, v_k.

Note that condition (a) is much easier to verify than condition (b). We refer to $\{u_k, v_k\}$ as a *first-order stationary point* if it satisfies condition (a), since this is equivalent to a zero subgradient on the (nonzero) u_k's and v_k's (cf. Proposition 2 in [80]).

Although we are not focusing on sparse PCA here, it may be interesting to compare Corollary 2 to what we get for sparse PCA, particularly as PCA may be a fundamentally more difficult problem. A dual problem to (Equation 3.7) is

$$\arg\max_{Z \in \mathbf{R}^{p\times p}} \; -\|Z\| \text{ s.t } \|\widehat{\Sigma} - Z\|_{*,s}^* \leq \lambda.$$

The following lemma gives (redundant) optimality conditions:

Lemma 8. *P solves* (Equation 3.7) *if and only if $\|P\|_* = 1$ and there exists $Z \in \mathbf{R}^{p\times p}$ such that*

1. $\|\widehat{\Sigma} - Z\|_{*,s}^* \leq \lambda$,
2. $\langle \widehat{\Sigma} - Z, P\rangle_{\mathrm{HS}} = \lambda\|P\|_{*,s}$,

3. $\langle Z, P \rangle_{\mathrm{HS}} = \|Z\| = \|Z\|\|P\|_*$, *and*

4. $\|Z\| = \langle \widehat{\Sigma}, P \rangle_{\mathrm{HS}} - \lambda \|P\|_{*,s}$.

In the PCA case, the semidefinite version of the problem is somewhat simpler due to the fact that the nuclear norm becomes a trace. If we solve

$$\widehat{P} = \underset{P \succeq 0}{\arg\min} \ -\langle \widehat{\Sigma}, P \rangle_{\mathrm{HS}} + \lambda \Theta_s(P) \text{ s.t. } \mathrm{tr}(P) \leq 1,$$

we get similar theoretical error guarantees as Theorem 6. Furthermore, $P = \sum_k u_k \otimes u_k$ solves this optimization program and $\{u_k\}$ is an optimal factorization with respect to Θ_s if and only if P is feasible and, for all $u \in \mathbf{R}^p$.

$$\langle \widehat{\Sigma} u, u \rangle + \left(\lambda \sum_k \theta_s(u_k, u_k) - \langle \widehat{\Sigma}, P \rangle_{\mathrm{HS}} \right) \|u\|_2^2 \leq \theta_s(u, u).$$

3.4.2 A first factored algorithm, a computational snag, and a heuristic

The results of the previous section give a simple abstract recipe for finding a global optimum of (Equation 3.6):

1. We optimize (Equation 3.8) over a fixed number r of rank-1 factors (i.e., vectors $u_1, \ldots, u_r, v_1, \ldots, v_r$) until we reach a first-order stationary point (by satisfying condition (a) in Corollary 2). Note that whenever condition (a) is not satisfied, it is easy to find a descent direction, since we can simply rescale the vectors u_k, v_k in a similar manner to the second part of the proof of Lemma 7.

2. At a first-order stationary point, if condition (b) in Corollary 2 holds, we have reached the global minimum. Otherwise, as in the first part of the proof of Lemma 7, there exists $\tilde{u}, \tilde{v} \in \mathbf{R}^p$ such that $\frac{1}{n} \sum_{i=1}^n (y_i - \langle X_i, B \rangle_{\mathrm{HS}}) \langle X_i \tilde{u}, \tilde{v} \rangle > \lambda \theta_s(\tilde{u}, \tilde{v})$. We set $(u_{r+1}, v_{r+1}) = (\epsilon \tilde{u}, \epsilon \tilde{v})$ for $\epsilon > 0$ small enough to decrease the objective function and go back to step 1.

The algorithm is guaranteed to terminate with a finite r by [80, Theorem 2].

The most difficult part to implement is step 2. Checking condition (b) requires maximizing a bilinear form on vectors u, v under a bound on $\theta_s(u, v)$. If we could maximize this for general bilinear forms, we could also solve sparse PCA (see Section 3.4.1), so we suspect it is not possible. However, this does not preclude positive results that exploit the particular structure of the phase retrieval problem.

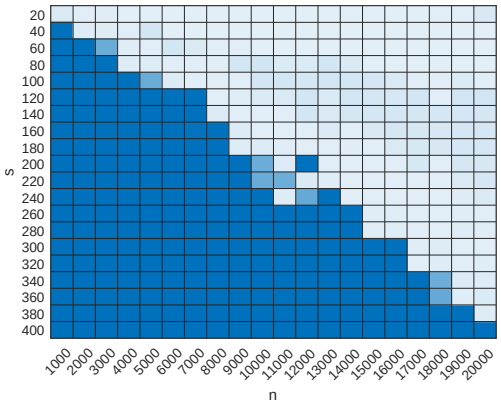
To implement a practical algorithm, we take a very simple shortcut: instead of checking condition (b) over *all* vectors $u, v \in \mathbf{R}^p$, we check it over *1-sparse* vectors. We simply calculate whether any element of $\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})X_i$ is greater than $(1 + 1/s)\lambda$. Although we have not yet found a robust theoretical justification, we will see in the next section that this heuristic works reasonably well in practice. We summarize our high-level practical algorithm in Algorithm 1.

Algorithm 1 High-level sparse phase retrieval algorithm

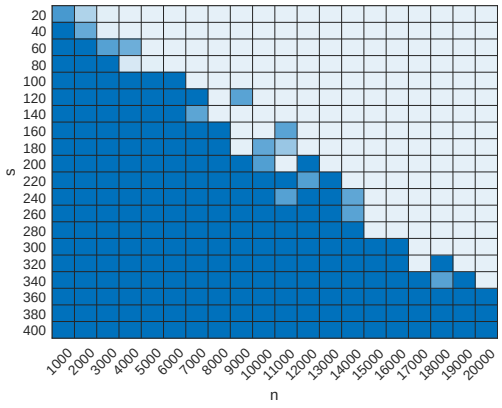
$r \leftarrow 1$
Initialize u_1, v_1 (e.g., some spectral algorithm)
while not Converged **do**
    Optimize (Equation 3.8) over $\{u_1, \dots u_r\}, \{v_1, \dots, v_r\}$ with first-order method until condition (a) in Corollary 2 is satisfied
    $Z \leftarrow \frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})X_i$, where $B = \sum_{k=1}^r u_k \otimes v_k$
    if $Z_{ij} > (1 + 1/s)\lambda$ for any $i, j \in \{1, \dots, p\}$ **then**
        $r \leftarrow r + 1$
        $u_{r+1} \leftarrow \epsilon e_j, v_{r+1} \leftarrow \epsilon e_i$, where $\epsilon > 0$ is sufficiently small to decrease objective function.
    else
        Converged $\leftarrow$ true
    end if
end while
return $\{u_1, \dots, u_r\}, \{v_1, \dots, v_r\}$

3.4.3 Simulation results

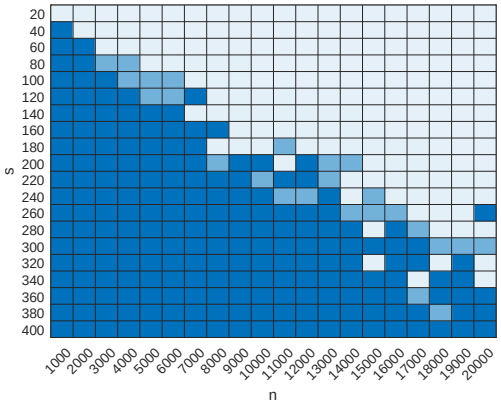
We implemented Algorithm 1 in MATLAB and ran a variety of simulations to illustrate its performance with respect to both sample complexity and noise performance. The interested



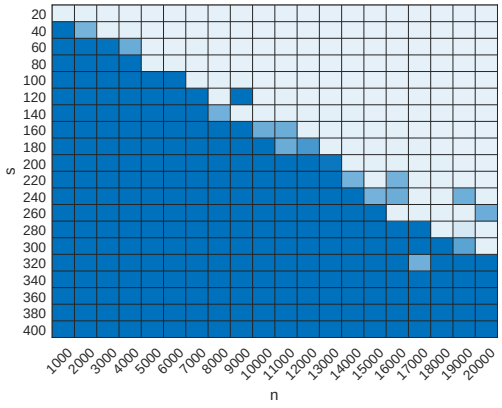
(a) Our algorithm



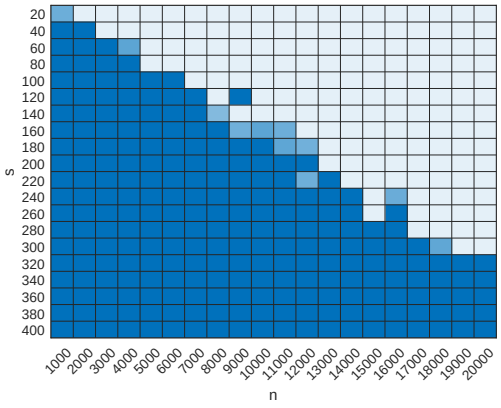
(b) SWF [61]



(c) GAMP [65]

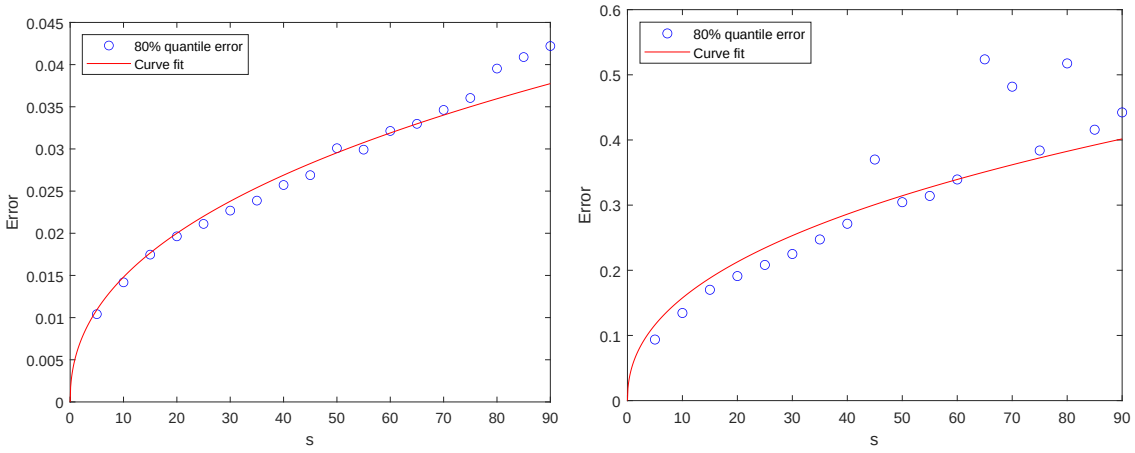


(d) SPARTA [60]



(e) CoPRAM [63]

Figure 3.1: Phase transition plots. Colors represent 80% quantile error over 10 trials (darker colors correspond to higher error). We used $p = 20{,}000$, $\|\beta^*\|_2 = 1$, and $\sigma = 0.02$. All algorithms were run on the same data.



(a) Gaussian noise ($\|\beta^*\|_2 = 1$, $\sigma = 0.05$) (b) Poisson noise ($\|\beta^*\|_2 = 20$)

Figure 3.2: Plot of $\|\hat{\beta} - \beta^*\|_2$ vs. s (80% quantile over 10 trials). All simulations use $p = 10{,}000$ and $n = 6{,}000$. Blue circles are actual data; the red curves are of the form $c\sqrt{s \log \frac{ep}{s}}$, where the scaling factor c is chosen to give minimum mean absolute deviation.

reader can view our code[6] to see more details, but some of the more salient features are the following:

- Line 5 of Algorithm 1 is implemented with alternating minimization over $U = [u_1 \cdots u_r] \in \mathbf{R}^{p \times r}$ and $V = [v_1 \cdots v_r] \in \mathbf{R}^{p \times r}$.
- After each alternating minimization step, we "rebalance" U and V (i.e., rescale each u_k, v_k to force $\theta_s(u_k, u_k) = \theta_s(u_k, v_k) = \theta_s(v_k, v_k)$).
- Each minimization problem over U or V is convex, and we solve it with an accelerated proximal gradient descent algorithm.
- The proximal step requires solving a convex problem of the form

$$\arg\min_{y \in \mathbf{R}^p} \langle x, y \rangle + \frac{a}{2}\|y\|_2^2 + \frac{b}{2}\|y\|_1^2$$

for arbitrary $x \in \mathbf{R}^p$. The solution is a soft-thresholding of x with a data-dependent threshold; the threshold can be found by an efficient iterative algorithm (e.g., a

[6] https://github.com/admcrae/spr2021

Newton algorithm). This problem is closely related to calculating the *dual norm* to $\|y\| \coloneqq \sqrt{\|y\|_2^2 + \frac{1}{s}\|y\|_1^2}$ for arbitrary $s > 0$, i.e., $\|x\|^* = \max_{\|y\| \leq 1} \langle x, y \rangle$. This may be an interesting object for further theoretical study.

All of our simulations used i.i.d. Gaussian measurement vectors $x \sim \mathcal{N}(0, I_p)$.

1. Figure 3.1 shows phase transition diagrams of performance versus sample size n and sparsity s for our algorithm and a variety of alternatives. Note that qualitatively, all these algorithms have similar performance in terms of sample complexity. Interestingly, all of them appear only to require (within a log factor) a number of samples *linear* in the sparsity s. This demonstrates a gap between the empirical performance of all these algorithms and the best theoretical guarantees that have been proved so far.

2. Figure 3.2 shows plots of the error versus sparsity s for both Gaussian noise and Poisson noise. Note that in both cases, the error roughly follows the predicted $\sqrt{s \log(p/s)}$ scaling.

3.5 Conclusion

We have shown that estimators for sparse phase retrieval and sparse PCA obtained by solving a convex program ((Equation 3.6) for sparse phase retrieval and (Equation 3.7) for sparse PCA) with the abstract mixed atomic norm (Equation 3.5) as a regularizer satisfy optimal statistical guarantees in terms of sample complexity and error. For sparse phase retrieval, we have derived a practical heuristic algorithm whose performance matches that of existing state-of-the-art algorithms.

Our work suggests new methods for analyzing these problems (and others with similar sparse factored structure, such as sparse blind deconvolution). It also suggests interesting new research directions in sparse recovery and in optimization. For example, it would be very useful to study *why* our heuristic approach appears to work well for sparse phase

retrieval as well as whether it is possible to do even better. A related problem is to prove that sparse phase retrieval has linear sample complexity with practical algorithms (or that it doesn't, along with why current empirical results seem to suggest otherwise). Similarly, the atomic matrix norm (along with other similar norms) invites further analysis, particularly in how well we can optimize it (where this may depend on the structure of the problem in which it is used). The interplay between statistical guarantees and computational complexity theory (e.g., in sparse PCA) may be very interesting here.

CHAPTER 4

SAMPLE COMPLEXITY AND EFFECTIVE DIMENSION FOR REGRESSION ON MANIFOLDS

In this chapter,[1] we consider the theory of regression on a manifold using reproducing kernel Hilbert space methods. Manifold models arise in a wide variety of modern machine learning problems, and our goal is to help understand the effectiveness of various implicit and explicit dimensionality-reduction methods that exploit manifold structure. Our first key contribution is to establish a novel nonasymptotic version of the Weyl law from differential geometry. From this we are able to show that certain spaces of smooth functions on a manifold are effectively finite-dimensional, with a complexity that scales according to the manifold dimension rather than any ambient data dimension. Finally, we show that given (potentially noisy) function values taken uniformly at random over a manifold, a kernel regression estimator (derived from the spectral decomposition of the manifold) yields minimax-optimal error bounds that are controlled by the effective dimension.

4.1 Introduction

High-dimensional data is ubiquitous in modern machine learning. Examples include images (2-D and 3-D), document texts, DNA, and neural recordings. In many cases, the number of dimensions in the data is much larger than the number of actual data samples. Traditional statistical methods cannot handle such cases, so researchers have turned to a variety of explicit dimensionality-reduction techniques—which make inference more tractable—and to tools such as neural networks that often implicitly transform the data into a much lower-dimensional feature space. These techniques inherently assume that the data have an *intrinsic* dimension that is much lower than that of the data's original representation. Our goal in

[1] This work is published in [83].

this work is to show that the difficulty of a supervised learning problem depends only on this intrinsic dimension and not on the (potentially much larger) ambient dimension. In particular, we consider the common assumption that the data lie on a low-dimensional *manifold* embedded in Euclidean space (see [84, 85, 86, 87] for some of the many example applications).

As an illustration of the kind of results we hope to obtain, we first consider a simple example: a function on the circle S^1 (or, equivalently, a periodic function on the real line). Specifically, suppose that we want to estimate a function f^* on the circle from random samples. In general, it is intractable to estimate an arbitrary function from finitely many samples, but it becomes possible if we assume f^* is structured. For example, f^* may exhibit a degree of smoothness, which can be readily characterized via the *Fourier series* for f^*. Specifically, recall that we can write f^* as the Fourier series sum $f^*(x) = a_0 + \sum_{\ell \geq 1}(a_\ell \cos(2\pi \ell x) + b_\ell \sin(2\pi \ell x))$. One common notion of smoothness in signal processing is that f^* is *bandlimited*, meaning that this sum can be truncated at some largest frequency Ω. In this case, f^* lies in a subspace of dimension at most $p(\Omega) = 2\lfloor \Omega/2\pi \rfloor + 1$. We know (see, e.g., [1, Chapter 12] or [88]) that we can recover such a function exactly, with high probability, from $n \gtrsim p(\Omega) \log p(\Omega)$ samples placed uniformly at random. If there is measurement noise, the squared L_2 error due to noise scales like $\frac{p(\Omega)}{n}\sigma^2$. In higher dimensions (say, on the torus T^m), an Ω-bandlimited function lies in a space of dimension $p(\Omega) = O(\Omega^m)$, and the number of random samples required scales accordingly.

Another model for smoothness is that f^*, rather than being bandlimited, has exponentially decaying frequency components. For example, suppose the Fourier coefficients satisfy $\sum_\ell e^{t\ell^2}(a_\ell^2 + b_\ell^2) < \infty$ for some $t > 0$ (this is roughly equivalent to f^* being the convolution of a Gaussian function with an arbitrary function in L_2). The space of such functions is infinite-dimensional, but any function in it can be approximated as Ω-bandlimited to within an error of size $O(e^{-c\Omega^2 t})$, which should enable us to recover a close approximation to f^* from $O(p(\Omega) \log p(\Omega))$ samples.

In this work, we provide precise analogs of these sample complexity results in the general case of a function on an arbitrary manifold $\mathcal{M}$ with dimension m. As on the circle or torus, an L_2 function $f(x)$ on a Riemannian manifold has a *spectral decomposition* into modes $u_\ell(x)$ corresponding to vibrational frequencies ω_ℓ for all non-negative integers ℓ; these modes are the eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$. Our first key contribution (described in Theorem 8) is a nonasymptotic version of the *Weyl law* from differential geometry: this states that, for large enough Ω, the set $\mathcal{H}_\Omega^{\text{bl}}$ of Ω-bandlimited functions on $\mathcal{M}$ (functions composed of modes with frequencies below Ω) has dimension $\dim(\mathcal{H}_\Omega^{\text{bl}}) \leq C_m \operatorname{vol}(\mathcal{M})\, \Omega^m =: p(\Omega)$. Thus the number of degrees of freedom scales according to the *manifold dimension* m rather than a larger ambient dimension.

Our second key contribution is an error bound for recovering functions on $\mathcal{M}$ from randomly-placed samples using kernel regression. We show in Theorem 9 that if we take $n \gtrsim p(\Omega) \log p(\Omega)$ samples of f^*, we can recover any Ω-bandlimited function with error

$$\frac{\|\hat{f} - f^*\|_{L_2}^2}{\operatorname{vol}(\mathcal{M})} \lesssim \frac{p(\Omega)}{n}\sigma^2,$$

which is precisely the error rate for parametric regression in a $D(\Omega)$-dimensional space. Our results extend further to approximately-bandlimited functions: for example, if f^* satisfies $\sum_\ell a_\ell^2 e^{t\omega_\ell^2} < \infty$, where $f^* = \sum_\ell a_\ell u_\ell$, then, again with $n \gtrsim p(\Omega) \log p(\Omega)$ samples, we get (Theorem 10)

$$\frac{\|\hat{f} - f^*\|_{L_2}^2}{\operatorname{vol}(\mathcal{M})} \lesssim \frac{p(\Omega)}{n}\sigma^2 + O(e^{-c\Omega^2 t}).$$

Both bounds are minimax optimal in the presence of noise.

These results follow from our Theorem 7, which is a more general result on regression in a reproducing kernel Hilbert space. Theorems 9 and 10 adapt this result to a specific choice of kernel.

This chapter is organized as follows. Sections section 4.2 and section 4.3 describe our framework, survey the relevant literature, and compare it to our results. Section 4.4 contains

our main theoretical results. The proofs are in the appendices in the supplementary material. The key technical results are Theorem 7, which is proved via empirical risk minimization and operator concentration inequalities, and Lemma 9 (used to prove Theorem 8), which is proved via heat kernel comparison results on manifolds of bounded curvature.

4.2 Framework and notation

4.2.1 Kernel regression and interpolation

Kernels provide a convenient and popular framework for nonparametric function estimation. They allow us to treat the evaluation of a nonlinear function as a *linear* operator on a Hilbert space, and they give us a computationally feasible way to estimate such a function (which is often in an infinite-dimensional space) from a finite set of samples. Here, we review some of the key ideas that we will need in analyzing kernel methods.

Let S be an arbitrary set, and suppose $k\colon S \times S \to \mathbf{R}$ is a positive definite kernel. Let $\mathcal{H}$ be its associated reproducing kernel Hilbert space (RKHS), characterized by the identity $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $x \in S$.

Now, suppose we have $X_1, \dots, X_n \in S$, $f^* \in \mathcal{H}$ is an unknown function, and we observe $Y_i = f^*(X_i) + \xi_i$ for $i = 1, \dots, n$, where the ξ_i's represent noise. A common estimator for f^* is the regularized empirical risk minimizer

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}} \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \alpha\|f\|_{\mathcal{H}}^2, \tag{4.1}$$

where $\alpha \geq 0$ is a regularization parameter. The solution to the optimization problem (Equation 4.1) is

$$\hat{f}(x) = \sum_{i=1}^{n} a_i k(x, X_i), \tag{4.2}$$

where $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathbf{R}^n$ is given by

$$\boldsymbol{a} = (n\alpha \boldsymbol{I}_n + \boldsymbol{K})^{-1}\boldsymbol{Y},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \in \mathbf{R}^n$, $\boldsymbol{K}$ is the kernel matrix on $X_1, \ldots, X_n$ defined by $\boldsymbol{K}_{ij} = k(X_i, X_j)$, and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

In general, $\hat{f}$ corresponds to a *ridge regression* estimate of f^*. The limiting case $\alpha = 0$ can be recast as the problem

$$\hat{f} = \underset{f \in \mathcal{H}}{\arg\min} \ \|f\|_{\mathcal{H}} \text{ s.t. } Y_i = f(X_i), \ i = 1, \ldots, n.$$

In this case, if the X_i's are distinct, then $\hat{f}$ interpolates the measured values of f^*.

4.2.2 Kernel integral operator and eigenvalue decomposition

A common tool for analyzing kernel interpolation and regression, which will play a central role in our analysis in Section 4.4, is the eigenvalue decomposition of a kernel's associated integral operator. The integral operator $\mathcal{T}$ is defined for functions f on S by

$$(\mathcal{T}(f))(x) = \int_S k(x, y) f(y) \, d\mu(y),$$

where μ is a measure on S. Under certain assumptions[2] on S, μ, and k, $\mathcal{T}$ is a well-defined operator on $L_2(S)$, is compact and positive definite with respect to the L_2 inner product, and has eigenvalue decomposition

$$\mathcal{T}(f) = \sum_{\ell=1}^{\infty} t_\ell \langle f, v_\ell \rangle_{L_2} v_\ell, \ f \in L_2(S),$$

where the eigenvalues $\{t_\ell\}$ are arranged in decreasing order and converge to 0, and the eigenfunctions $\{v_\ell\}$ are an orthonormal basis for $L_2(S)$. We also have $k(x, y) = \sum_{\ell=1}^{\infty} t_\ell v_\ell(x) v_\ell(y)$, where the convergence is uniform and in L_2.

This eigendecomposition plays an important role in characterizing the RKHS $\mathcal{H}$ associated with the kernel k. Combining this expression for k with the identity $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} =$

[2] E.g., S is a compact metric space; μ is strictly positive, finite, and Borel; and k is continuous [89].

$f(x)$, we can derive the fact that, for all $f, g \in \mathcal{H}$,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\langle f, v_\ell \rangle_{L_2} \langle g, v_\ell \rangle_{L_2}}{t_\ell}.$$

This implies that $\langle f, g \rangle_{L_2} = \langle \mathcal{T}^{1/2}(f), \mathcal{T}^{1/2}(g) \rangle_{\mathcal{H}}$ for all $f, g \in L_2(S)$. Thus $\mathcal{T}^{1/2}$ is an isometry from $L_2(S)$ to $\mathcal{H}$, and so for any $f \in \mathcal{H}$, we can write $f = \mathcal{T}^{1/2}(f_0)$, where $\|f_0\|_{L_2} = \|f\|_{\mathcal{H}}$. This implies that, for any $p \geq 1$, the projection of f onto $(\mathrm{span}\{v_1, \ldots, v_p\})^\perp$ has L_2 norm at most $\sqrt{t_{p+1}}\|f\|_{\mathcal{H}}$. Hence the decay of the eigenvalues $\{t_\ell\}$ of $\mathcal{T}$ characterizes the "effective dimension" of $\mathcal{H}$ in L_2, which will be a fundamental building block for our analysis.

4.2.3 Spectral decomposition of a manifold and related kernels

We now turn to our specific problem of regression on a manifold, considering how an RKHS framework can help us. The book [90] is an excellent reference for the material in this section.

A smooth, compact Riemannian manifold $\mathcal{M}$ (without boundary) can be analyzed via the *spectral decomposition* of its Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ (we will often call it the Laplacian for short). This operator is defined as $\Delta_{\mathcal{M}} f := -\operatorname{div}(\nabla f)$. In $\mathbf{R}^m$, it is simply the operator $-\sum_{i=1}^{m} \frac{\partial^2}{\partial x_i^2}$. The Laplacian can be diagonalized as

$$\Delta_{\mathcal{M}} f = \sum_{\ell=0}^{\infty} \lambda_\ell \langle f, u_\ell \rangle_{L_2} u_\ell,$$

where $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots$, the sequence $\lambda_\ell \to \infty$ as $\ell \to \infty$, and $\{u_\ell\}$ is an orthonormal basis for $L_2(\mathcal{M})$ (all integrals are with respect to the standard volume measure on $\mathcal{M}$).

The eigenvalues $\{\lambda_\ell\}$ are the squared resonant frequencies of $\mathcal{M}$, and the eigenfunctions $\{u_\ell\}$ are the vibrating modes, since solutions to the wave equation $f_{tt} + \Delta_{\mathcal{M}} f = 0$ on $\mathcal{M}$

have the form

$$f(t,x) = \sum_{\ell=0}^{\infty} (a_\ell \sin\sqrt{\lambda_\ell} t + b_\ell \cos\sqrt{\lambda_\ell} t) u_\ell(x).$$

The classical Weyl law (e.g., [90, p. 9]) says that, if $\mathcal{M}$ has dimension m, then, asymptotically,

$$|\{\ell : \lambda_\ell \le \lambda\}| \sim c_m \operatorname{vol}(\mathcal{M}) \lambda^{m/2}$$

as $\lambda \to \infty$, where $c_m = (2\pi)^{-m} V_m$, with V_m denoting the volume of the unit ball in $\mathbf{R}^m$.

Using the spectral decomposition of the Laplacian, any number of kernels can be defined by

$$k(x,y) = \sum_{\ell=0}^{\infty} g(\lambda_\ell) u_\ell(x) u_\ell(y)$$

for some function g. With this construction, the integral operator of k has eigenvalue decomposition $\mathcal{T}(f) = \sum_{\ell \ge 0} g(\lambda_\ell) \langle f, u_\ell \rangle_{L_2} u_\ell$, hence, per Section 4.2.2, $\|f\|_{\mathcal{H}}^2 = \sum_{\ell \ge 0} \frac{\langle f, u_\ell \rangle^2}{g(\lambda_\ell)}$.

Our results could, in principle, apply to many kernels with the above form, but we will primarily consider *bandlimited kernels* and the *heat kernel*. The bandlimited kernel with bandlimit $\Omega > 0$ is

$$k_\Omega^{\mathrm{bl}}(x,y) = \sum_{\lambda_\ell \le \Omega^2} u_\ell(x) u_\ell(y),$$

which is the reproducing kernel of the space of bandlimited functions on $\mathcal{M}$:

$$\mathcal{H}_\Omega^{\mathrm{bl}} = \left\{ f \in L_2(\mathcal{M}) \colon f \in \operatorname{span}\{u_\ell \colon \lambda_\ell \le \Omega^2\} \right\}$$

with $\|f\|_{\mathcal{H}_\Omega^{\mathrm{bl}}} = \|f\|_{L_2}$ for $f \in \mathcal{H}_\Omega^{\mathrm{bl}}$. The heat kernel is a natural counterpart to the common Gaussian radial basis function on $\mathbf{R}^m$. Detailed treatments can be found in [90, 91]. We will define it for $t > 0$ as

$$k_t^{\mathrm{h}}(x,y) = \sum_{\ell=0}^{\infty} e^{-\lambda_\ell t/2} u_\ell(x) u_\ell(y).$$

Its corresponding RKHS is

$$\mathcal{H}_t^{\mathrm{h}} = \left\{ f \in L_2(\mathcal{M}) \colon \|f\|_{\mathcal{H}_t^{\mathrm{h}}}^2 = \sum_{\ell=0}^{\infty} e^{\lambda_\ell t/2} \langle f, u_\ell \rangle_{L_2}^2 < \infty \right\}.$$

The heat kernel k_t^{h} gets its name from the fact that it is the fundamental solution to the heat equation $f_t + \frac{1}{2}\Delta_{\mathcal{M}} f = 0$ on $\mathcal{M}$. The heat kernel on $\mathbf{R}^m$ is $k_t^{\mathrm{h}}(x,y) = \frac{1}{(2\pi t)^{m/2}} e^{-\|x-y\|^2/2t}$.

4.3 Related work

4.3.1 Dimensionality reduction and low-dimensional structure

There is an extensive literature on the use of low-dimensional manifold structure in machine learning. Perhaps most prominently, nonlinear dimensionality-reduction techniques that exploit manifold structure have been developed, such as [92, 93, 94, 95, 96]. More recently, there has been explicit inclusion of manifold models into neural network architectures [97, 98, 99, 100]. However, none of this research provides nonasymptotic performance guarantees.

On the other hand, the field of high-dimensional statistics provides many theoretical guarantees for low-dimensional data models. For example, there are extensive bodies of theory for models such as sparsity [101, 102] and low-rank structure [103]. One can view low-dimensional manifold models as a more powerful generalization of such structures. One interesting work that bridges the gap between manifold models and high-dimensional statistics is [104], which is another explicit dimensionality-reduction technique. Another similar line of work is the study of algebraic variety models (e.g., [105]), which are also nonlinear and low-dimensional.

While the great success of the many implicit and explicit dimensionality-reducing methods provides empirical evidence for the possibility of exploiting manifold structure, there are still very large gaps in our theoretical understanding of when and why these methods can be effective.

4.3.2 Manifold regression and kernels

Regression on manifold domains has been explored in a number of previous works. The closely-related problem of density estimation is considered in [106, 107]. Particularly relevant to our work, [106] uses the same bandlimited kernel and heat kernel that we highlight (and it analyzes the spectral decomposition of these kernels via the asymptotic Weyl law). It is primarily interested in the power of the error rate that can be obtained by assuming the function (density) of interest has a certain number of derivatives; in particular, it shows that $\|\hat{f} - f^*\|_{L_2}^2 \lesssim n^{-2s/(m+2s)}$ if f has s bounded derivatives. Both works, like ours, assume explicit knowledge of the manifold.

Perhaps more relevant to practical applications, [108] seeks to provide a manifold-agnostic algorithm via local linear approximations to the data manifold; however, it is also primarily interested in asymptotic error rates. The paper [109] examines related methods asymptotically in more detail. Another manifold-agnostic method similar in spirit to ours is that of [110], who consider kernel estimation with (Euclidean) Gaussian radial basis functions. They obtain the optimal $n^{-2s/(m+2s)}$ rate for s-smooth regression functions; however, their assumptions are quite different from ours in that their regression functions must have *smooth extensions* to (a neighborhood in) the embedding space. Similarly, [111] obtain the optimal rate for functions that are s-smooth (in the manifold calculus, similarly to our assumptions) using a neural-network–type architecture. However, they implicitly assume that the manifold is C^∞-embedded in Euclidean space.

In [112, 113], the authors explore Gaussian process models (which are closely related to kernel methods) on a manifold.

The error rate $\|\hat{f} - f^*\|_{L_2}^2 \lesssim n^{-2s/(m+2s)}$ is standard (and minimax optimal) in nonparametric statistics. However, our function model and results are quite different in nature. The regression functions we consider are *infinitely* smooth, and we show that the estimation of these functions is much like a *finite-dimensional* regression problem; not only do we get an n^{-1} error rate (as we do when we take $s \to \infty$ above), but the constant in front of this

rate and the minimum number of samples needed are proportional to the finite effective dimension.

Finally, we also note that the idea of using a kernel that can be expressed in terms of the spectral decomposition of a manifold's Laplacian also has precedent. In addition to [106], the paper [114] suggests using such kernels for interpolation in Sobolev spaces on a manifold.

4.3.3 General kernel interpolation and regression

Regression is a strict superset of interpolation; interpolation typically assumes that we sample function values exactly (i.e., there is no noise), while regression allows for (and often assumes) noise.

There is a substantial literature on the use of a kernel for interpolation of functions in an RKHS (often, in this literature, referred to as the "native space" of the kernel). A fairly comprehensive survey can be found in [115]. Distinct from our work, most of this literature considers *deterministic* samples of the function of interest. Given (deterministic) sample locations $\{X_1, \ldots, X_n\} \subset S$, results in this literature tend to have the form $\|\hat{f} - f^*\|_\infty \leq g(h_X)\|f^*\|_{\mathcal{H}}$, where $h_X = \max_{x \in S} \min_{i \in \{1,\ldots,n\}} d(x, X_i)$, and $g(h)$ is a function that decreases to 0 as $h \to 0$ at a rate that depends on the properties of the kernel k (typically as a power or exponentially). Some recent work applying kernel interpolation theory to manifolds is [116, 117, 118].

Much of the literature on (noisy) RKHS regression primarily considers the case when the eigenvalues of the integral operator (described in Section 4.2.1) decay as $t_\ell \lesssim \ell^{-b}$. In [119, 120, 121], it is shown that the minimax optimal error rate is $\|f^* - \hat{f}\|_{L_2}^2 \lesssim n^{-b/(b+1)}$. Many other recent papers have explored this rate of convergence in a variety of settings [122, 123, 124, 125]. Several of these include more general spectral regularization algorithms, suggested by [126]. Some interesting recent extensions consider a variety of algorithms that may be more practical for large-data situations. These include iterative methods [127, 128,

129] and distributed algorithms [130, 131, 132].

Another set of results (which are the most similar to ours) uses a regularized effective dimension $p_\alpha = \sum_\ell \frac{t_\ell}{\alpha + t_\ell}$, where α is the regularization parameter. This is considered in [133] and greatly refined in [134]. Variations on these results can be found in [135]. See Section 4.4.1 for further discussion and comparison to our results. The earlier report [136] resembles our work in its analysis of truncated operators. We note that in the case of power-law eigenvalue decay, these results (and ours) recover the $n^{-b/(b+1)}$ error rate.

It is interesting to note that the squared error rate $n^{-b/(b+1)}$ can recover the standard rate for regression of s-smooth functions on manifolds. The Sobolev space of order s is the RKHS of the kernel $\sum_\ell (1 + \lambda_\ell)^{-s} u_\ell(x) u_\ell(y)$. By the Weyl law, its eigenvalues decay according to $t_\ell \approx \ell^{-2s/m}$; plugging $2s/m$ in for b recovers the standard rate $n^{-2s/(m+2s)}$.

4.4 Main theoretical results

4.4.1 Dimensionality in RKHS regression

Here we present our main results for general regression and interpolation in an RKHS. Our results also apply to the slightly more general setting of learning in an arbitrary Hilbert space (see, e.g., [134]), but we do not explore this here. We continue to use the notation established in Sections 4.2.1 and 4.2.2, and we further assume that $\mu(S) = 1$ (since μ is finite, we can always obtain this by a rescaling). We assume that the function samples we take are uniformly distributed on S:

Assumption 3. The sample locations $X_1, \ldots, X_n$ are i.i.d. according to μ.

Since $\mathcal{H}$ is, in general, infinite-dimensional, there is typically no hope of recovering an arbitrary $f^* \in \mathcal{H}$ to within a small error in $\mathcal{H}$-norm from a finite number of measurements. However, the discussion in Section 4.2.2 suggests a more feasible goal. Since any set of functions bounded in $\mathcal{H}$-norm can be approximated within an arbitrarily small L_2 error in a finite-dimensional subspace of L_2, as long as the number of measurements is proportional to

this loosely-defined "effective dimension" of $\mathcal{H}$, we have hope of recovering f^* accurately in an L_2 sense.

Let $p > 0$ be a fixed integer dimension. Let $G = \operatorname{span}\{v_1, \dots, v_p\} \subset \mathcal{H} \cap L_2(S)$, and let $G^\perp$ be its orthogonal complement in $L_2(S)$ and $\mathcal{H}$. We denote by $\mathcal{T}_G$ and $\mathcal{T}_{G^\perp}$ the restrictions of $\mathcal{T}$ onto G and $G^\perp$, respectively. We make the following assumptions on the eigenvalues and eigenfunctions of $\mathcal{T}$:

Assumption 4. We have $\sum_{\ell=1}^{p} v_\ell^2(x) \le K_p$ and $\sum_{\ell=p+1}^{\infty} t_\ell v_\ell^2(x) \le R_p$ uniformly over almost every $x \in S$ for some constants K_p and R_p independent of x.

This says that the energy of the eigenfunctions of $\mathcal{T}$ is reasonably spread out over the domain S—for the basis $\{v_1, \dots, v_p\}$, this is a type of incoherence assumption. If the eigenfunctions are well-behaved, we can expect $K_p \approx p$ and $R_p \approx \operatorname{tr} \mathcal{T}_{G^\perp}$. This holds in our original example of the Fourier series on the circle, since the sinusoid basis functions are bounded by an absolute constant. Our "pointwise" Weyl law in Theorem 8 shows that we have similar behavior for the spectral decomposition of a manifold. Note that K_p in Assumption 4 is identical to the quantity $K(p)$ in [88], which uses similar methods to handle a much simpler problem.

Assumption 5. For some $\gamma, \gamma' \ge 0$, we have $\frac{\operatorname{tr} \mathcal{T}_{G^\perp}}{t_{p+1}} \le \gamma p$ and $\frac{R_p}{t_{p+1}} \le \gamma' K_p$.

This assumption greatly simplifies the notation of our results and is always true with an appropriate choice of γ and γ'. γ is often small when t_{p+1} is in the decaying "tail" of eigenvalues. If the eigenvalues decay like $t_\ell \approx \ell^{-b}$, we can take $\gamma \approx (b-1)^{-1}$. Note that a similar assumption appears in [137]. If $K_p \approx p$ and $R_p \approx \operatorname{tr} \mathcal{T}_{G^\perp}$, then $\gamma \approx \gamma'$.

With these assumptions in place, we can state our main theorem for RKHS regression:

Theorem 7. *Suppose Assumptions 3 to 5 hold. Let $\delta \in (0,1)$. If*

$$n \ge (7 \vee 3\gamma') K_p \log \frac{(2 \vee 4\gamma)p}{\delta},$$

then the following hold for the kernel estimate $\hat{f}$ with regularization parameter $\alpha \geq 0$:

1. *If there is no noise, that is, $Y_i = f^*(X_i)$ for each i, then, with probability at least $1 - \delta$, uniformly in f^*,*

$$\|\hat{f} - f^*\|_{L_2} \leq (\sqrt{2\alpha} + 6\sqrt{t_{p+1}})\|f^*\|_{\mathcal{H}}.$$

2. *Now suppose that $Y_i = f(X_i) + \xi_i$, where the ξ_i's are i.i.d., zero-mean, sub-exponential random variables with variance σ^2 and are independent of the X_i's. If we additionally have*

$$\frac{n}{\log^2 n} \geq C(1 \vee \gamma')\frac{K_p}{p}\frac{\|\xi\|_{\psi_1}^2}{\sigma^2},$$

where C is a universal constant, and $\alpha \geq 54 t_{p+1}$, then, with probability at least $1 - 2\delta$, uniformly in f^*,

$$\|\hat{f} - f^*\|_{L_2} \leq (\sqrt{2\alpha} + 6\sqrt{t_{p+1}})\|f^*\|_{\mathcal{H}} + 4\left(1 + \frac{\sqrt{\gamma}}{8}\right)\frac{\sqrt{p} + 2\sqrt{\log 4/\delta}}{\sqrt{n}}\sigma.$$

Our results guarantee an L_2 recovery error bounded by two terms: (1) a "bias" depending on the next tail eigenvalue t_{p+1} and the regularization coefficient α, and (2) a "variance" term that behaves similarly to the error found in p-dimensional regression. When $K_p \approx p$, this result yields the $n \gtrsim p \log p$ sample complexity that we expect. If $\mathcal{H}$ is, in fact, p-dimensional (which our framework can handle with $t_\ell = 0$ for $\ell > p$), this result recovers standard p-dimensional regression bounds such as those in [88].

We assume i.i.d. noise for simplicity, but our result could easily be extended beyond this case. Note that if the noise is Gaussian, the ratio $\|\xi\|_{\psi_1}^2/\sigma^2$ is an absolute constant.

For interpolation ($\alpha = 0$) in the noiseless case, this theorem yields $\|\hat{f} - f^*\|_{L_2} \leq 6\sqrt{t_{p+1}}\|f^*\|_{\mathcal{H}}$. In the noisy case, the lower bound on α can be relaxed to get a result with worse constants. We obtain qualitatively similar results whenever $\alpha \gtrsim t_{p+1}$. The assumptions and results of [134] (specialized to our setting) are comparable to Theorem 7

when we set $\alpha \approx t_{p+1}$. However, our results have the advantage of applying even in infinite-dimensional settings with no regularization: the regularized effective dimension $p_\alpha = \sum_\ell \frac{t_\ell}{\alpha + t_\ell}$ from their work would be infinite if $\alpha = 0$.

Although we do not explore it here, we note that one could generalize our approach to the case where the sampling measure differs from that under which the L_2 norm is calculated. We could simply bound the ratio (Radon-Nikodym derivative) between the two measures, or we could perform leverage-score sampling to mitigate the need for bounding the eigenfunctions (see, e.g., [137] for similar ideas).

In the presence of noise, Theorem 7 is minimax optimal over the set $\{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq r\}$ for any $r > 0$ if p is chosen so that $\frac{p}{n}\sigma^2 \approx t_{p+1} r^2$. In this case,

$$\left\{ f \in \operatorname{span}\{v_1, \ldots, v_p\} : \|f\|_{L_2} \lesssim \sqrt{\frac{p}{n}}\sigma \right\} \subset \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq r\},$$

and the minimax rate (with, say, Gaussian noise) over the left-hand set is well-known to be $\sqrt{\frac{p}{n}}\sigma$.

4.4.2 Manifold function estimation

We now describe how we can leverage Theorem Theorem 7 to establish sample complexity bounds for regression on a manifold. Suppose, again, that $\mathcal{M}$ is an m-dimensional smooth, compact Riemannian manifold. To study the eigenvalues and eigenfunctions of the Laplacian $\Delta_{\mathcal{M}}$, we consider the heat kernel k_t^{h}. Our key tool is the following fact:

Lemma 9. *Let $\epsilon \in (0, 2/3)$. Suppose the sectional curvature of $\mathcal{M}$ is bounded above by κ. For $t \leq \frac{6\epsilon}{(m-1)^2\kappa}$ and all $x \in \mathcal{M}$,*

$$k_t^{\mathrm{h}}(x, x) \leq \frac{1+\epsilon}{(2\pi t)^{m/2}}.$$

This is a precise quantification of the well-known asymptotic behavior of the heat kernel as $t \to 0$ (see, e.g., [90, Section VI]). It is derived in Section B.2 from a novel set of more

general upper and lower bounds for the heat kernel on a manifold of bounded curvature; we note that these may be of independent interest.

Our nonasymptotic Weyl law is a simple consequence of Lemma 9:

Theorem 8. *If $\mathcal{M}$ has sectional curvature bounded above by κ, and $\epsilon \in (0, 2/3)$, then, for all $x \in \mathcal{M}$ and $\lambda \geq \frac{m(m-1)^2\kappa}{6\epsilon}$,*

$$N_x(\lambda) \coloneqq \sum_{\lambda_\ell \leq \lambda} u_\ell^2(x) \leq \frac{2(1+\epsilon)\sqrt{m}}{(2\pi)^m} V_m \lambda^{m/2}.$$

With appropriate rescaling by $\mathrm{vol}(\mathcal{M})$, this gives us a bound on the constant K_p from Section 4.4.1. Since this result bounds the eigenfunctions, it is a type of "local Weyl law" (see, e.g., [138]). Integrating this result over $\mathcal{M}$ gives a nonasymptotic version of the traditional Weyl law. Our bound is within the modest factor $2(1+\epsilon)\sqrt{m}$ of the optimal asymptotic law. For simplicity, we will take $\epsilon = 1/2$ in what follows, but slightly better constants could be obtained with smaller ϵ.

The following result for the finite-dimensional bandlimited kernel is a straightforward consequence of Theorems 7 and 8:

Theorem 9. *Suppose the sectional curvature of $\mathcal{M}$ is bounded above by κ. Let $\Omega^2 \geq \frac{m(m-1)^2\kappa}{3}$, and suppose $f^* \in \mathcal{H}_\Omega^{\mathrm{bl}}$. Let $\hat{f}$ be the kernel regression estimate with kernel k_Ω^{bl}.*[3]

Let $\delta \in (0,1)$, and suppose $n \geq 7p \log \frac{2p}{\delta}$, where

$$p = p(\Omega) \coloneqq \frac{3\sqrt{m}\, V_m}{(2\pi)^m} \mathrm{vol}(\mathcal{M})\, \Omega^m. \tag{4.3}$$

Under the same noise assumptions as in Theorem 7, if $\frac{n}{\log^2 n} \geq C\|\xi\|_{\psi_1}^2/\sigma^2$, then, with

[3]The calculation of this estimate is somewhat different than usual, since the rank of the kernel matrix $\boldsymbol{K}$ is at most the dimension of $\mathcal{H}_\Omega^{\mathrm{bl}}$. We do not use regularization, but we use the Moore-Penrose pseudoinverse of $\boldsymbol{K}$ instead of $\boldsymbol{K}^{-1}$.

probability at least $1 - 2\delta$*, uniformly in* f^**,*

$$\frac{\|\hat{f} - f^*\|_{L_2}}{\sqrt{\mathrm{vol}(\mathcal{M})}} \leq 4\frac{\sqrt{p} + 2\sqrt{\log 4/\delta}}{\sqrt{n}}\sigma.$$

To analyze the heat kernel, which has an infinite number of nonzero eigenvalues, we need the following additional corollary of Lemma 9, which will let us bound the constant R_p from Section 4.4.1:

Lemma 10. *For* $\epsilon \in (0, 2/3), t \leq \frac{6\epsilon}{(m-1)^2\kappa}$*,* $\lambda \geq m/t$*, and all* $x \in \mathcal{M}$*,*

$$\sum_{\lambda_\ell \geq \lambda} e^{-\lambda_\ell t/2} u_\ell^2(x) \leq e^{-\lambda t/2} \frac{2(1+\epsilon)\sqrt{m}}{(2\pi)^m} V_m \lambda^{m/2}.$$

From this, we obtain the following result:

Theorem 10. *Suppose the sectional curvature of* $\mathcal{M}$ *is bounded above by* κ*. Let* $t \leq \frac{3}{(m-1)^2\kappa}$*, and suppose* $f^* \in \mathcal{H}_t^{\mathrm{h}}$*. Fix* $\Omega^2 \geq m/t$*, and let* $\hat{f}$ *be the kernel regression estimate of* f^* *with kernel* k_t^{h} *and regularization parameter* $\alpha \geq 54\frac{e^{-\Omega^2 t/2}}{\mathrm{vol}(\mathcal{M})}$*.*

Let $\delta \in (0, 1)$*, and suppose* $n \geq 7p \log \frac{4p}{\delta}$*, with* p *defined as in* (Equation 4.3)*.*

Under the same noise assumptions as in Theorem 7, if $\frac{n}{\log^2 n} \geq C\|\xi\|_{\psi_1}^2/\sigma^2$*, then, with probability at least* $1 - 2\delta$*, uniformly in* f^**,*

$$\frac{\|\hat{f} - f^*\|_{L_2}}{\sqrt{\mathrm{vol}(\mathcal{M})}} \leq \left(\sqrt{2\alpha} + 6\sqrt{\frac{e^{-\Omega^2 t/2}}{\mathrm{vol}(\mathcal{M})}}\right)\|f^*\|_{\mathcal{H}_t^{\mathrm{h}}} + \frac{9}{2}\frac{\sqrt{p} + 2\sqrt{\log 4/\delta}}{\sqrt{n}}\sigma.$$

These results illustrate how we can exploit the effective finite dimension of spaces of smooth functions on manifolds in regression. This function space dimension (and hence the sample complexity of regression) grows exponentially in the *manifold dimension*, rather than in the larger ambient data dimension, if $\mathcal{M}$ is embedded in a higher-dimensional space. In practice, the true bandlimited or heat kernels may be difficult to compute. It is an interesting open question whether we can obtain similar results for manifold-agnostic algorithms (the

work of [108], although it does not apply to our function classes, is an interesting potential starting point).

As discussed in Section 4.4.1, our general regression result Theorem 7 is similar to prior results [133, 134], but it has the advantage of applying even without regularization in the noiseless case. However, we note that one could obtain results in many ways comparable (minus this advantage) to Theorems 9 and 10 by plugging Theorem 8 and Lemma 10 into those previous regression results. We could not do this with classical power-law results such as [119, 120, 121], since our eigenvalue decay is *exponential* rather than power-law.

Since the (classical) Weyl law also *lower* bounds the complexity of spaces of bandlimited functions, then, as discussed in Section 4.4.1, Theorems 9 and 10 (for the optimally chosen value of Ω) are minimax optimal when there is noise. Furthermore, the requirement $n \gtrsim p \log p$ is necessary in general: if we consider the torus T^m, recovering arbitrary Ω-bandlimited functions requires every point on T^m to be within distance $O(1/\Omega)$ of a sample point; considering a uniform grid on T^m and a coupon collector argument makes it clear that $n \gtrsim O(\Omega^m \log \Omega^m)$ randomly sampled points are required.

As mentioned in Section 4.4.1 for general kernel learning, these results could be extended to consider nonuniform sampling over the manifold.

There are also some very interesting connections between kernel methods and neural networks. The recent works [139, 140] show that trained multi-layer neural networks approach, in the infinite-width limit, a kernel regression function with a "neural tangent kernel" that depends on the initialization distribution of the weights and the network architecture. This follows literature on the connections between Gaussian processes (closely related to kernel methods) and wide neural networks (see, e.g., [141, 142]). It would be very interesting to explore any potential connections between these and the kernels considered in this work, which are derived from a manifold's spectral decomposition.

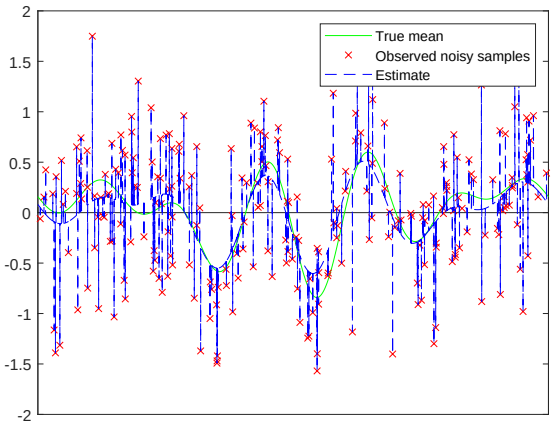
CHAPTER 5
HARMLESS INTERPOLATION IN REGRESSION AND CLASSIFICATION WITH STRUCTURED FEATURES

In this chapter,[1] we analyze the phenomenon of harmless interpolation in regression and classification. Overparametrized neural networks tend to perfectly fit noisy training data yet generalize well on test data. Inspired by this empirical observation, recent work has sought to understand this phenomenon of *benign overfitting* or *harmless interpolation* in the much simpler linear model. Previous theoretical work critically assumes that either the data features are statistically independent or the input data is high-dimensional; this precludes general nonparametric settings with structured feature maps. In this chapter, we present a general and flexible framework for upper bounding regression and classification risk in a reproducing kernel Hilbert space. A key contribution is that our framework describes precise sufficient conditions on the data Gram matrix under which harmless interpolation occurs. Our results recover prior independent-features results (with a much simpler analysis), but they furthermore show that harmless interpolation can occur in more general settings such as features that are a bounded orthonormal system. Furthermore, our results show an asymptotic separation between classification and regression performance in a manner that was previously only shown for Gaussian features.

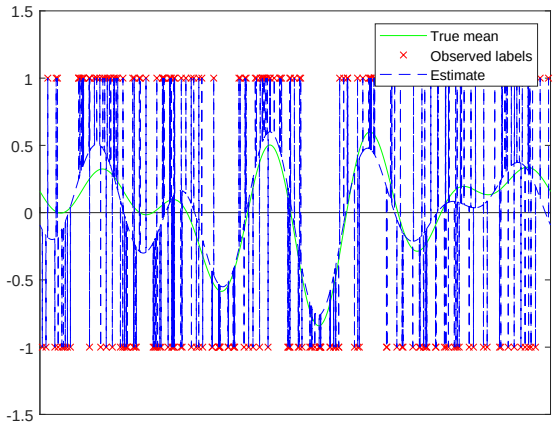
5.1 Introduction

Overparametrized neural networks tend to perfectly fit, or *interpolate*, noisy training data. Somewhat surprisingly, these overparametrized networks also tend to generalize well [144]. More recently, this phenomenon of "harmless interpolation" was also empirically demonstrated in the much simpler model families of kernel machines [145] and overparametrized

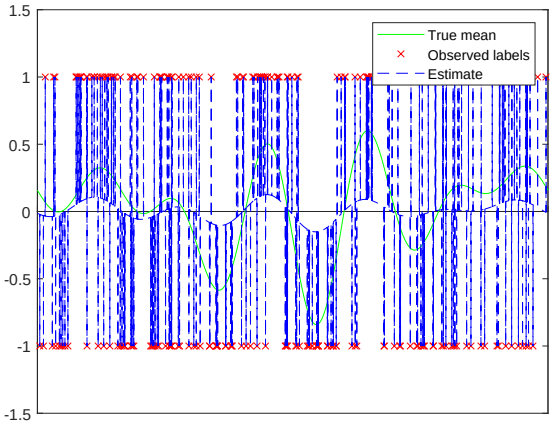
[1] This work is published in [143].



(a) Gaussian-noise observations



(b) Binary labels—good regression performance



(c) Binary labels—poor regression but good classification performance

Figure 5.1: Interpolation in various regimes. This uses the bi-level Fourier series framework of Section 5.4.

linear models [146]. These observations have motivated a large body of research that aims to develop a mathematical understanding of the generalization properties of interpolating solutions and the impact of fitting noise (see Section 5.1.2 for more related work).

While these theoretical results represent significant progress, they come with some caveats. Most notably, harmless interpolation has only been shown under (a) strong assumptions on the feature distribution or (b) high dimension of the input data. For example, the strongest guarantees on harmless interpolation assume that the features consist of independent random variables (or are a linear transformation of such a vector). Similarly, the positive results on consistency of kernel interpolation require the dimension of the input

data to grow with the size of the training set.

To see why these assumptions may not be realistic, consider the problem of simple linear regression using a Fourier series model $f(x) = \sum_\ell a_\ell e^{i2\pi\ell x}$ for a function f on the interval $[0, 1]$, where ℓ may range over all integers or, as we will later assume, a subset $\{-d, \ldots, d\}$. Here the input data dimension is 1, and the features are given by $v_\ell(x) = e^{\mathbf{j}2\pi\ell x}$. If x is uniformly distributed, the features $\{v_\ell(x)\}_\ell$ (all evaluated at the same random x), though *uncorrelated*, are not *independent*. In this (and many other) examples, the input data can be low-dimensional and the features may not be independent. Whether harmless interpolation is possible with high-dimensional feature maps on such *constant*-dimensional data remains an open question. As a first effort, [147] show that harmless interpolation can occur with structured feature maps with uniformly spaced data, but whether this can be shown for the more realistic case of randomly-sampled data has remained open.

A second question is how the interpolation phenomenon applies to the *classification* problem. For example, [148, 149] show that the max-margin support vector machine can achieve good performance even when the corresponding regression task does not. These results require the very strong assumption of independent (sub)Gaussian features. Whether this asymptotic separation between regression and classification tasks exists in more general kernel settings is not addressed by this literature.

5.1.1 Our contributions

In this work, we provide new non-asymptotic risk bounds for both regression and classification tasks with the standard Hilbert-norm regularizer under minimal regularity assumptions. Our results apply for an arbitrarily small amount of regularization (including the interpolating regime) and are summarized below.

Harmless interpolation in kernel regression. For the regression task, we obtain new non-asymptotic risk bounds on the mean-squared-error of the Hilbert-norm regularized estimator, which includes the cases of kernel ridge regression and minimum-Hilbert-norm

interpolation. In Section 5.2.2, we give error bounds for fixed sample locations. In Section 5.2.3, we give a variety of concentration results from random sampling that, when combined with our fixed-sample theorems, yield high-probability guarantees of harmless interpolation. Our results imply harmless interpolation in significantly more general settings than previous works (see Section 5.1.2 for a comparison to prior work). Our results recover existing independent-feature results (e.g., [150]) but also apply to other examples such as bounded orthonormal systems (BOSs). BOSs include many popular feature ensembles such as sinusoids and Chebyshev polynomials. Figure 5.1a shows an example of a function estimate that yields strong regression performance for the case of sinusoidal Fourier basis features.

Asymptotic separation between kernel classification and regression. We next analyze the classification error of the minimum-Hilbert-norm interpolator of binary labels. Although good regression performance implies good classification performance (see Figure 5.1b), the reverse is not true. In Section 5.3, we derive a simple bound on classification error that can be much tighter than the bound on regression error, and we present another fixed-sample error bound useful for bounding the regression risk. Then, for the case of bounded orthonormal system features, we demonstrate an asymptotic separation between the regression and classification tasks. Figure 5.1c illustrates how the minimum-norm label interpolator can have poor regression performance but good classification performance.

5.1.2 Related work

Harmless interpolation. Recent work has shown that the "harmless interpolation" phenomenon becomes more pronounced with increased (effective) overparameterization when the minimum-Hilbert-norm interpolator is used in kernel regression [177, 178] or the minimum-norm interpolator is used in linear regression [150, 179, 180, 147, 162, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191] in a variety of models. See [192, 193, 194] for recent surveys of this line of work.

All of these models make at least one of the following assumptions: (a) independence of features [150, 180, 147, 195], (b) sub-Gaussianity in the feature vector [150, 162], (c) high data dimension [180, 181, 182, 177, 178, 183, 184, 185, 186, 187, 188, 189, 190], or (d) explicit structure in the kernel operator/feature map [179, 177, 178, 147, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191]. For specific kernels like the Laplace kernel, statistically consistent interpolation may actually *require* growing data dimension with the number of training examples [151], as the data dimension fundamentally alters the eigenvalues of the Laplace kernel integral operator. In contrast, our results do not explicitly posit any of these assumptions. Our sufficient conditions for harmless interpolation are expressed purely in terms of the eigenvalues of the kernel integral operator and do not require special structure either on the eigenfunctions or the integral operator itself.

Classification versus regression. General techniques from statistical learning theory (e.g., [152, 153]) do not differentiate between classification and regression tasks. However, the idea that classification is easier than regression is well-known: the main idea is we do not need near-zero bias, but rather how much signal is recovered only needs to be large relative to the variance. This idea goes back to [154], and has primarily been used to obtain faster non-asymptotic rates for classification relative to regression in a number of scenarios [155, 156, 157]. A separation in statistical consistency between the two tasks was shown more recently in [148]. Similar sharp analyses for classification error have also been provided for the related high-dimensional linear discriminant analysis setting [149, 158, 159]. These results all make restrictive assumptions of Gaussianity, independent sub-Gaussian features, or Gaussian mixture models; the most fine-grained analyses [148, 158] require Gaussian design. With our more general analysis, we show that the previous restrictive assumptions can be avoided and demonstrate that the separation between classification and regression consistency is a general phenomenon. Although our error expressions are less sharp nonasymptotically than those that assume Gaussian features, the consistency implications are nearly identical.

General kernel regression. Finally, our work continues a substantial literature on general linear and RKHS regression. Space limitations prevent a comprehensive review, but we note that our analysis techniques most closely resemble the approach of [133, 134], who analyze explicitly regularized ridge regression under random design with minimal assumptions on the data distribution. Other notable works are [119, 120], which also use techniques based on the kernel integral operator. These works assume a power-law eigenvalue decay to get power-law regression error bounds. Our results apply to more general kernels with an arbitrary eigenvalue decay and give a more refined bias-variance decomposition of error. Significantly, none of these works analyze interpolating solutions in the presence of noise.

5.2 Kernel regression

Our results are presented in terms of reproducing kernel Hilbert space regression (with traditional linear regression as a special case). We first introduce the analytical framework and then present our main results.

5.2.1 Kernel regression introduction

We first review the general theory of regression in reproducing kernel Hilbert spaces. A more thorough introduction to kernel theory can be found in many standard references, such as [160], [115], and Chapter 12 of [161].

Let X be a set, and let $\mathcal{H}$ be a real reproducing kernel Hilbert space over X with kernel $k\colon X \times X \to \mathbf{R}$. For $f \in \mathcal{H}$ and $x \in X$, we have $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$, where $k_x \coloneqq k(\cdot, x)$. Note that this implies that $k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}}$.

Suppose $f^* \in \mathcal{H}$, and we observe $y_i = f^*(x_i) + \xi_i$, $i = 1, \dots, n$, where $x_1, \dots, x_n \in X$ are sample points, and the ξ_i's represent noise or other measurement error. We use the kernel

ridge regression estimate

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \alpha \|f\|_{\mathcal{H}}^2,$$

where $\alpha \geq 0$ is a regularization term. When $\alpha \to 0$, we get the minimum-Hilbert-norm interpolator,

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \text{ s.t. } f(x_i) = y_i \; \forall i = 1, \ldots, n.$$

By the standard kernel regression formula, we have $\hat{f}(x) = \sum_{i=1}^{n} \hat{z}_i k(x, x_i)$ where the vector $\hat{z} = (\alpha I_n + K)^{-1} y$, and K is the kernel Gram matrix with $K_{ij} = \langle k_{x_i}, k_{x_j} \rangle_{\mathcal{H}} = k(x_i, x_j)$. We denote by $\mathcal{A} \colon \mathcal{H} \to \mathbf{R}^n$ the *sampling operator*, which is defined by $(\mathcal{A}(f))_i = f(x_i) = \langle f, k_{x_i} \rangle_{\mathcal{H}}$. The adjoint of the sampling operator is given by $\mathcal{A}^*(z) = \sum_{i=1}^{n} z_i k_{x_i}$ for all $z \in \mathbf{R}^n$. Then the Gram matrix is $K = \mathcal{A}\mathcal{A}^*$, and we can write the kernel regression estimate in terms of the standard ridge regression formulas:

$$\hat{f} = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} y = (\alpha \mathcal{I}_{\mathcal{H}} + \mathcal{A}^*\mathcal{A})^{-1} \mathcal{A}^* y.$$

Note that, in general, the second expression is only well-defined if $\alpha > 0$ (since $\mathcal{A}$ is rank-deficient if $\mathcal{H}$ is infinite-dimensional).

We analyze two terms in this estimator. The first is the estimator that would be obtained in the absence of noise, which is given by

$$\hat{f}_0 \coloneqq \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A} f^* = (\alpha \mathcal{I}_{\mathcal{H}} + \mathcal{A}^*\mathcal{A})^{-1} \mathcal{A}^*\mathcal{A} f^*.$$

The second is the contribution to the estimate due to noise, which we denote by the function $\epsilon(x)$. We have

$$\epsilon = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1} \xi,$$

where $\xi = (\xi_1, \ldots, \xi_n)$. This leads to a standard decomposition in the error of the estimator

$\hat{f}$ in terms of its bias and variance.

To characterize the test error, we need a sampling model. Let μ be a probability measure on X. We then define the kernel integral operator $\mathcal{T}$ as

$$(\mathcal{T}(f))(x) = \int_X k(x,y)f(y)\, d\mu(y)$$

with respect to the measure μ. Under mild regularity/continuity conditions (see, e.g., [89] for a thorough analysis), we have the eigenvalue decomposition

$$\mathcal{T}(f) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle v_\ell, f\rangle_{L_2} v_\ell,$$

where $\{v_\ell\}_{\ell=1}^{\infty}$ is an orthonormal basis for $L_2(X,\mu)$, and $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$ are the eigenvalues of $\mathcal{T}$ arranged in decreasing order. Furthermore, we have

$$k(x,y) = \sum_{\ell=1}^{\infty} \lambda_\ell v_\ell(x) v_\ell(y).$$

We can handle the finite-dimensional case by setting $\lambda_\ell = 0$ for $\ell > d$, where $d = \dim(\mathcal{H})$ (furthermore, the standard linear regression case can be recovered with $X = \mathbf{R}^d$ and $k(x,y) = \langle x, y\rangle_{\ell_2}$). Note that in order to interpolate an arbitrary set of samples, we need the dimension d to be at least the number of samples n (otherwise, the linear system is overdetermined).

We will also use the following well-known fact throughout our analysis: for any $f, g \in L_2$, we have

$$\langle f, g\rangle_{L_2} = \langle \mathcal{T}^{1/2} f, \mathcal{T}^{1/2} g\rangle_{\mathcal{H}}.$$

Hence, $\mathcal{T}^{1/2}$ is an isometry from L_2 to $\mathcal{H}$. Note that this implies that for every $f \in \mathcal{H}$,

$$\|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\langle f, v_\ell\rangle_{L_2}^2}{\lambda_\ell}.$$

Intuitively, we expect that if f has small/bounded $\mathcal{H}$-norm, most of its energy is captured by components corresponding to relatively large eigenvalues. Therefore, it is feasible to recover an accurate (in L_2) estimate of f, even though f lies in an infinite-dimensional space.

We will assume $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mu$, i.e., the training examples are drawn from the same measure as the test example $x \sim \mu$. Since we are evaluating a regression task, we wish to bound the squared (excess) prediction loss $\mathbf{E}(\hat{f}(x) - f^*(x))^2 = \|\hat{f} - f^*\|_{L_2}^2$. We will provide non-asymptotic upper bounds on $\|\hat{f} - f^*\|_{L_2}^2$ as a function of the number of training examples n. We will also focus on understanding scenarios for which we obtain statistical consistency, i.e., $\|\hat{f} - f^*\|_{L_2}^2 \to 0$ as $n \to \infty$.

5.2.2 Main results for deterministic sample locations

To state our main results, we introduce some additional notation. Here and for the rest of this section, p will be a fixed integer that we can tune in our analysis. We divide the function space $L_2(X, \mu)$ into two parts: $G = \operatorname{span}\{v_1, \ldots, v_p\}$ denotes the space spanned by the first p eigenfunctions of $\mathcal{T}$, and $G^\perp$ denotes its orthogonal complement (in both L_2 and $\mathcal{H}$). Accordingly, we split our sampling operator into two parts: $\mathcal{A}_G = \mathcal{A}|_G$ and $\mathcal{R} = \mathcal{A}|_{G^\perp}$. Intuitively, if p is chosen such that $\lambda_{p+1}, \lambda_{p+2}, \ldots$ are relatively small, we expect G to contain most of the energy in any given function $f \in \mathcal{H}$. A key fact is that the Gram matrix can be decomposed as $\mathcal{A}\mathcal{A}^* = \mathcal{A}_G\mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*$. The dimension p is similar to (but more flexible than) the regularization-dependent effective dimension in [133, 134].

Since $p = \dim(G)$ is finite, we can recover a function in G from a finite number of samples. We state this quantitatively by analyzing the restricted sampling operator $\mathcal{A}_G$. To state concentration results on G in terms of the L_2 norm, we denote $\mathcal{C} = \mathcal{A}_G$, and we let $\mathcal{C}^* = \mathcal{T}_G^{-1}\mathcal{A}_G^*$ be its adjoint with respect to the L_2 inner product. Note that $\frac{1}{n}\,\mathbf{E}\,\mathcal{A}_G^*\mathcal{A}_G = \mathcal{T}_G$, where $\mathcal{T}_G = \mathcal{T}|_G$. Therefore,

$$\frac{1}{n}\,\mathbf{E}\,\mathcal{C}^*\mathcal{C} = \mathcal{I}_G,$$

where $\mathcal{I}_G$ is the identity operator on G. Provided that $n \gg p$, we expect $\frac{1}{n}\mathcal{C}^*\mathcal{C} \approx \mathcal{I}_G$. We

will analyze how closely this holds later; we first state *deterministic* results that depend on the error in this approximation.

The second key approximation regards the "remainder Gram matrix" $\mathcal{R}\mathcal{R}^*$. Previous interpolation literature has assumed that this matrix is approximately a multiple of the identity I_n (or is in some sense "well-conditioned"). We will again analyze how accurately this holds later, but for now, we will state our main results assuming that $\alpha I_n + \mathcal{R}\mathcal{R}^*$ is upper and lower bounded by multiples of the identity. There is no requirement that $\alpha \geq 0$; in principle, our framework applies to negative regularization [162], but we do not explore this aspect in detail.

Finally, we will assume, for simplicity and brevity, that $f^* \in G$ exactly. If this did not hold, there would be another term in the "bias" error bound whose size is directly proportional to the size of $\mathcal{P}_{G^\perp}(f^*)$, which in turn is negligible provided that $f^* \in \mathcal{H}$ (i.e., f^* has bounded $\mathcal{H}$-norm). Note that kernel methods run into fundamental approximation-theoretic limitations in the absence of a bounded-$\mathcal{H}$-norm assumption [163, 164, 165].

Theorem 11 (Bias)**.** *Suppose that*

1. $\alpha_L I_n \preceq \alpha I_n + \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$ *for some numbers* $\alpha_U \geq \alpha_L > 0$*, and*
2. $\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \leq c$ *for some* $c < 1$.

Let $\bar{\alpha} = \frac{2\alpha_U \alpha_L}{\alpha_U + \alpha_L}$ *be the harmonic mean of* α_U *and* α_L*. Then, for any* $f^* \in G$*, we have*

$$\left\|\hat{f}_0 - f^*\right\|_{L_2} \lesssim \min\left\{\sqrt{\lambda_1}, \frac{1}{1-c}\frac{\bar{\alpha}}{n\sqrt{\lambda_p}}, \frac{1}{1-c}\sqrt{\frac{\bar{\alpha}}{n}}\right\} \times \left(1 + \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\right)\|f^*\|_{\mathcal{H}}.$$

Theorem 12 (Variance)**.** *Suppose the conditions of Theorem 11 hold, and let* $\tilde{\alpha} = \frac{\alpha_U + \alpha_L}{2}$*. Furthermore, suppose the* ξ_i*'s are zero-mean and independent with variance bounded by* σ^2*.*

Then

$$\mathbf{E}_\xi \|\epsilon\|_{L_2}^2 \lesssim \sigma^2 \left(\frac{\alpha_U}{\alpha_L} + 1\right)^2 \left(\frac{p}{n} + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2}\right). \tag{5.1}$$

Section 5.2.4 contains simplified proof sketches of Theorems 11 and 12; we provide complete proofs in Section C.2 in the supplementary material. The reader should note that our proofs consist of relatively simple linear algebra. Compare this, for example, to [150] or [148], where the analysis depends delicately on the independence (or, in the latter case, even Gaussianity) of the features via rather complicated matrix manipulations.

We could also obtain a high-probability (with respect to ξ) bound on the variance (if, e.g., the ξ_i's are sub-Gaussian), but we omit this to preserve the clarity and simplicity of the result. We outline how one could do this in Section C.2.2.

5.2.3 Operator concentration results

We now state operator concentration results on three important quantities: (a) the deviation of the residual Gram matrix $\mathcal{R}\mathcal{R}^*$ from a multiple of the identity, (b) the quantity $\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})$ which appears in the variance bound, and (c) the deviation of $\frac{1}{n}\mathcal{C}^*\mathcal{C}$ from $\mathcal{I}_G$. All proofs are contained in Section C.3 in the supplementary material. We begin with our most general results that apply under minimal assumptions.

General residual concentration

Let $k^R(x,y) = \sum_{\ell>p} \lambda_\ell v_\ell(x) v_\ell(y)$ be the reproducing kernel restricted to $G^\perp$.

Lemma 11 (Generic residual Gram matrix)**.**

$$\mathbf{E}\|\mathcal{R}\mathcal{R}^* - (\mathrm{tr}\,\mathcal{T}_{G^\perp})I_n\|^2 \lesssim n^2\,\mathrm{tr}(\mathcal{T}_{G^\perp}^2) + \|k^R(\cdot,\cdot) - \mathrm{tr}\,\mathcal{T}_{G^\perp}\|_\infty^2,$$

where $\|k^R(\cdot,\cdot) - \mathrm{tr}\,\mathcal{T}_{G^\perp}\|_\infty = \sup_x\{|k^R(x,x) - \mathrm{tr}\,\mathcal{T}_{G^\perp}|\}$.

Note for this result to give $\alpha_L I_n \preceq \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$ where α_U/α_L is bounded, we need $\operatorname{tr} \mathcal{T}_{G^\perp} \gtrsim n\sqrt{\operatorname{tr}(\mathcal{T}_{G^\perp}^2)}$. Even when $\{\lambda_\ell\}_{\ell>p}$ are all equal (see Section 5.3.1), we need $\dim \mathcal{H} = d \gtrsim n^2$. While this may seem restrictive, it is not possible to do better without additional assumptions on the features. In Section C.4, we show that in the case of Fourier features, $\lambda_{\max}(\mathcal{R}\mathcal{R}^*)/\lambda_{\min}(\mathcal{R}\mathcal{R}^*) \gtrsim \frac{n^4}{\tau^2 d^2}$ with probability at least $1 - e^{-\tau}$, and thus $d \gtrsim n^2$ is necessary. This can be significantly relaxed when the features are independent, as shown in Section 5.2.3.

To bound the variance, we will use the following expectation throughout the rest of this chapter:

Lemma 12 (Generic trace bound on $\mathcal{R}^*\mathcal{R}$)**.**

$$\mathbf{E} \operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R}) = n \operatorname{tr}(\mathcal{T}_{G^\perp}^2) = n \sum_{\ell>p} \lambda_\ell^2.$$

Note that Lemma 12, Theorem 12, and the approximate identity $\mathcal{R}\mathcal{R}^* \approx (\operatorname{tr} \mathcal{T}_{G^\perp}) I_n$ combine to bound the variance error as $\|\epsilon\|_{L_2}^2 \lesssim \frac{p}{n} + n\Big(\sum_{\ell>p} \lambda_\ell^2\Big) / \Big(\sum_{\ell>p} \lambda_\ell\Big)^2$. This is identical to the bound provided in [150].

Bounded orthonormal system

Our results show that harmless interpolation can occur in much more general settings than independent and/or sub-Gaussian features. An important class of features that are not independent or sub-Gaussian is a *bounded orthonormal system* (BOS).

On the subspace G defined before, the basis $v_1, \ldots, v_p$ is a BOS if it is an orthonormal basis in L_2 (as we have already assumed) and, further, we have

$$\sum_{\ell=1}^{p} v_\ell^2(x) \le Cp$$

μ-almost surely in x for some constant $C \ge 1$. Equivalently, for all $f \in G$, $\|f\|_\infty^2 \le$

$Cp\|f\|_{L_2}^2$.

This assumption is satisfied by many popular choices of features including sinusoids (see Section 5.4), Chebyshev polynomials, and the standard Euclidean basis on $\mathbf{R}^d$. One can also often show that kernel eigenfunctions satisfy this property, such as when the data lie on a low-dimensional manifold [83].

It is easy to derive concentration inequalities for bounded orthonormal systems via matrix/operator concentrations results for sums of bounded independent random matrices (e.g., [49]—see our supplementary material for details). A bound that is useful for our purposes is the following:

Lemma 13 (BOS sampling operator on G)**.** *If G is spanned by a bounded orthonormal system with constant C, then, for $t > 0$, with probability at least $1 - e^{-t}$,*

$$\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \lesssim \sqrt{\frac{Cp(t + \log p)}{n}} + \frac{Cp(t + \log p)}{n}.$$

Thus if $n \gtrsim Cp \log p$, we can have, say, $\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \leq 1/4$ (or any other small constant) with high probability.

In general, the $Cp \log p$ sample complexity is optimal under the BOS assumption. As a simple example, consider the following basis $\{v_1, \ldots, v_p\}$ on $\mathbf{R}^p$ (written as functions on $\{1, \ldots, p\}$): for uniquely determined constants c_1 and c_2, set the measure to be $\mu(\{j\}) = \frac{1}{Cp}$ for $j < p$ and $\mu(\{p\}) = c_1$, and set $v_\ell = \sqrt{Cp}\delta_\ell$ for $\ell < p$ and $v_p = c_2\delta_p$. One can easily verify by a coupon collector argument that we need $O(Cp \log p)$ samples from μ merely to sample every coordinate at least once.

Independent features

To compare to prior work, we list independent-feature concentration results that can be plugged into our Theorem 11. Suppose that for $x \sim \mu$, the features $\{v_\ell(x)\}$ are independent random variables. The key benefit this gives us is that we can now write the residual Gram

matrix $\mathcal{R}\mathcal{R}^*$ as a sum of *independent* random rank-1 matrices. To see this, define the vectors

$$w_\ell = (v_\ell(x_1), v_\ell(x_2), \dots, v_\ell(x_n)) \in \mathbf{R}^n.$$

We have already been assuming that the entries of each w_ℓ are independent (since they only depend on the independent variables x_i), but an independent features assumption implies that the entire set of random vectors $\{w_\ell\}_{\ell \geq 1}$ is independent. We can then write

$$\mathcal{R}\mathcal{R}^* = \sum_{\ell > p} \lambda_\ell w_\ell \otimes w_\ell.$$

We state a formal result for sub-Gaussian independent features. We expect similar results hold for much weaker tail conditions.

Lemma 14 (Independent features residual Gram matrix)**.** *Suppose the features $\{v_\ell(x)\}_{\ell \geq 1}$ are zero-mean, independent, and sub-Gaussian. Then, for $t > 0$, with probability at least $1 - e^{-t}$,*

$$\|\mathcal{R}\mathcal{R}^* - (\operatorname{tr} \mathcal{T}_{G^\perp}) I_n\| \lesssim \sqrt{(n+t) \operatorname{tr}(\mathcal{T}_{G^\perp}^2)} + (n+t)\lambda_{p+1}.$$

The zero-mean assumption is for simplicity and can easily be relaxed at the cost of a more complicated theorem statement. Note that this is stronger than Lemma 11 in two ways: first, the bound holds with exponentially high probability as opposed to being merely in expectation. Second, we have effectively replaced the n^2 in Lemma 11 by n, greatly reducing the amount of overparametrization we need.

Note for this result to give $\alpha_L I_n \preceq \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$ where α_U / α_L is bounded, we need

$$n \lesssim \frac{\operatorname{tr} \mathcal{T}_{G^\perp}}{\lambda_{p+1}} = \frac{1}{\lambda_{p+1}} \sum_{\ell > p} \lambda_\ell$$

(this also gives us $\sqrt{n \operatorname{tr}(\mathcal{T}_{G^\perp}^2)} \lesssim \operatorname{tr} \mathcal{T}_{G^\perp}$ by Cauchy-Schwartz). This is identical to the requirement that $r_{k^*}(\Sigma) \geq bn$ in [150].

We can also obtain slightly improved results (vs. the BOS assumption) for concentration of $\mathcal{C}^*\mathcal{C}$:

Lemma 15 (Sampling operator on G under independent features)**.** *With probability at least* $1 - e^{-t}$*,*

$$\left\| \frac{1}{n}\mathcal{C}^*\mathcal{C} - \mathcal{I}_G \right\|_{L_2} \lesssim \sqrt{\frac{p+t}{n}} + \frac{p+t}{n}.$$

Thus we only require $n \gtrsim p$ to obtain the required concentration. For a proof, see, for example, [53, Section 4.6].

5.2.4 Informal proof sketch (deterministic)

Here we outline the basic proof structure of Theorems 11 and 12. We will perform the analysis as though $\alpha I_n + \mathcal{R}\mathcal{R}^* = \bar{\alpha} I_n$ and $\mathcal{C}^*\mathcal{C} = n\mathcal{I}_G$ (equivalently, $\mathcal{A}_G^*\mathcal{A} = n\mathcal{T}_G$), and we will write "$\approx$" where we make these substitutions. The main additional steps we need are to quantify the error due to these approximations.

Bias term (from signal)

Note that since $f^* \in G$, we have

$$\begin{aligned}
\hat{f}_0 &= \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}_G f^* \\
&\approx \mathcal{A}^*(\bar{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G f^* \\
&= \mathcal{A}^*\mathcal{A}_G(\bar{\alpha}\mathcal{I}_G + \mathcal{A}_G^*\mathcal{A}_G)^{-1} f^* \\
&\approx \begin{bmatrix} \mathcal{A}_G^* \\ \mathcal{R}^* \end{bmatrix} \mathcal{A}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1} f^*.
\end{aligned}$$

From this we obtain

$$\begin{aligned}\mathcal{P}_G(\hat{f}_0) &= \mathcal{A}_G^*\mathcal{A}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}f^* \\ &\approx n\mathcal{T}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}f^* \\ &= \overline{\mathcal{S}}f^*,\end{aligned}$$

where $\overline{\mathcal{S}} \coloneqq n\mathcal{T}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}$ is the idealized "survival" operator, representing the extent to which the original signal f^* is preserved. We then have $f^* - \mathcal{P}_G(\hat{f}_0) \approx \overline{\mathcal{B}}f^*$, where $\overline{\mathcal{B}} \coloneqq \mathcal{I}_G - \overline{\mathcal{S}} = \bar{\alpha}(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}$ is the idealized "bias" operator. One can verify that

$$\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \lesssim \min\left\{\sqrt{\lambda_1}, \frac{\bar{\alpha}}{n\sqrt{\lambda_p}}, \sqrt{\frac{\bar{\alpha}}{n}}\right\}.$$

This bounds $\|\mathcal{P}_G(\hat{f}_0) - f^*\|_{L_2}$ for Theorem 11; the formal theorem has an extra factor of $1/(1-c)$ which comes from the approximation errors (recall that c is determined by how accurate our idealizing approximation are—see the statement of Theorem 11 for the precise definition).

Next, note that $\mathcal{P}_{G^\perp}(\hat{f}_0) \approx \mathcal{R}^*\mathcal{C}\frac{\overline{\mathcal{B}}}{\bar{\alpha}}f^*$, where we have substituted $\mathcal{C}$ for $\mathcal{A}_G$. Because

$$\|\mathcal{R}^*\|_{\ell_2\to L_2} \le \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*\|_{\ell_2\to\mathcal{H}} \lesssim \sqrt{\lambda_{p+1}\bar{\alpha}},$$

we have

$$\begin{aligned}\left\|\mathcal{R}^*\mathcal{C}\frac{\overline{\mathcal{B}}}{\bar{\alpha}}\right\|_{\mathcal{H}\to L_2} &\le \|\mathcal{R}^*\|_{\ell_2\to L_2}\|\mathcal{C}\|_{L_2\to\ell_2}\frac{\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2}}{\bar{\alpha}} \\ &\lesssim \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2},\end{aligned}$$

which allows us to bound $\|\mathcal{P}_{G^\perp}(\hat{f}_0)\|_{L_2}$.

Variance term (from noise)

Making similar approximations as above (with $\tilde{\alpha}$ instead of $\bar{\alpha}$—the distinction comes from the approximation arguments we use in the formal proof), we have

$$\begin{aligned}
\epsilon &= \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}*)^{-1}\xi \\
&\approx \begin{bmatrix} \mathcal{A}_G^* \\ \mathcal{R}^* \end{bmatrix} (\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\xi \\
&= \begin{bmatrix} (\tilde{\alpha}\mathcal{I}_G + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{A}_G^*\xi \\ \mathcal{R}^*(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\xi \end{bmatrix} \\
&\approx \begin{bmatrix} (\tilde{\alpha}\mathcal{T}_G^{-1} + n\mathcal{I}_G)^{-1}\mathcal{C}^*\xi \\ \mathcal{R}^*(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\xi \end{bmatrix}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{E}_\xi\|\epsilon\|_{L_2}^2 &\approx \sigma^2 \left\| \begin{bmatrix} (\tilde{\alpha}\mathcal{T}_G^{-1} + n\mathcal{I}_G)^{-1}\mathcal{C}^* \\ \mathcal{R}^*(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1} \end{bmatrix} \right\|_{HS,\ell_2\to L_2}^2 \\
&\lesssim \sigma^2 \left(\frac{1}{n^2}\operatorname{tr}_{L_2}(\mathcal{C}^*\mathcal{C}) + \frac{1}{\tilde{\alpha}^2}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R}) \right) \\
&\approx \sigma^2 \left(\frac{1}{n^2}\operatorname{tr}_{L_2}(n\mathcal{I}_G) + \frac{1}{\tilde{\alpha}^2}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R}) \right) \\
&= \sigma^2 \left(\frac{p}{n} + \frac{1}{\tilde{\alpha}^2}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R}) \right).
\end{aligned}$$

The factor of α_U/α_L comes from the approximation arguments.

5.3 Kernel classification

We now consider the case of classification, in which the observation y is a (noisy) binary label in $\{-1, 1\}$ with a distribution depending on x. Our approach is to perform ordinary linear/kernel regression on the binary labels y_i with the squared loss function. Although

this seems counter-intuitive, recent results (e.g., [166]) have shown that training with the squared-loss is highly competitive with the more common cross-entropy loss function in practice. Separately, recent results have also shown that regression on binary labels is, in some interesting overparametrized cases, equivalent to the maximum-margin SVM (e.g., [148, 167]). Inspired by these findings, we study the minimum-ℓ_2-norm interpolator of the binary labels $\{y_i\}_{i=1}^n$ and its ensuing *classification* error.

Through the lens of regression, our target function f^* is now replaced by

$$\eta^*(x) \coloneqq \mathbf{E}(y \mid x) = 2\,\mathrm{P}(y = 1 \mid x) - 1.$$

The label noise is $\xi = y - \eta^*(x)$. Note that $\mathbf{E}[\xi|x] = 0$ by definition, and $\mathrm{var}(\xi \mid x) = 1 - (\eta^*)^2(x)$. Our assumption on the label noise model is that $\eta^* \in G$.

The regression procedure yields an estimator $\hat{\eta}$ of η^*. Then, our classification rule is given by $\hat{y} = \mathrm{sign}(\hat{\eta})$. Given a probability distribution μ over x, the *excess risk* of the classification rule with respect to the Bayes-optimal classifier is given by

$$\mathcal{E} \coloneqq \mathrm{P}(\hat{y} \neq y) - \mathrm{P}(y \neq \mathrm{sign}(\eta^*)).$$

Standard calculations (see [154]) give

$$\mathcal{E} = \int |\eta^*(x)| \mathbf{1}_{\{\mathrm{sign}(\hat{\eta}(x)) \neq \mathrm{sign}(\eta^*(x))\}} d\mu(x).$$

Thus the excess risk is the average of the sign error of $\hat{\eta}$ versus η^* modulated by how distinguishable the two classes are (which is represented by $|\eta^*|$).

To bound $\mathcal{E}$, we decompose our estimate $\hat{\eta}$ as

$$\hat{\eta} = s\eta^* + \hat{\eta}_r, \tag{5.2}$$

where s is a parameter that we can tune in our analysis, and $\hat{\eta}_r$ is the residual. If $s > 0$, we have

$$\{\text{sign}(\hat{\eta}) \neq \text{sign}(\eta^*)\} \subseteq \{|\hat{\eta}_r| \geq s|\eta^*|\},$$

so

$$\mathcal{E} \leq \frac{1}{s} \int |\hat{\eta}_\perp(x)| \, d\mu(x) = \frac{\|\hat{\eta}_r\|_{L_1}}{s} \leq \frac{\|\hat{\eta}_r\|_{L_2}}{s}, \tag{5.3}$$

where the norm inequality is due to the fact that μ is a probability measure. For reasons that will shortly become clear, we will call s the *survival factor* and $\hat{\eta}_r$ the *residual*.

A first possible choice for the quantities in (Equation 5.2) is $s = 1$ and $\hat{\eta}_r = \hat{\eta} - \eta^*$. This choice yields $\mathcal{E} \leq \|\hat{\eta} - \eta\|_{L_1}$; therefore, small regression error implies small excess classification risk. However, we are interested in cases in which the regression error is not small but the classification error is. To use the bound (Equation 5.3), we would need to show that we can have the ratio $\|\hat{\eta}_r\|_{L_2}/s$ be very small with a different choice of $s \ll 1$.

We now show how this can work. We recall the idealized "survival" and "bias" operators $\overline{\mathcal{S}}$ and $\overline{\mathcal{B}}$ from Section 5.2.4. Note that to bound the regression error we show that $\hat{\eta} \approx \overline{\mathcal{S}}(\eta^*)$ and that $\|\eta^* - \overline{\mathcal{S}}\eta^*\|_{L_2} = \|\overline{\mathcal{B}}\eta^*\|_{L_2}$ is small. For the classification problem, an interesting new possibility arises. As a simple example, suppose all the first p eigenvalues $\lambda_1, \ldots, \lambda_p$ are identically 1. Then $\overline{\mathcal{S}} = \frac{n}{\bar{\alpha}+n}\mathcal{I}_G$. If $\bar{\alpha} \ll n$, then the ideal bias $\overline{\mathcal{B}} = \frac{\bar{\alpha}}{\bar{\alpha}+n}\mathcal{I}_G$ will be small. However, what if $\alpha \gtrsim n$, in which case the bias is not small? We cannot get small regression error, but for classification, we can apply (Equation 5.3) while choosing $s = \frac{n}{\bar{\alpha}+n}$. Then, as long as

$$\|\hat{\eta} - \overline{\mathcal{S}}\eta^*\|_{L_2} \ll \frac{n}{\bar{\alpha} + n},$$

we will have small excess classification risk. In Section 5.3.1, we use this observation to provide sufficient conditions for classification consistency, and demonstrate that these conditions are significantly weaker than the ones needed to be regression-consistent. This apporach is qualitatively similar to that of [148], which provides a slightly sharper bound but relies on a special form of η^* and Gaussianity of the features. Their techniques do not

easily extend to a more general setting.

To combine this framework with our previous interpolation results, note that, under our new notation, $\hat{\eta} = \hat{\eta}_0 + \epsilon$, where $\hat{\eta}_0 = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*$ and $\epsilon = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi$. We will show that $\hat{\eta}_0 \approx \overline{\mathcal{S}}\eta^*$ and ϵ is small. For the first objective, we present here a more refined version of Theorem 11 that bounds the error to $\overline{\mathcal{S}}\eta^*$ rather than η^* itself. This will be used to characterize the classification error in Section 5.3.1.

Lemma 16 (More refined bias estimate)**.** *Under the conditions of Theorem 11 (assuming c is bounded away from 1 so that $(1-c)^{-1}$ is subsumed into the constants),*

$$\|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2} \lesssim \left(c + \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\right) \times \min\left\{\lambda_1, \frac{\bar{\alpha}}{n\sqrt{\lambda_p}}, \sqrt{\frac{\bar{\alpha}}{n}}\right\}\|\eta^*\|_{\mathcal{H}}.$$

The proof of Lemma 16 is an easy modification of the proof of Theorem 11 (see Section C.2.1).

5.3.1 Bi-level ensemble asymptotic analysis

We now examine the implications of this refined classification analysis in a bounded orthonormal system (BOS). In particular, suppose that the eigenfunctions are all bounded (e.g., a Fourier series for periodic functions on an interval). For the eigenvalues, we consider the bi-level ensemble as defined in [148] with non-negative parameters n, β, q, r (where $\beta > 1$ and $r < 1$). This ensemble contains $d = n^\beta$ features, of which $p = n^r$ have "large" eigenvalues, and the remaining $d - p$ eigenvalues are small and their relative magnitude depends on the parameter q. Specifically, we set

$$\lambda_\ell = \begin{cases} 1, & 1 \leq \ell \leq p \\ n^{-(\beta - r - q)}, & p < \ell \leq d. \end{cases} \tag{5.4}$$

We require $q < \beta - r$ to ensure that the "small" eigenvalues are actually smaller than 1.

Corollary 3. *Consider the bi-level ensemble with parameters n, β, q, r, and suppose that the eigenfunctions are all bounded by an absolute constant. Further, suppose that $\beta > 2$ and $r < 1$, and $\eta^* \in G$. Then we obtain the following asymptotic results as $n \to \infty$:*

1. *If $q < 1 - r$, as $n \to \infty$, $\|\hat{\eta} - \eta^*\|_{L_2} \to 0$ in probability, and therefore both regression and classification are consistent.*

2. *If $q > 1 - r$, $\|\hat{\eta}\|_{L_2} \to 0$ in probability, and therefore regression is inconsistent for nonzero η^*.*

3. *If $q < \frac{3}{2}(1 - r)$ and $\beta > 2(r + q)$, excess classification risk $\mathcal{E} \to 0$ in probability, that is, classification is consistent.*

Corollary 3 is proved in Section C.5. Note that we have an asymptotic separation between classification and regression when $1 - r < q < \frac{3}{2}(1 - r)$. This is comparable to the results of [148], which allow slightly larger q and smaller β but require much stronger feature assumptions.

We use the bi-level ensemble model in (Equation 5.4) for simplicity; however, we can obtain non-asymptotic bounds on the classification risk under more general assumptions. For a fixed value of $p := n^r$, our analysis allows the tail eigenvalues corresponding to indices $p < \ell \leq d$ to be non-uniform. The requirement that the top p eigenvalues are the same is somewhat more stringent; in general, when the eigenvalues are different, the survival operator $\overline{\mathcal{S}}$ is not a multiple of the identity. This could lead to qualitatively different behavior, as now $\hat{\eta}$ may be distorted from η^* due to differences in the eigenvalues of $\mathcal{T}_G$. This problem disappears in the case that either (a) η^* is proportional to a single eigenfunction or (b) the first p eigenvalues of $\mathcal{T}$ are identical (both of which hold in [148]). We analyze further the extent to which we can bound the classification risk when *neither* of these assumptions holds in Section C.6. Our analysis method requires λ_p to be close to λ_1 to obtain significant gains for classification over regression.

5.4 Numerical experiments

We now perform numerical experiments to demonstrate how the parameters β, r, and q of the bi-level ensemble model affect regression and classification performance. We consider the case of Fourier features $v_\ell(x) = e^{\mathrm{j}2\pi\ell x}$ for $\ell = -d, \dots, d$ over $x \in [0,1]$ with the uniform sampling measure, and the bi-level ensemble as defined in (Equation 5.4). The corresponding kernel function is

$$\begin{aligned} k(x,y) &= \sum_{\ell=-d}^{d} \lambda_\ell v_\ell(x)\overline{v_\ell(y)} \\ &= (1 - n^{-(\beta-r-q)})D_p(x-y) \\ &\quad + n^{-(\beta-r-q)}D_d(x-y), \end{aligned}$$

where $D_m(t) = \frac{\sin[(2m+1)\pi t]}{\sin(\pi t)}$ is the Dirichlet sinc function. We consider three cases for the bi-level ensemble parameters: $(\beta, r, q) = (2.6, 0.3, 0.3)$, $(2.6, 1/3, 5/6)$, and $(2.6, 0.8, 0.45)$. We sweep over several values of n between 10 and 3162. For each n, we generate an $\eta^* \in \text{span}\{v_\ell\}_{\ell=-p}^{p}$, scaled such that $\max_{x\in[0,1]} |\eta^*(x)| = 1$.

We first attempt to reconstruct $\eta^*(x)$ from noisy samples $y_i^{\text{reg}} = \eta^*(x_i) + \xi_i$ for $i = 1, \dots, n$ where ξ_i are i.i.d. $\mathcal{N}(0,1)$. We use the kernel ridge regression estimator $\widehat{\eta^*}^{\text{reg}} = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1}y^{\text{reg}}$ with a regularization parameter of $\alpha = 10^{-3}$. We then measure the relative L_2 error of the estimate, i.e., $\mathcal{E}^{\text{reg}} = \|\eta^* - \widehat{\eta}^{\text{reg}}\|_{L_2}^2 / \|\eta^*\|_{L_2}^2$.

We also attempt to reconstruct $\eta^*(x)$ from binary observations $y_i^{\text{class}} = +1$ with probability $(1+\eta^*(x_i))/2$ and -1 with probability $(1-\eta^*(x_i))/2$ for $i = 1, \dots, n$. We use the estimator $\widehat{\eta}^{\text{class}} = \mathcal{A}^*(\alpha I_n + \mathcal{A}\mathcal{A}^*)^{-1}y^{\text{class}}$ with a regularization parameter of $\alpha = 10^{-3}$. We then measure the relative excess risk, i.e.,

$$\mathcal{E}^{\text{class}} = \frac{\int |\eta^*(x)| \mathbf{1}_{\{\text{sign}(\hat{\eta}^{\text{class}}(x)) \neq \text{sign}(\eta^*(x))\}}\, dx}{\int |\eta^*(x)|\, dx}.$$

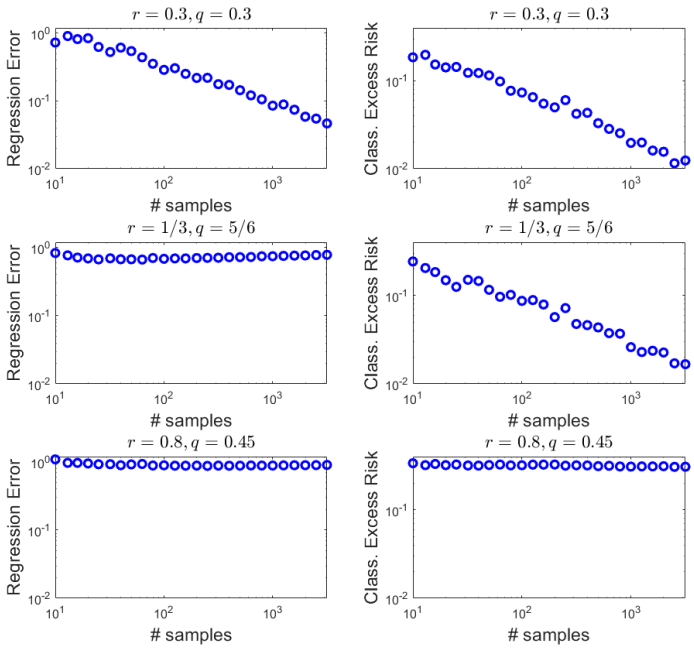


Figure 5.2: Relative L_2 errors versus n and the relative classification excess risks versus n for each of the three sets of bi-level ensemble parameters (averaged over 100 trials).

The above procedure is repeated over 100 trials. In Figure 5.2, we plot the relative L_2 error (averaged over 100 trials) versus n and the relative excess risk (averaged over 100 trials) versus n for each of the three sets of values for β, r, q. In the first case where $\beta = 2.6$, $r = 0.3$, and $q = 0.3$, we have $r + q < 1$ and both $\mathcal{E}^{\text{reg}}$ and $\mathcal{E}^{\text{class}}$ decrease as n increases. In the second case where $\beta = 2.6$, $r = 1/3$, and $q = 5/6$, we have $1 - r < q < \frac{3}{2}(1 - r)$ and $\beta > 2r + 2q$, and $\mathcal{E}^{\text{class}}$ decreases as n increases, but $\mathcal{E}^{\text{reg}}$ does not decrease as n increases. In the third case where $\beta = 2.6$, $r = 0.8$, and $q = 0.45$, we have that $r + q > 1$, and $q > \frac{3}{2}(1 - r)$, and both $\mathcal{E}^{\text{reg}}$ and $\mathcal{E}^{\text{class}}$ do not decrease as n increases.

5.5 Discussion

In this work we showed that under minimal assumptions on the data and feature map (a) harmless interpolation of noise in data is possible, and (b) we can be classification-consistent in high-dimensional regimes where we are not regression-consistent. Important future directions include considering more general function models (e.g., any $f^* \in \mathcal{H}$ or even $f^* \notin \mathcal{H}$), better understanding the implications of distortion among the top eigenfunctions in classifica-

tion error, and improving the non-asymptotic rates for classification risk from Section 5.3.1. Another intriguing question is whether there is an equivalence between interpolating binary labels and the max-margin SVM (as shown in [148, 167]) in the more general settings considered in this work. Finally, it would be very interesting to study whether our upper bounds (particularly for classification) can be matched by non-asymptotic lower bounds.

Appendices

APPENDIX A

PHASE RETRIEVAL AND PCA ANALYSIS

A.1 Detailed analysis of mixed norm

In this section, we explore several important properties of the mixed norm $\|\cdot\|_{*,s}$.

First, we show that matrices with small mixed norm can be written as a convex combination of sparse rank-1 matrices.

Lemma 17. *For any matrix A, we can write $A = \sum a_i u_i \otimes v_i$, where each u_i and v_i has unit ℓ_2 norm and is s-sparse, and $\sum |a_i| \leq 2\|A\|_{*,s}$.*

Proof. Because $\|\cdot\|_{*,s}$ is defined as an atomic norm over rank-1 atoms, it suffices to prove the result for rank-1 A. Therefore, we will show that any rank-1 matrix $x \otimes y$ can be written as $x \otimes y = \sum u_i \otimes v_i$, where each u_i and v_i is s-sparse, and $\sum \|u_i\|_2 \|v_i\|_2 \leq 2\theta_s(x, y)$.

Indeed, a standard result from sparsity theory (see, e.g., Exercise 10.3.7 in [53]) says that any vector z can be written as $z = \sum z_i$, where each z_i is s-sparse, and $\sum \|z_i\|_2 \leq \|z\|_2 + \frac{1}{\sqrt{s}}\|z\|_1$. Applying this to both x and y, we have

$$x \otimes y = \left(\sum_i x_i\right)\left(\sum_j y_j\right) = \sum_{i,j} x_i \otimes y_j,$$

where each x_i and y_j is s-sparse, and

$$\begin{aligned}\sum_{i,j} \|x_i\|_2 \|y_j\|_2 &= \left(\sum_i \|x_i\|_2\right)\left(\sum_j \|y_j\|_2\right) \\ &\leq \left(\|x\|_2 + \frac{\|x\|_1}{\sqrt{s}}\right)\left(\|y\|_2 + \frac{\|y\|_1}{\sqrt{s}}\right) \\ &\leq 2\theta_s(x, y).\end{aligned}$$

□

This result is useful because it immediately implies the following:

Corollary 4. *For any matrix Z,*

$$\sup_{\|A\|_{*,s}\le 1} \langle Z, A\rangle_{\mathrm{HS}} \le 2 \sup_{\substack{\|u\|_2,\|v\|_2\le 1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle.$$

Next, we consider the subgradients of $\|\cdot\|_{*,s}$ on $\mathbf{R}^{p\times p}$ at a point $B = \beta\otimes\beta$. Let $I \subset \{1,\dots,p\}$ denote the indices for which the entries of β are nonzero. With some abuse of notation, we also write I as the subspace of $\mathbf{R}^{p\times p}$ whose matrices are zero except at entries $(i,j)\in I\times I$. We also denote $T = \{x\otimes\beta + \beta\otimes y : x, y\in\mathbf{R}^p\}$.

According to [80, Proposition 1], a matrix $W \in \partial\|B\|_{*,s}$ if the following two properties hold:

1. $\langle W\beta, \beta\rangle = \theta_s(\beta,\beta)$, and
2. $\langle Wu, v\rangle \le \theta_s(u,v)$ for all $u, v\in\mathbf{R}^p$.

The matrix $W_\beta \coloneqq \frac{\beta^{\otimes 2}}{\|\beta\|_2^2} + \frac{1}{s}(\operatorname{sign}\beta)^{\otimes 2}$ clearly satisfies these properties. However, as with the subgradients of the ordinary nuclear norm, a much broader set of matrices satisfies these properties:

Lemma 18. *Suppose β is s-sparse. Any matrix of the form $W = W_\beta + W_{\beta,\perp} \in \partial\|B\|_{*,s}$ if $W_{\beta,\perp}\in T^\perp\cap I^\perp$, and $\langle W_{\beta,\perp}u, v\rangle \le \frac{1}{2}\theta_s(u,v)$ for all $u, v\in\mathbf{R}^p$.*

Proof. Note that the first subgradient property above holds because it does for W_β and $W_{\beta,\perp}\in T^\perp$. To show that property 2 above holds, note that for any $u, v\in\mathbf{R}^p$,

$$\begin{aligned}
\langle W_\beta u, v\rangle &= \frac{\langle\beta,u\rangle\langle\beta,v\rangle}{\|\beta\|_2^2} + \frac{1}{s}\langle\operatorname{sign}\beta, u\rangle\langle\operatorname{sign}\beta, v\rangle\\
&\le \|\mathcal{P}_\beta u\|_2\|\mathcal{P}_\beta v\|_2 + \frac{1}{s}\|\mathcal{P}_I u\|_1\|\mathcal{P}_I v\|_1\\
&\le \frac{1}{2}\left(\|\mathcal{P}_\beta u\|_2^2 + \|\mathcal{P}_\beta v\|_2^2 + \frac{\|\mathcal{P}_I u\|_1^2}{s} + \frac{\|\mathcal{P}_I v\|_1^2}{s}\right),
\end{aligned}$$

where $\mathcal{P}_\beta$ denotes the projection onto the (1-dimensional) subspace spanned by β. Using the shorthand $\beta^\perp$ to denote the subspace of $\mathbf{R}^p$ orthogonal to β, note that

$$\mathcal{P}_{T^\perp \cap I^\perp}(u \otimes v) = \mathcal{P}_{I^\perp}(u) \otimes \mathcal{P}_{I^\perp}(v) + \mathcal{P}_{I \cap \beta^\perp}(u) \otimes \mathcal{P}_{I^\perp}(v) + \mathcal{P}_{I^\perp}(u) \otimes \mathcal{P}_{I \cap \beta^\perp}(v).$$

Then

$$\begin{aligned}
\langle W_{\beta,\perp} u, v \rangle &= \langle W_{\beta,\perp}, u \otimes v \rangle_{\mathrm{HS}} \\
&= \langle W_{\beta,\perp}, \mathcal{P}_{T^\perp \cap I^\perp}(u \otimes v) \rangle_{\mathrm{HS}} \\
&= \langle W_{\beta,\perp}, \mathcal{P}_{I^\perp}(u) \otimes \mathcal{P}_{I^\perp}(v) \rangle_{\mathrm{HS}} + \langle W_{\beta,\perp}, \mathcal{P}_{I \cap \beta^\perp}(u) \otimes \mathcal{P}_{I^\perp}(v) \rangle_{\mathrm{HS}} \\
&\quad + \langle W_{\beta,\perp}, \mathcal{P}_{I^\perp}(u) \otimes \mathcal{P}_{I \cap \beta^\perp}(v) \rangle_{\mathrm{HS}} \\
&\leq \frac{1}{2}\left[\theta_s(\mathcal{P}_{I^\perp}(u), \mathcal{P}_{I^\perp}(v)) + \theta_s(\mathcal{P}_{I \cap \beta^\perp}(u), \mathcal{P}_{I^\perp}(v)) + \theta_s(\mathcal{P}_{I^\perp}(u), \mathcal{P}_{I \cap \beta^\perp}(v))\right] \\
&= \frac{1}{2}\left[\|\mathcal{P}_{I^\perp}(u)\|_2^2 + \|\mathcal{P}_{I^\perp}(v)\|_2^2 + \frac{\|\mathcal{P}_{I^\perp}(u)\|_1^2}{s} + \frac{\|\mathcal{P}_{I^\perp}(v)\|_1^2}{s}\right] \\
&\quad + \frac{1}{4}\left[\|\mathcal{P}_{I \cap \beta^\perp}(u)\|_2^2 + \|\mathcal{P}_{I \cap \beta^\perp}(v)\|_2^2 + \frac{\|\mathcal{P}_{I \cap \beta^\perp}(u)\|_1^2}{s} + \frac{\|\mathcal{P}_{I \cap \beta^\perp}(v)\|_1^2}{s}\right] \\
&\leq \frac{1}{2}\Bigg[\|\mathcal{P}_{I^\perp}(u)\|_2^2 + \|\mathcal{P}_{I^\perp}(v)\|_2^2 + \frac{\|\mathcal{P}_{I^\perp}(u)\|_1^2}{s} + \frac{\|\mathcal{P}_{I^\perp}(v)\|_1^2}{s} \\
&\quad + \|\mathcal{P}_{I \cap \beta^\perp}(u)\|_2^2 + \|\mathcal{P}_{I \cap \beta^\perp}(v)\|_2^2\Bigg] \\
&= \frac{1}{2}\left[\|\mathcal{P}_{\beta^\perp}(u)\|_2^2 + \|\mathcal{P}_{\beta^\perp}(v)\|_2^2 + \frac{\|\mathcal{P}_{I^\perp}(u)\|_1^2}{s} + \frac{\|\mathcal{P}_{I^\perp}(v)\|_1^2}{s}\right],
\end{aligned}$$

where the last inequality follows from the fact that the ℓ_1 norm of an s-sparse vector is at most $\sqrt{s}$ times the ℓ_2 norm. Since

$$\|\mathcal{P}_I(u)\|_1^2 + \|\mathcal{P}_{I^\perp}(u)\|_1^2 \leq \left(\|\mathcal{P}_I(u)\|_1 + \|\mathcal{P}_{I^\perp}(u)\|_1\right)^2 = \|u\|_1^2,$$

we obtain

$$\langle (W_\beta + W_{\beta,\perp})u, v \rangle \leq \theta_s(u, v).$$

□

It is much easier to verify the following:

Lemma 19. *Every matrix of the form $W = W_\beta + W^\perp$, where either*

1. *$W^\perp \in I^\perp$ and $\|W^\perp\|_{\infty,\infty} \le 1/s$, i.e., $\langle W^\perp u, v\rangle \le \frac{1}{s}\|u\|_1\|v\|_1 = \frac{1}{s}\|u \otimes v\|_{1,1}$ for all $u, v \in \mathbf{R}^p$ or*

2. *$W^\perp \in T^\perp$ and $\|W^\perp\| \le 1$, i.e., $\langle W^\perp u, v\rangle \le \|u\|_2\|v\|_2$ for all $u, v \in \mathbf{R}^p$,*

satisfies $W \in \partial\|\beta\|_{*,s}$.

Proof. Since $W^\perp \perp \beta\otimes\beta$ for both choices of $W^\perp$, the first subgradient property ($\langle W\beta, \beta\rangle = \theta_s(\beta, \beta)$) clearly holds.

To prove case 1, let $W^\perp \in I^\perp$ and$\|W^\perp\|_{\infty,\infty} \le 1/s$. Then

$$\begin{aligned}
\langle Wu, v\rangle &= \langle W_\beta u, v\rangle + \langle W^\perp, u \otimes v\rangle_{\mathrm{HS}} \\
&\le \frac{\|u\|_2^2 + \|v\|_2^2}{2} + \frac{\|\mathcal{P}_I(u \otimes v)\|_{1,1}}{s} + \frac{\|\mathcal{P}_{I^\perp}(u \otimes v)\|_{1,1}}{s} \\
&= \frac{\|u\|_2^2 + \|v\|_2^2}{2} + \frac{1}{s}\|u \otimes v\|_{1,1} \\
&\le \frac{\|u\|_2^2 + \|v\|_2^2}{2} + \frac{\|u\|_1^2 + \|v\|_1^2}{2s} \\
&= \theta_s(u, v).
\end{aligned}$$

Similarly, for case 2, let $W^\perp \in T^\perp$ and $\|W^\perp\| \le 1$. Then

$$\begin{aligned}
\langle Wu, v\rangle &= \langle W_\beta u, v\rangle + \langle W^\perp, u \otimes v\rangle_{\mathrm{HS}} \\
&\le \frac{\|\mathcal{P}_\beta(u)\|_2^2 + \|\mathcal{P}_\beta(v)\|_2^2}{2} + \frac{\|u\|_1^2 + \|v\|_1^2}{2s} + \|\mathcal{P}_{\beta^\perp}(u)\|_2\|\mathcal{P}_{\beta^\perp}(v)\|_2 \\
&\le \frac{\|\mathcal{P}_\beta(u)\|_2^2 + \|\mathcal{P}_\beta(v)\|_2^2}{2} + \frac{\|u\|_1^2 + \|v\|_1^2}{2s} + \frac{\|\mathcal{P}_{\beta^\perp}(u)\|_2^2 + \|\mathcal{P}_{\beta^\perp}(v)\|_2^2}{2} \\
&= \theta_s(u, v).
\end{aligned}$$

□

A.2 An empirical process bound

We will use the following bound several times:

Lemma 20. *Let $(\gamma_1, \epsilon_1), \ldots, (\gamma_n, \epsilon_n)$ be i.i.d. copies of (γ, ϵ), where, for some σ^2, M, and η, we have, for all $u \in \mathbf{R}^p$, that $\epsilon\langle\gamma, u\rangle$ is zero-mean,*

$$\mathbf{E}(\epsilon\langle\gamma, u\rangle^2)^2 \leq \sigma^2\|u\|_2^2,$$

and

$$\|\epsilon\langle\gamma, u\rangle^2\|_\alpha \leq M\alpha^{\eta+1}\|u\|_2^2$$

for all $\alpha \geq 3$. Let $Z = \frac{1}{n}\sum_{i=1}^n \epsilon_i\gamma_i \otimes \gamma_i$. Then, for any integer $s \geq 1$, with probability at least $1 - e^{-s}(s/p)^s$,

$$\sup_{\substack{\|u\|_2=\|v\|_2=1 \\ \|u\|_0,\|v\|_0\leq s}} \langle Zu, v\rangle \lesssim \sigma\sqrt{\frac{s\log(ep/s)}{n}} + \frac{M}{n^{1-c}}\Big(s\log\frac{ep}{s}\Big)^{\eta+1},$$

where $c \approx \frac{1}{s\log(ep/s)}$.

Furthermore, taking $\epsilon = 1$ in the above moment bounds, the same empirical process bound holds for the matrix

$$Z = \frac{1}{n}\sum_{i=1}^n \gamma_i \otimes \gamma_i - \mathbf{E}\,\gamma \otimes \gamma.$$

Proof. We only prove the first bound, since the second can be proved similarly and also follows from the first by a symmetrization argument.

We first consider the random variable $\langle Zu, v\rangle$ for fixed unit-norm u and v. We have

$$\langle Zu, v\rangle = \frac{1}{n}\sum_{i=1}^n \epsilon_i\langle\gamma_i, u\rangle\langle\gamma_i, v\rangle.$$

This is the sum of independent, zero-mean random variables. Note that by the Cauchy-

Schwartz inequality,

$$\mathbf{E}(\epsilon_i \langle \gamma_i, u \rangle \langle \gamma_i, v \rangle)^2 \leq \sigma^2$$

and, for $\alpha \geq 3$,

$$\|\epsilon_i \langle \gamma_i, u \rangle \langle \gamma_i, v \rangle\|_\alpha \leq M \alpha^{\eta+1}.$$

Then, by [168, Theorem 3.1], for any $\delta > 0$, with probability at least $1 - \delta$,

$$\langle Zu, v \rangle \lesssim \sigma \sqrt{\frac{\log \delta^{-1}}{n}} + \frac{M \alpha^{\eta+1}}{n^{1-1/\alpha}} \delta^{-1/\alpha}.$$

We then use a covering argument similar to that in [169]. Let J_1 and J_2 be any two subspaces of s-sparse vectors in $\mathbf{R}^p$. The unit sphere S_{J_i} in J_i can be covered within a resolution of $1/4$ by at most 9^s points ([53, Corollary 4.2.13], for example). Let $\mathcal{N}_{J_1}, \mathcal{N}_{J_2}$ be optimal $1/4$-covering sets. For each $x \in S_{J_i}$, let $n_i(x)$ be the closest point in $\mathcal{N}_{J_i}$. Then

$$\begin{aligned} \sup_{\substack{u \in S_{J_1} \\ v \in S_{J_2}}} \langle Zu, v \rangle &= \sup_{\substack{u \in S_{J_1} \\ v \in S_{J_2}}} \langle Z n_1(u), n_2(v) \rangle + \langle Z(u - n_1(u)), v \rangle + \langle Z n_1(u), v - n_2(v) \rangle \\ &\leq \max_{\substack{u \in \mathcal{N}_{J_1} \\ v \in \mathcal{N}_{J_2}}} \langle Zu, v \rangle + \frac{1}{2} \sup_{\substack{u \in S_{J_1} \\ v \in S_{J_2}}} \langle Zu, v \rangle, \end{aligned}$$

so

$$\sup_{\substack{u \in S_{J_1} \\ v \in S_{J_2}}} \langle Zu, v \rangle \leq 2 \max_{\substack{u \in \mathcal{N}_{J_1} \\ v \in \mathcal{N}_{J_2}}} \langle Zu, v \rangle.$$

Let

$$\mathcal{N} = \bigcup_{s\text{-sparse } J_1,\, J_2} \mathcal{N}_{J_1} \times \mathcal{N}_{J_2}.$$

Clearly,

$$\sup_{\substack{\|u\|_2=\|v\|_2=1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle = \sup_{s\text{-sparse } J_1,\, J_2} \sup_{\substack{u\in S_{J_1}\\ v\in S_{J_2}}} \langle Zu, v\rangle$$

$$\le 2 \max_{(u,v)\in\mathcal{N}} \langle Zu, v\rangle.$$

There are $\binom{p}{s} \le \left(\frac{ep}{s}\right)^s$ s-sparse subspaces of $\mathbf{R}^p$, so $|\mathcal{N}| \le \left(9^s\left(\frac{ep}{s}\right)^s\right)^2$.

By a union bound and substituting δ above with $\delta/|\mathcal{N}|$, we then have, for any $\delta > 0$, with probability at least $1-\delta$,

$$\sup_{\substack{\|u\|_2=\|v\|_2=1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle \lesssim \sigma\sqrt{\frac{s\log(ep/s)}{n} + \frac{\log\delta^{-1}}{n}} + \frac{M\alpha^{\eta+1}}{n^{1-1/\alpha}}\left(\frac{Cp}{s}\right)^{2s/\alpha}\delta^{-1/\alpha}.$$

Taking $\delta = e^{-s}(s/p)^s$ and $\alpha \approx s\log\frac{Cp}{s}$, we get, with probability at least $1 - e^{-s}(s/p)^s$,

$$\sup_{\substack{\|u\|_2=\|v\|_2=1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle \lesssim \sigma\sqrt{\frac{s\log(ep/s)}{n}} + \frac{M}{n^{1-c}}\left(s\log\frac{ep}{s}\right)^{\eta+1}.$$

□

A.3 Proof of sparse phase retrieval error bound

Throughout this section, let T and I be as in Section A.1 for $\beta = \beta^*$. First, we need the following lower bound on the empirical L_2 loss:

Lemma 21. *Let $x_1, \ldots, x_n$ be i.i.d. copies of a random vector x satisfying Assumption 1, and let $X_i = x_i \otimes x_i$. Suppose*

$$n \gtrsim s\log\frac{ep}{s},$$

and let $C \ge 1$ be a fixed constant. With probability at least $1 - e^{-n}$, the following event

holds: For all $A \in \mathbf{R}^{p\times p}$ such that

$$\|\mathcal{P}_{I^\perp}(A)\|_{*,s} + \|\mathcal{P}_I(A)\|_* \le C\|A\|_F,$$

we have

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle_{\mathrm{HS}}^2 \gtrsim \|A\|_F^2,$$

where the constant in the lower bound depends on C.

Proof. If $X = x \otimes x$, by a straightforward calculation, for any $p \times p$ matrix A,

$$\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2 = \sum_{i\neq j} A_{ii}A_{jj}\,\mathbf{E}(x^i)^2(x^j)^2 + 2\sum_{i\neq j} A_{ij}^2\,\mathbf{E}(x^i)^2(x^j)^2 + \sum_i A_{ii}^2\,\mathbf{E}(x^i)^4.$$

Using the facts that $\mathbf{E}(x^i)^2 = 1$ for each i and x_i and x_j are independent when $i \neq j$, we have

$$\begin{aligned}\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2 &= \sum_{i,j} A_{ii}A_{jj} + 2\sum_{i\neq j} A_{ij}^2 + \sum_i A_{ii}^2(\mathbf{E}(x^i)^4 - 1)\\ &\ge (\operatorname{tr} A)^2 + \min\{2, \mathbf{E}(x^1)^4 - 1\}\|A\|_F^2\\ &\gtrsim \|A\|_F^2.\end{aligned}$$

The last inequality uses the assumption that $\mathbf{E}(x^1)^4 > 1$.

By the Hanson-Wright inequality for sub-Gaussian vectors [170], we have

$$\mathbf{E}(\langle X, A\rangle_{\mathrm{HS}}^2 - \mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2)^2 \lesssim \|A\|_F^4,$$

so $\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^4 \lesssim (\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2)^2$. By the Paley-Zygmund inequality, we then have, for some $c_1, c_2 > 0$,

$$\inf_{A\in\mathbf{R}^{p\times p}} \mathrm{P}(|\langle X, A\rangle_{\mathrm{HS}}| \ge c_1\|A\|_F) \ge c_2.$$

The remainder of the proof is a small-ball argument ([171]; see also [172] for an excellent

introduction).

Let

$$S = \{A \in \mathcal{S}_p : \|A\|_F = 1;\ \|\mathcal{P}_{I^\perp}(A)\|_{*,s} + \|\mathcal{P}_I(A)\|_* \le C\}.$$

We will prove that

$$\inf_{A\in S} \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle_{\mathrm{HS}}^2 \ge c$$

with high probability for some constant $c > 0$.

By [172, Proposition 5.1], for any $t > 0$, we have, with probability at least $1 - e^{-t^2/2}$,

$$\inf_{A\in S} \sqrt{\frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle_{\mathrm{HS}}^2} \gtrsim c_1 c_2 - 2\,\mathbf{E}\sup_{A\in S}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle X_i, A\rangle_{\mathrm{HS}}\right) - \frac{1}{\sqrt{n}} c_1 t,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables independent of everything else.

Writing $A = \mathcal{P}_I(A) + \mathcal{P}_{I^\perp}(A)$ and setting $Z = \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i$, we have, by Corollary 4,

$$\langle Z, A\rangle_{\mathrm{HS}} \lesssim (\|\mathcal{P}_{I^\perp}(A)\|_{*,s} + \|\mathcal{P}_I(A)\|_*) \sup_{\substack{\|u\|_2=\|v\|_2=1 \\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle.$$

Combining the assumptions on A with Lemma 20, we have

$$\mathbf{E}\sup_{A\in S}\langle Z, A\rangle_{\mathrm{HS}} \lesssim \sqrt{\frac{s\log(ep/s)}{n}} + \frac{s\log(ep/s)}{n}.$$

Choosing n large enough and $t \approx \sqrt{n}$ completes the proof.

□

Now, we are ready to prove the main theorem.

Proof of Theorem 5. Let $\widehat{B}$ be the solution to (Equation 3.6). Writing $F(B)$ as the objective function, the convexity of the optimization problem implies that

$$0 \le \langle \nabla F(\widehat{B}), B^* - \widehat{B}\rangle_{\mathrm{HS}} = \frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, \widehat{B}\rangle_{\mathrm{HS}})\langle X_i, \widehat{B} - B^*\rangle_{\mathrm{HS}} + \lambda\langle W_{\widehat{B}}, B^* - \widehat{B}\rangle_{\mathrm{HS}},$$

for any $W_{\widehat{B}} \in \partial\|\widehat{B}\|_{*,s}$. By the monotonicity of (sub)gradients of convex functions, we have that, for any $W \in \partial\|B^*\|_{*,s}$, $\langle W - W_{\widehat{B}}, B^* - \widehat{B}\rangle_{\mathrm{HS}} \geq 0$, and therefore

$$0 \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle X_i, \widehat{B}\rangle_{\mathrm{HS}})\langle X_i, \widehat{B} - B^*\rangle_{\mathrm{HS}} + \lambda\langle W, B^* - \widehat{B}\rangle_{\mathrm{HS}}.$$

Let $H = \widehat{B} - B^*$. Using the fact that $(y_i - \langle X_i, \widehat{B}\rangle_{\mathrm{HS}})\langle X_i, \widehat{B} - B^*\rangle_{\mathrm{HS}} = \xi_i\langle X_i, H\rangle_{\mathrm{HS}} - \langle X_i, H\rangle_{\mathrm{HS}}^2$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i, H\rangle_{\mathrm{HS}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X, H\rangle_{\mathrm{HS}} + \lambda\langle W, -H\rangle_{\mathrm{HS}}.$$

From Lemmas 18 and 19 and the convexity of subgradients, we can take

$$W = W_1 + W_2 + W_3 + W_4,$$

where $W_1 = \frac{\beta^* \otimes \beta^*}{\|\beta^*\|_2^2} + \frac{(\operatorname{sign}\beta^*)^{\otimes 2}}{s}$, W_2 satisfies $\langle W_2, H\rangle_{\mathrm{HS}} = \frac{1}{4}\|\mathcal{P}_{T^\perp \cap I^\perp}(H)\|_{*,s}$, W_3 satisfies $\langle W_3, H\rangle_{\mathrm{HS}} = \frac{1}{4s}\|\mathcal{P}_{I^\perp}\|_{1,1}$, and W_4 satisfies $\langle W_3, H\rangle_{\mathrm{HS}} = \frac{1}{4}\|\mathcal{P}_{I \cap T^\perp}(H)\|_*$.

Note that $W_1 \in I$, and $\|W_1\|_F \leq 2$, so $\langle W_1, H\rangle_{\mathrm{HS}} \leq 2\|\mathcal{P}_I(H)\|_F$. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\langle X_i, H\rangle_{\mathrm{HS}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X, H\rangle_{\mathrm{HS}} \\ + \lambda\left(2\|\mathcal{P}_I(H)\|_F - \frac{1}{4}\|\mathcal{P}_{T^\perp \cap I^\perp}(H)\|_{*,s} - \frac{1}{4s}\|\mathcal{P}_{I^\perp}(H)\|_{1,1} - \frac{1}{4}\|\mathcal{P}_{I \cap T^\perp}(H)\|_*\right).\end{aligned} \tag{A.1}$$

Note that because $\mathcal{P}_{T \cap I^\perp}(H)$ has the form $\beta^* \otimes u + u \otimes \beta^*$ where $u \in I^\perp$, we have

$$\begin{aligned}\|\mathcal{P}_{T \cap I^\perp}(H)\|_{*,s} &= \|\mathcal{P}_{T \cap I^\perp}(H)\|_* + \frac{1}{s}\|\mathcal{P}_{T \cap I^\perp}(H)\|_{1,1} \\ &\leq \|\mathcal{P}_{T \cap I^\perp}(H)\|_* + \frac{1}{s}\|\mathcal{P}_{I^\perp}(H)\|_{1,1} + \|\mathcal{P}_{T^\perp \cap I^\perp}(H)\|_{*,s}\end{aligned}$$

and therefore

$$\begin{aligned}\|\mathcal{P}_{I^\perp}(H)\|_{*,s} &\leq \|\mathcal{P}_{T\cap I^\perp}(H)\|_{*,s} + \|\mathcal{P}_{T^\perp\cap I^\perp}(H)\|_{*,s} \\ &\leq \|\mathcal{P}_{T\cap I^\perp}(H)\|_* + \frac{1}{s}\|\mathcal{P}_{I^\perp}(H)\|_{1,1} + 2\|\mathcal{P}_{T^\perp\cap I^\perp}(H)\|_{*,s}.\end{aligned}$$

For large enough λ chosen as in the theorem statement, Corollary 4 and Lemma 20 give, with high probability,

$$\begin{aligned}\sum_{i=1}^n \xi_i\langle X, H\rangle_{\mathrm{HS}} &\leq \frac{\lambda}{16}(\|\mathcal{P}_I(H)\|_* + \|\mathcal{P}_{I^\perp}(H)\|_{*,s}) \\ &\leq \frac{\lambda}{16}\left(\|\mathcal{P}_I(H)\|_* + \|\mathcal{P}_{T\cap I^\perp}(H)\|_* + \frac{1}{s}\|\mathcal{P}_{I^\perp}(H)\|_{1,1} + 2\|\mathcal{P}_{T^\perp\cap I^\perp}(H)\|_{*,s}\right).\end{aligned}$$

On this event, using (Equation A.1) and noting that $\|\mathcal{P}_I(H)\|_F \leq \|H\|_F$, $\|\mathcal{P}_{T\cap I^\perp}(H)\|_* \leq \sqrt{2}\|H\|_F$, and

$$\|\mathcal{P}_{I\cap T^\perp}(H)\|_* \geq \|\mathcal{P}_I(H)\|_* - \|\mathcal{P}_{I\cap T}(H)\|_* \geq \|\mathcal{P}_I(H)\|_* - \sqrt{2}\|H\|_F,$$

we have

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, H\rangle_{\mathrm{HS}}^2 \lesssim \lambda(C\|H\|_F - \|\mathcal{P}_{I^\perp}(H)\|_{*,s} - \|\mathcal{P}_I(H)\|_*)$$

for a modest constant C. Since the left-hand side is nonnegative, $\|\mathcal{P}_{I^\perp}\|_{*,s} + \|\mathcal{P}_I(H)\|_* \leq C\|\mathcal{P}_I(H)\|_F$. By Lemma 21, we then have, with high probabability, $\frac{1}{n}\sum_{i=1}^n \langle X_i, H\rangle_{\mathrm{HS}}^2 \gtrsim \|H\|_F^2$. Thus

$$\|H\|_F^2 \lesssim \lambda\|H\|_F,$$

which completes the proof. □

A.4 Proof of sparse PCA error bound

Proof of Theorem 6. By a similar argument to that in the proof of Theorem 5 in Section A.3, the solution to (Equation 3.7) satisfies

$$\langle \widehat{\Sigma}, -H \rangle_{\mathrm{HS}} \leq \lambda \langle W, -H \rangle_{\mathrm{HS}},$$

for $H = \widehat{P} - P_1$ and any $W \in \partial \|P_1\|_{*,s}$. Choosing W as in the proof of Theorem 5, we obtain

$$\langle \widehat{\Sigma}, -H \rangle_{\mathrm{HS}} \leq \lambda \left(2\|\mathcal{P}_I(H)\|_F - \frac{1}{4}\|\mathcal{P}_{T^\perp \cap I^\perp}(H)\|_{*,s} - \frac{1}{4s}\|\mathcal{P}_{I^\perp}(H)\|_{1,1} - \frac{1}{4}\|\mathcal{P}_{I \cap T^\perp}(H)\|_* \right).$$

We first consider the difference between $\langle \widehat{\Sigma}, H \rangle_{\mathrm{HS}}$ and $\langle \Sigma, H \rangle_{\mathrm{HS}}$. Since the distribution of $\widehat{\Sigma}$ is independent of μ, we assume, without loss of generality, that $\mu = 0$. We write $x_i = \Sigma^{1/2} z_i$, where $z_i \sim \mathcal{N}(0, I_p)$, and $\Sigma^{1/2} = \sqrt{\sigma_1} v_1 \otimes v_1 + \Sigma_2^{1/2}$. We therefore want to bound

$$\langle \widehat{\Sigma} - \Sigma, H \rangle_{\mathrm{HS}} = \langle \Sigma^{1/2}(Z - I_p - \bar{z} \otimes \bar{z})\Sigma^{1/2}, H \rangle_{\mathrm{HS}},$$

where $Z = \frac{1}{n}\sum_{i=1}^n z_i \otimes z_i$ and $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$.

Denote by $\mathcal{P}_{(v_1 \otimes v_1)^\perp}$ the orthogonal projection onto the orthogonal complement of $v_1 \otimes v_1$. We can write

$$H = \langle H v_1, v_1 \rangle v_1 \otimes v_1 + \mathcal{P}_{(v_1 \otimes v_1)^\perp}(H).$$

First, for all $t \leq n$, with probability at least $1 - e^{-t}$,

$$\begin{aligned}
\left|\langle \widehat{\Sigma} - \Sigma, v_1 \otimes v_1 \rangle_{\mathrm{HS}}\right| &= \sigma_1 \left| \frac{1}{n} \sum_{i=1}^{n} (\langle z_i, v_1 \rangle^2 - 1) - \langle \bar{z}, v_1 \rangle^2 \right| \\
&\leq \sigma_1 \left| \frac{1}{n} \sum_{i=1}^{n} (\langle z_i, v_1 \rangle^2 - 1) \right| + \langle \bar{z}, v_1 \rangle^2 \\
&\lesssim \sigma_1 \left(\sqrt{\frac{t}{n}} + \frac{t}{n} \right) \\
&\lesssim \sigma_1 \sqrt{\frac{t}{n}},
\end{aligned}$$

where the second-to-last inequality follows from applying a Bernstein inequality to the sum and an ordinary Gaussian tail bound to the $\mathcal{N}(0, 1/n)$ random variable $\langle \bar{z}, v_1 \rangle$.

To analyze the remainder, we write

$$\langle \mathcal{P}_{(v_1 \otimes v_1)^\perp}(H), \widehat{\Sigma} - \Sigma \rangle_{\mathrm{HS}} = \langle H, \mathcal{P}_{(v_1 \otimes v_1)^\perp}(\widehat{\Sigma}) - \Sigma_2 \rangle_{\mathrm{HS}}.$$

Since

$$\mathcal{P}_{(v_1 \otimes v_1)^\perp}(x \otimes x) = \sqrt{\sigma_1} \langle z, v_1 \rangle (v_1 \otimes (\Sigma_2^{1/2} z) + (\Sigma_2^{1/2} z) \otimes v_1) + (\Sigma_2^{1/2} z)^{\otimes 2},$$

we obtain, if $t' \leq n$, with probability at least $1 - e^{-t'}$,

$$\begin{aligned}
\langle (\mathcal{P}_{(v_1 \otimes v_1)^\perp}(\widehat{\Sigma}) - \Sigma_2)u, v\rangle &= \frac{1}{n}\sum_{i=1}^n \sqrt{\sigma_1}\langle z_i, v_1\rangle \Big(\langle v_1, u\rangle \langle z_i, \Sigma_2^{1/2} v\rangle + \langle v_1, v\rangle \langle z_i, \Sigma_2^{1/2} u\rangle \Big) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \langle (z_i \otimes z_i - I_p)\Sigma_2^{1/2} u, \Sigma_2^{1/2} v\rangle \\
&\quad - \sqrt{\sigma_1}\langle \bar{z}, v_1\rangle \Big(\langle v_1, u\rangle \langle \bar{z}, \Sigma_2^{1/2} v\rangle + \langle v_1, v\rangle \langle \bar{z}, \Sigma_2^{1/2} u\rangle \Big) \\
&\quad - \langle (\bar{z} \otimes \bar{z})\Sigma_2^{1/2} u, \Sigma_2^{1/2} v\rangle \\
&\lesssim (\sqrt{\sigma_1 \sigma_2} + \sigma_2)\left(\sqrt{\frac{t'}{n}} + \frac{t'}{n} \right) \|u\|_2 \|v\|_2 \\
&\lesssim \sqrt{\sigma_1 \sigma_2}\sqrt{\frac{t'}{n}} \|u\|_2 \|v\|_2
\end{aligned}$$

for fixed $u, v \in \mathbf{R}^p$, where we have used the fact that all terms except the last are zero-mean. Applying the argument of Lemma 20, if $n \gtrsim s \log \frac{ep}{s}$, then, with probability at least $1 - e^{-s}(s/p)^s$,

$$\sup_{\substack{\|u\|_2 = \|v\|_2 = 1 \\ \|u\|_0, \|v\|_0 \leq s}} \langle (\mathcal{P}_{(v_1 \otimes v_1)^\perp}(\widehat{\Sigma}) - \Sigma_2)u, v\rangle \lesssim \sqrt{\sigma_1 \sigma_2}\sqrt{\frac{s \log(ep/s)}{n}}.$$

Using Corollary 4, we have with probability at least $1 - e^{-t} - e^{-s}(s/p)^s$

$$\langle \widehat{\Sigma} - \Sigma, H\rangle_{\mathrm{HS}} \lesssim \sqrt{\frac{t}{n}}\sigma_1 |\langle Hv_1, v_1\rangle| + \sqrt{\frac{s \log(ep/s)}{n}}\sqrt{\sigma_1 \sigma_2}\|H\|_{*,s}.$$

Note that $|\langle Hv_1, v_1\rangle| = 1 - \langle \widehat{P} v_1, v_1\rangle$. Also,

$$\begin{aligned}
\|H\|_{*,s} &\leq \|\mathcal{P}_I(H)\|_{*,s} + \|\mathcal{P}_{I^\perp}(H)\|_{*,s} \\
&\lesssim \|\mathcal{P}_I(H)\|_* + \|\mathcal{P}_{I^\perp}(H)\|_{*,s},
\end{aligned}$$

where the second inequality is due to the fact that any matrix in I has nonzero entries in at most s columns and rows.

Combining everything so far, we have

$$\begin{aligned}
&\langle \Sigma, P_1 - \widehat{P} \rangle_{\mathrm{HS}} - \sigma_1 \sqrt{\frac{t}{n}} (1 - \langle \widehat{P} v_1, v_1 \rangle) \\
&\quad \lesssim \sqrt{\sigma_1 \sigma_2} \sqrt{\frac{s \log(ep/s)}{n}} (\|\mathcal{P}_I(H)\|_* + \|\mathcal{P}_{I^\perp}(H)\|_{*,s}) \\
&\qquad + \lambda \bigg(\|H\|_F - \|\mathcal{P}_{T^\perp \cap I^\perp}(H)\|_{*,s} - \frac{1}{s} \|\mathcal{P}_{I^\perp}(H)\|_{1,1} - \|\mathcal{P}_{I \cap T^\perp}(H)\|_* \bigg).
\end{aligned}$$

By a similar argument to that in the proof of Theorem 5, choosing $\lambda \gtrsim \sqrt{\sigma_1 \sigma_2} \sqrt{\frac{s \log(ep/s)}{n}}$ gives

$$\langle \Sigma, P_1 - \widehat{P} \rangle_{\mathrm{HS}} - \sigma_1 \sqrt{\frac{t}{n}} (1 - \langle \widehat{P} v_1, v_1 \rangle) \lesssim \lambda \|H\|_F.$$

Now, note that if $\sqrt{\frac{t}{n}} \lesssim \frac{\sigma_1 - \sigma_2}{\sigma_1}$, then

$$\begin{aligned}
&\langle \Sigma, P_1 - \widehat{P} \rangle_{\mathrm{HS}} - \sigma_1 \sqrt{\frac{t}{n}} (1 - \langle \widehat{P} v_1, v_1 \rangle) \\
&\quad = \sigma_1 (1 - \langle \widehat{P} v_1, v_1 \rangle) - \langle \Sigma_2, \widehat{P} \rangle_{\mathrm{HS}} - \sigma_1 \sqrt{\frac{t}{n}} (1 - \langle \widehat{P} v_1, v_1 \rangle) \\
&\quad \geq \sigma_1 \left(1 - \sqrt{\frac{t}{n}} \right) (1 - \langle \widehat{P} v_1, v_1 \rangle) - \sigma_2 \|\mathcal{P}_{T^\perp}(\widehat{P})\|_* \\
&\quad = \left(\sigma_1 - \sigma_2 - \sigma_1 \sqrt{\frac{t}{n}} \right) (1 - \langle \widehat{P} v_1, v_1 \rangle) + \sigma_2 (1 - \langle \widehat{P} v_1, v_1 \rangle - \|\mathcal{P}_{T^\perp}(\widehat{P})\|_*) \\
&\quad \gtrsim (\sigma_1 - \sigma_2)(1 - \langle \widehat{P} v_1, v_1 \rangle) + \sigma_2 (1 - \langle \widehat{P} v_1, v_1 \rangle - \|\mathcal{P}_{T^\perp}(\widehat{P})\|_*) \\
&\quad \geq (\sigma_1 - \sigma_2)(1 - \langle \widehat{P} v_1, v_1 \rangle)
\end{aligned}$$

where the last inequality follows from

$$\begin{aligned}
0 &\geq \|\widehat{P}\|_* - 1 \\
&= \|\widehat{P}\|_* - \|P_1\|_* \\
&\geq \langle \widehat{P} - P_1, v_1 \otimes v_1 \rangle_{\mathrm{HS}} + \|\mathcal{P}_{T^\perp}(\widehat{P})\|_* \\
&= \langle \widehat{P} v_1, v_1 \rangle + \|\mathcal{P}_{T^\perp}(\widehat{P})\|_* - 1.
\end{aligned}$$

Therefore, requiring $n \gtrsim \left(\frac{\sigma_1}{\sigma_1-\sigma_2}\right)^2 t$, we have

$$1 - \langle \widehat{P} v_1, v_1 \rangle \lesssim \frac{\lambda}{\sigma_1 - \sigma_2} \|H\|_F. \tag{A.2}$$

Note that we can write

$$\widehat{P} = a v_1 \otimes v_1 + v_1 \otimes u + w \otimes v_1 + \mathcal{P}_{T^\perp}(\widehat{P}),$$

where $a = \langle \widehat{P} v_1, v_1 \rangle$ and $u, v \perp v_1$. Then

$$1 \geq \|\widehat{P}\|_*^2 \geq \|\widehat{P}\|_F^2 = a^2 + \|u\|_2^2 + \|w\|_2^2 + \|\mathcal{P}_{T^\perp}(\widehat{P})\|_F^2,$$

and therefore

$$\begin{aligned} \|H\|_F^2 &= (1-a)^2 + \|u\|_2^2 + \|w\|_2^2 + \|\mathcal{P}_{T^\perp}(\widehat{P})\|_F^2 \\ &\leq (1-a)^2 + 1 - a^2 \\ &= 2(1-a). \end{aligned}$$

Combining this with (Equation A.2), we obtain

$$\|H\|_F^2 \lesssim \frac{\lambda}{\sigma_1 - \sigma_2} \|H\|_F,$$

from which the result follows.

□

A.5 Proof of Poisson variance/moment bounds

If x satisfies Assumption 1 and, conditioned on x, $y \sim \mathrm{Poisson}(\langle x, \beta^* \rangle^2)$, then

$$\begin{aligned}
\mathbf{E}\, \xi^2 \langle x, u \rangle^4 &= \mathbf{E}\big[\mathbf{E}[\xi^2 \mid x] \langle x, u \rangle^4\big] \\
&= \mathbf{E} \langle x, \beta^* \rangle^2 \langle x, u \rangle^4 \\
&\lesssim \|\beta^*\|_2^2.
\end{aligned}$$

Also,

$$\begin{aligned}
\|\xi \langle x, u \rangle^2\|_\alpha &= \big(\mathbf{E}\big|\xi \langle x, u \rangle^2\big|^\alpha\big)^{1/\alpha} \\
&= \big(\mathbf{E}\big[\mathbf{E}[|\xi|^\alpha \mid x] |\langle x, u \rangle|^{2\alpha}\big]\big)^{1/\alpha} \\
&\lesssim \sqrt{\alpha} \big(\mathbf{E} |\langle x, \beta^* \rangle|^\alpha |\langle x, u \rangle|^{2\alpha}\big)^{1/\alpha} + \alpha \|\langle x, u \rangle^2\|_\alpha \\
&\lesssim \alpha^2 (\|\beta^*\|_2 + 1),
\end{aligned}$$

where the first inequality uses the standard (centered) Poisson moment bound

$$\|Z - \mathbf{E}\, Z\|_\alpha \lesssim \sqrt{\alpha \lambda} + \alpha$$

if $Z \sim \mathrm{Poisson}(\lambda)$.

APPENDIX B

MANIFOLD REGRESSION ANALYSIS

B.1 Proof of general RKHS results

We write $\mathcal{P}_G$ and $\mathcal{P}_{G^\perp}$ for the projections in L_2 and $\mathcal{H}$ onto G and its orthogonal complement $G^\perp$, respectively.

For brevity, we denote by P_n the empirical measure given by the n independent samples of the variables (X, ξ), i.e., $P_n w = \frac{1}{n}\sum_{i=1}^n w(X_i, \xi_i)$. For example, if $h\colon S \to \mathbf{R}$ is a function, $P_n h^2 = \frac{1}{n}\sum_{i=1}^n h^2(X_i)$, and $P_n \xi h = \frac{1}{n}\sum_{i=1}^n \xi_i h(X_i)$.

We use the following lemmas in our proof of Theorem 7:

Lemma 22. *Let $\delta \in (0,1)$. If*

$$n \geq \max\{7, 3\gamma'\} K_p \log \frac{\max\{2, 4\gamma\} p}{\delta},$$

then, with probability at least $1 - \delta$,

$$P_n f^2 \geq \frac{1}{2}\|f\|_{L_2}^2 - 3\sqrt{t_{p+1}}\|f\|_{L_2}\|f\|_{\mathcal{H}}$$

for all $f \in \mathcal{H}$.

Lemma 23. *There is a universal constant C such that, if*

$$\frac{n}{\log^2 n} \geq C(1 \vee \gamma')\frac{K_p}{p}\frac{\|\xi\|_{\psi_1}^2}{\sigma^2},$$

then, with probability at least $1-\delta$*,*

$$|P_n \xi f| \leq \frac{3}{2}\sigma \cdot \left(\frac{\sqrt{p} + 2\sqrt{\log 4/\delta}}{\sqrt{n}} \|f\|_{L_2} + \frac{\sqrt{\operatorname{tr} T_{G^\perp}} + 2\sqrt{t_{p+1} \log 4/\delta}}{\sqrt{n}} \|f\|_{\mathcal{H}} \right)$$
$$\leq \frac{3}{2}\sigma \cdot \left(\frac{\sqrt{p} + 2\sqrt{\log 4/\delta}}{\sqrt{n}} \right) \left(\|f\|_{L_2} + \sqrt{\gamma t_{p+1}} \|f\|_{\mathcal{H}} \right).$$

for all $f \in \mathcal{H}$*.*

With these, we prove the main result:

Proof of Theorem 7. We write our objective function as

$$F(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \alpha \|f\|_{\mathcal{H}}^2.$$

$\hat{f}$ satisfies $\nabla F(\hat{f}) = 0$. Noting that

$$\frac{1}{2} \nabla F(f) = -\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i)) k(\cdot, X_i) + \alpha f,$$

we have

$$\begin{aligned}
0 &= \frac{1}{2} \langle \nabla F(\hat{f}), f^* - \hat{f} \rangle_{\mathcal{H}} \\
&= \langle *, \alpha \hat{f} - \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}(X_i)) k(\cdot, X_i) \rangle_{\mathcal{H}} f^* - \hat{f} \\
&= \alpha \langle \hat{f}, f^* - \hat{f} \rangle_{\mathcal{H}} - \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}(X_i))(f^*(X_i) - \hat{f}(X_i)) \\
&= \alpha \langle \hat{f}, f^* - \hat{f} \rangle_{\mathcal{H}} + \frac{1}{n} \sum_{i=1}^{n} \left[(Y_i - f^*(X_i))(\hat{f}(X_i) - f^*(X_i)) - (\hat{f}(X_i) - f^*(X_i))^2 \right] \\
&= \alpha \langle \hat{f}, f^* - \hat{f} \rangle_{\mathcal{H}} + P_n \xi (\hat{f} - f^*) - P_n (\hat{f} - f^*)^2.
\end{aligned} \tag{B.1}$$

Let E_1 and E_2 denote the events of Lemmas 22 and 23. For part 1 of the theorem, we assume that E_1 holds, which occurs with probability at least $1-\delta$. For part 2, we assume

$E_1 \cap E_2$ holds, which occurs with probability at least $1 - 2\delta$. In what follows, we treat the two cases the same (and assume $\alpha > 0$), since we can simply take $\sigma = 0$ and the limit $\alpha \downarrow 0$ for part 1.

Let $e_2 = \|\hat{f} - f^*\|_{L_2}$ and $e_{\mathcal{H}} = \|\hat{f} - f^*\|_{\mathcal{H}}$. On $E_1 \cap E_2$, (Equation B.1) implies

$$\frac{1}{2}e_2^2 \leq \sigma(ae_2 + be_{\mathcal{H}}) + ce_2e_{\mathcal{H}} + \alpha\langle\hat{f}, f^* - \hat{f}\rangle_{\mathcal{H}},$$

where $a = \frac{3}{2}\frac{\sqrt{p}+2\sqrt{\log 4/\delta}}{\sqrt{n}}$, $b = \sqrt{\gamma t_{p+1}}a$, and $c = 3\sqrt{t_{p+1}}$. First, note that

$$\langle\hat{f}, f^* - \hat{f}\rangle_{\mathcal{H}} = \langle f^*, f^* - \hat{f}\rangle_{\mathcal{H}} - e_{\mathcal{H}}^2 \leq \|f^*\|_{\mathcal{H}}e_{\mathcal{H}} - e_{\mathcal{H}}^2,$$

so

$$\sigma be_{\mathcal{H}} + \alpha\langle\hat{f}, f^* - \hat{f}\rangle_{\mathcal{H}} \leq (\sigma b + \alpha\|f^*\|_{\mathcal{H}})e_{\mathcal{H}} - \alpha e_{\mathcal{H}}^2 \leq \frac{(\sigma b + \alpha\|f^*\|_{\mathcal{H}})^2}{\alpha}.$$

To control the error term $ce_2e_{\mathcal{H}}$, we need a more explicit bound on $e_{\mathcal{H}}$. Because $P_n(\hat{f} - f^*)^2 \geq 0$, (Equation B.1) gives

$$e_{\mathcal{H}}^2 \leq \|f^*\|_{\mathcal{H}}e_{\mathcal{H}} + \frac{1}{\alpha}P_n\xi(\hat{f} - f^*) \leq \|f^*\|_{\mathcal{H}}e_{\mathcal{H}} + \frac{\sigma}{\alpha}(ae_2 + be_{\mathcal{H}}).$$

Because $x^2 \leq a + bx$ implies $x \leq \sqrt{a} + b$, we then have

$$e_{\mathcal{H}} \leq \|f^*\|_{\mathcal{H}} + \frac{\sigma b}{\alpha} + \sqrt{\frac{\sigma ae_2}{\alpha}}.$$

Putting everything together, we have

$$\frac{1}{2}e_2^2 \leq \frac{(\sigma b + \alpha\|f^*\|_{\mathcal{H}})^2}{\alpha} + \sigma ae_2 + ce_2\left(\|f^*\|_{\mathcal{H}} + \frac{\sigma b}{\alpha} + \sqrt{\frac{\sigma ae_2}{\alpha}}\right).$$

$x^2 \le a + bx + cx^{3/2}$ implies $x \le \sqrt{a} + b + c^2$, so

$$\begin{aligned} e_2 &\le \sqrt{2}\frac{\sigma b}{\sqrt{\alpha}} + \sqrt{2\alpha}\|f^*\|_{\mathcal{H}} + 2\sigma a + 2c\|f^*\|_{\mathcal{H}} + 2\frac{\sigma c b}{\alpha} + 4\frac{c^2\sigma a}{\alpha} \\ &= (\sqrt{2\alpha} + 2c)\|f^*\|_{\mathcal{H}} + 2\sigma\left(a + \frac{b}{\sqrt{2\alpha}} + \frac{bc}{\alpha} + 2\frac{ac^2}{\alpha}\right). \end{aligned}$$

The result immediately follows by substituting our choices of a, b, and c and, if $\sigma \neq 0$, using the assumption that $\alpha \ge 54 t_{p+1}$. □

B.1.1 Proofs of key lemmas

Lemma 22 follows quickly from the following two concentration results:

Lemma 24. *If $\delta \in (0,1)$, and $n \ge 7K_p \log\frac{p}{\delta}$, then, with probability at least $1-\delta$, for all $f \in G$,*

$$P_n f^2 \ge \frac{1}{2}\|f\|_{L_2}^2.$$

Proof. Note that for all $f \in G$,

$$\begin{aligned} P_n f^2 &= \frac{1}{n}\sum_{i=1}^n f^2(X_i) \\ &= \frac{1}{n}\sum_{i=1}^n \langle Z(X_i), f\rangle_{L_2}^2 \\ &= \langle (P_n(Z \otimes_{L_2} Z))f, f\rangle_{L_2}, \end{aligned}$$

where we define $Z(X) = \sum_{\ell=1}^p v_\ell(X) v_\ell \in G$. The lemma will follow from a concentration result on $P_n(Z \otimes_{L_2} Z)$. Note that the operator $Z(X) \otimes_{L_2} Z(X) \succeq 0$ for all X, and, by Assumption 4, we have

$$\|Z(X) \otimes_{L_2} Z(X)\|_{L_2} = \|Z(X)\|_{L_2}^2 = \sum_{\ell=1}^p v_\ell^2(X) \le K_p$$

almost surely. Also, $\mathbf{E}\, P_n(Z \otimes_{L_2} Z) = \mathbf{E}\, Z(X) \otimes_{L_2} Z(X) = \mathcal{I}_G$. The matrix Chernoff

bound [49, Theorem 5.1.1] implies that, for all $\epsilon \in [0, 1)$,

$$\mathrm{P}(P_n(Z \otimes_{L_2} Z) \succeq (1-\epsilon)\mathcal{I}_G) \geq 1 - p\left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{n/K_p}.$$

Choosing $\epsilon = 1/2$ gives the result. □

Lemma 25. *If $\delta \in (0, 1)$, and $n \geq \frac{3R_p}{t_{p+1}} \log \frac{2 \operatorname{tr} T_{G^\perp}}{t_{p+1}\delta}$, then, with probability at least $1 - \delta$, for all $f \in G^\perp$,*

$$P_n f^2 \leq 2t_{p+1}\|f\|_{\mathcal{H}}^2.$$

Proof. Similarly to the proof of Lemma 24, for all $f \in G^\perp$,

$$P_n f^2 = \langle (P_n(W \otimes_{\mathcal{H}} W))f, f\rangle_{\mathcal{H}},$$

where $W(X) = \sum_{\ell > p} t_\ell v_\ell(X) v_\ell$. Note that $\mathbf{E}\, W(X) \otimes_{\mathcal{H}} W(X) = T_{G^\perp}$. By Assumption 4,

$$\|W(X) \otimes_{\mathcal{H}} W(X)\|_{\mathcal{H}} = \|W(X)\|_{\mathcal{H}}^2 = \sum_{\ell > p} t_\ell v_\ell^2(X) \leq R_p$$

almost surely. By [49, Theorem 7.2.1], if $\epsilon \geq R_p / n t_{p+1}$, then

$$\mathrm{P}(\|P_n(W \otimes_{\mathcal{H}} W)\|_{\mathcal{H}} \leq (1+\epsilon)t_{p+1}) \geq 1 - 2d_p\left(\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right)^{nt_{p+1}/R_p},$$

where $d_p = \operatorname{tr} T_{G^\perp}/t_{p+1}$. Choosing $\epsilon = 1$ gives the result. □

Proof of Lemma 22. Applying Lemmas 24 and 25 (with $\delta/2$ substituted for δ) and a union bound, we have, with probability at least $1 - \delta$,

$$\begin{aligned}\sqrt{P_n f^2} &\geq \sqrt{P_n(\mathcal{P}_G f)^2} - \sqrt{P_n(\mathcal{P}_{G^\perp} f)^2} \\ &\geq \frac{1}{\sqrt{2}}\|\mathcal{P}_G f\|_{L_2} - \sqrt{2t_{p+1}}\|\mathcal{P}_{G^\perp} f\|_{\mathcal{H}},\end{aligned}$$

so

$$\begin{aligned} P_n f^2 &\geq \frac{1}{2}\|\mathcal{P}_G f\|_{L_2}^2 - 2\sqrt{t_{p+1}}\|\mathcal{P}_G f\|_{L_2}\|\mathcal{P}_{G^\perp} f\|_{\mathcal{H}} \\ &\geq \frac{1}{2}\|f\|_{L_2}^2 - \frac{1}{2}\|\mathcal{P}_{G^\perp} f\|_{L_2}^2 - 2\sqrt{t_{p+1}}\|f\|_{L_2}\|f\|_{\mathcal{H}} \\ &\geq \frac{1}{2}\|f\|_{L_2}^2 - 3\sqrt{t_{p+1}}\|f\|_{L_2}\|f\|_{\mathcal{H}}. \end{aligned}$$

□

Proof of Lemma 23. Let B_2^G denote the L_2-unit ball in G, and let $B_{\mathcal{H}}^{G^\perp}$ denote the $\mathcal{H}$-unit ball in $G^\perp$. Note that for all $f \in \mathcal{H}$, we have

$$f \in \|f\|_{L_2} B_2^G + \|f\|_{\mathcal{H}} B_{\mathcal{H}}^{G^\perp},$$

where the plus sign denotes Minkowski addition. Therefore, because $|P_n \xi f|$ is sublinear in f, it suffices to bound

$$Z_1 \coloneqq \sup_{f \in B_2^G} |P_n \xi f|$$

and

$$Z_2 \coloneqq \sup_{f \in B_{\mathcal{H}}^{G^\perp}} |P_n \xi f|.$$

We present a complete proof for the bound of Z_1; the proof for Z_2 is similar.

First, note that

$$\begin{aligned} Z_1 &= \sup_{f \in B_2^G} |P_n \xi f| \\ &= \sup_{\sum_{\ell=1}^p a_\ell^2 \le 1} \left| P_n \left(\xi \sum_{\ell=1}^p a_\ell v_\ell \right) \right| \\ &= \sup_{\sum_{\ell=1}^p a_\ell^2 \le 1} \left| \sum_{\ell=1}^p a_\ell P_n(\xi v_\ell) \right| \\ &= \left(\sum_{\ell=1}^p P_n^2(\xi v_\ell) \right)^{1/2}, \end{aligned}$$

so

$$\mathbf{E}\, Z_1 \le \sqrt{\mathbf{E}\, Z_1^2} = \sqrt{\sum_{\ell=1}^p \mathbf{E}\, P_n^2(\xi v_\ell)} = \sigma \sqrt{\frac{p}{n}}.$$

We also have

$$\sup_{f \in B_2^G} \sum_{i=1}^n \mathbf{E}(\xi_i f(x_i))^2 = n\sigma^2.$$

Finally, note that

$$\sup_{f \in B_2^G} \|f\|_\infty \le \sqrt{K_p},$$

so

$$\left\| \max_i \sup_{f \in B_2^G} |\xi_i f(x_i)| \right\|_{\psi_1} \le \sqrt{K_p} \|\xi\|_{\psi_1} \log n.$$

Let $\eta \in (0,1)$. [173, Theorem 4] (with, in the notation of that paper, $\delta = 1$) implies that, with probability at least $1 - \delta/2$,

$$Z_1 \le \sigma \left((1+\eta) \sqrt{\frac{p}{n}} + 2 \sqrt{\frac{\log 4/\delta}{n}} \right) + \frac{C'_\eta \sqrt{K_p} \|\xi\|_{\psi_1} (\log n)(\log 12/\delta)}{n}$$

for a constant C'_η that only depends on η. By a similar argument, we have, with probability

at least $1 - \delta/2$,

$$Z_2 \leq \sigma \left((1+\eta)\sqrt{\frac{\operatorname{tr} T_{G^\perp}}{n}} + 2\sqrt{\frac{t_{p+1} \log 4/\delta}{n}} \right) + \frac{C'_\eta \sqrt{R_p} \|\xi\|_{\psi_1} (\log n)(\log 12/\delta)}{n}.$$

Fixing $\eta \in (0, 1/2)$ and choosing a suitable constant C to ensure n is large enough completes the proof. □

B.2 Proof of heat kernel approximation

In this appendix, we prove upper and lower bounds on the heat kernel diagonal values. Although we only use the upper bound in our paper, we include the lower bound also as both may be of independent interest.

The concepts from differential geometry used in this section can be found in, for example, [174, 175]. The key tools we will use in our analysis of how well the heat kernel is approximated by a Gaussian RBF are the following *comparison theorems*:

Lemma 26 ([91, Theorem 4.5.1])**.** *If the sectional curvature of an m-dimensional manifold $\mathcal{M}$ is bounded above by $K > 0$, then, for all $x, y \in \mathcal{M}$, $k_t^{\mathrm{h}}(x, y) \leq k_t^{h,K}(d_{\mathcal{M}}(x, y))$, where $k_t^{h,K}(r)$ is the (radially symmetric) heat kernel on the m-dimensional space of constant curvature K, and, if $K > 0$, we set $k_t^{h,K}(r) = k_t^{h,K}(\pi/\sqrt{K})$ for $r \geq \pi/\sqrt{K}$.*

Lemma 27 ([91, Theorem 4.5.2])**.** *If the Ricci curvature of $\mathcal{M}$ is bounded below by $(m-1)K$ for some constant K, then, for all $x, y \in \mathcal{M}$, $k_t^{\mathrm{h}}(x, y) \geq k_t^{h,K}(d_{\mathcal{M}}(x, y))$, where $k_t^{h,K}(r)$ is the heat kernel on the space of constant curvature K.*

A lower bound of K on sectional curvature implies a lower bound of $(m-1)K$ on the Ricci curvature tensor (see, e.g., the formula for $\mathrm{Ric}(v, v)$ in [175, p. 38]), so Lemma 27 also holds under the (stronger) assumption of a lower bound of K on sectional curvature.

The space of constant curvature $K > 0$ is the sphere $S_K^m = S^m/\sqrt{K}$, while the space of constant curvature $-K < 0$ is the scaled hyperbolic space $H_K^m = H^m/\sqrt{K}$. To apply

Lemmas 26 and 27, we need to find bounds for the heat kernel on the sphere and on hyperbolic space.

We will use the following result:

Lemma 28 ([176, Theorem 1])**.** *The heat kernel in hyperbolic space H^m has the radial representation*

$$k_t^{h,H^m}(r) = e^{-\frac{(m-1)^2 t}{8}} \left(\frac{r}{\sinh r}\right)^{\frac{m-1}{2}} \frac{e^{-r^2/2t}}{(2\pi t)^{m/2}} \times \mathbf{E}_r \exp\left(-\frac{(m-1)(m-3)}{8} \int_0^t \left(\frac{1}{\sinh^2 R_s} - \frac{1}{R_s^2}\right) \, ds\right),$$

where R_s is an m-dimensional Bessel process, and $\mathbf{E}_r$ denotes expectation conditioned on $R_t = r$.

A nearly identical argument to that in [176] gives a corresponding result for the sphere S^m for $m \geq 2$:

Lemma 29. *For all $m \geq 2$, the heat kernel on the sphere S^m has the radial representation*

$$k_t^{h,S^m}(r) = e^{\frac{(m-1)^2 t}{8}} \left(\frac{r}{\sin r}\right)^{\frac{m-1}{2}} \frac{e^{-r^2/2t}}{(2\pi t)^{m/2}} \times \mathbf{E}_r \exp\left(-\frac{(m-1)(m-3)}{8} \int_0^t \left(\frac{1}{\sin^2 R_s} - \frac{1}{R_s^2}\right) \, ds\right),$$

where, again, R_s is an m-dimensional Bessel process, and $\mathbf{E}_r$ denotes expectation conditioned on $R_t = r$.

For $m \geq 3$, the exponent in the integrands in the formula of Lemma 28 (resp. Lemma 29) is always positive (resp. negative), so we have the following simple bounds on the heat kernels on the standard spaces of constant curvature:

$$k_t^{h,H^m}(r) \geq e^{-\frac{(m-1)^2 t}{8}} \left(\frac{r}{\sinh r}\right)^{\frac{m-1}{2}} \frac{e^{-r^2/2t}}{(2\pi t)^{m/2}}, \tag{B.2}$$

and

$$k_t^{h,S^m}(r) \leq e^{\frac{(m-1)^2 t}{8}} \left(\frac{r}{\sin r}\right)^{\frac{m-1}{2}} \frac{e^{-r^2/2t}}{(2\pi t)^{m/2}}. \tag{B.3}$$

It is easily verified that $p_t^{S^m_K}(r) = p_{Kt}^{S^m}(\sqrt{K}r)$, with a similar formula for scaled hyperbolic space. We can summarize this in the following result:

Lemma 30. *Suppose $\mathcal{M}$ is an m-dimensional complete Riemannian manifold for $m \geq 3$.*

1. *Suppose $\mathcal{M}$ has Ricci curvature bounded below by $-(m-1)K_1$. Then, for all $x, y \in \mathcal{M}$, denoting $r = d(x,y)$,*

$$k_t^{\mathrm{h}}(x,y) \geq e^{-\frac{(m-1)^2}{8}K_1 t} \left(\frac{\sqrt{K_1}r}{\sinh(\sqrt{K_1}r)}\right)^{\frac{m-1}{2}} \frac{e^{-r^2/2t}}{(2\pi t)^{m/2}}.$$

2. *Suppose $\mathcal{M}$ has sectional curvature bounded above by K_2. Then, for $r < \pi/\sqrt{K_2}$, and for all $x, y \in \mathcal{M}$ such that $d(x,y) \geq r$,*

$$k_t^{\mathrm{h}}(x,y) \leq e^{\frac{(m-1)^2}{8}K_2 t} \left(\frac{\sqrt{K_2}r}{\sin(\sqrt{K_2}r)}\right)^{\frac{m-1}{2}} \frac{e^{-r^2/2t}}{(2\pi t)^{m/2}}.$$

We note that, for $r = 0$ and t small, these results are comparable to the well-known asymptotic expansion for the heat kernel, which depends on the scalar curvature at x (see, e.g., [90, Section VI.4]).

Finally, we specialize to the case $r = 0$ and simplify:

Proposition 1. *Let $\epsilon \leq 2/3$.*

1. *Under the conditions of Lemma 30.1, for $t \leq \frac{8\epsilon}{(m-1)^2 K_1}$ and all $x \in \mathcal{M}$,*

$$k_t^{\mathrm{h}}(x,x) \geq \frac{1-\epsilon}{(2\pi t)^{m/2}}.$$

2. *Under the conditions of Lemma 30.2, for* $t \leq \frac{6\epsilon}{(m-1)^2 K_2}$ *and all* $x \in \mathcal{M}$,

$$k_t^{\mathrm{h}}(x,x) \leq \frac{1+\epsilon}{(2\pi t)^{m/2}}.$$

Proof. From Lemma 30, we have

$$e^{-\frac{(m-1)^2}{8}K_1 t} \leq (2\pi t)^{-m/2} k_t^{\mathrm{h}}(x,x) \leq e^{\frac{(m-1)^2}{8}K_2 t}.$$

The result follows from noting that $e^{-s} \geq 1 - s$ for all $s \geq 0$, and $e^s \leq 1 + \frac{4}{3}s$ for $0 \leq s \leq 1/2$. □

Lemma 9 is a case of this last result, taking $K_2 = \kappa$.

B.3 Proof of non-asymptotic Weyl law estimates

Proof of Theorem 8. By Lemma 9, for all $\lambda \geq 0$ and $t \leq \frac{6\epsilon}{(m-1)^2\kappa}$,

$$\begin{aligned} e^{-\lambda t/2} N_x(\lambda) &= e^{-\lambda t/2} \sum_{\lambda_\ell \leq \lambda} v_\ell^2(x) \\ &\leq \sum_{\ell=0}^{\infty} e^{-\lambda_\ell t/2} v_\ell^2(x) \\ &= k_t^{\mathrm{h}}(x,x) \\ &\leq \frac{1+\epsilon}{(2\pi t)^{m/2}}. \end{aligned}$$

Taking $t = m/\lambda$, we get

$$\begin{aligned} N_x(\lambda) &\leq \frac{(1+\epsilon)e^{\lambda t/2}}{(2\pi t)^{m/2}} \\ &= \frac{1+\epsilon}{(4\pi)^{m/2}} \frac{e^{m/2}}{(m/2)^{m/2}} \lambda^{m/2} \\ &\leq \frac{1+\epsilon}{(4\pi)^{m/2}} \frac{2\sqrt{m}}{\Gamma\left(\frac{m}{2}+1\right)} \lambda^{m/2} \\ &= \frac{2(1+\epsilon)\sqrt{m}}{(2\pi)^m} V_m \lambda^{m/2}, \end{aligned}$$

where the second inequality uses Stirling's approximation. □

Proof of Lemma 10. For $c \in (0,1)$, note that

$$\begin{aligned} \sum_{\lambda_\ell \geq \lambda} e^{-\lambda_\ell t/2} v_\ell^2(x) &\leq e^{-(1-c)\lambda t/2} \sum_{\lambda_\ell \geq \lambda} e^{-c\lambda_\ell t/2} v_\ell^2(x) \\ &\leq e^{-(1-c)\lambda t/2} \sum_{k=0}^{\infty} e^{-c\lambda_\ell t/2} v_\ell^2(x) \\ &= e^{-(1-c)\lambda t/2} p_{ct}^{\mathcal{M}}(x,x) \\ &\leq e^{-\lambda t/2}(1+\epsilon) \frac{e^{c\lambda t/2}}{(2\pi c t)^{m/2}}. \end{aligned}$$

Choosing $c = m/\lambda t$, the remainder of the proof is identical to that of Theorem 8. □

B.4 Proof of manifold regression results

Proof of Theorems 9 and 10. To apply the framework of Sections 4.2.1 and 4.4.1, which assumes the set S has measure 1, we consider the normalized volume measure $d\widetilde{V} = dV/\operatorname{vol}\mathcal{M}$. With respect to $\widetilde{V}$, k_t^{h} has the eigenvalue decomposition

$$k_t^{\mathrm{h}}(x,y) = \frac{1}{\operatorname{vol}\mathcal{M}} \sum_\ell e^{-\lambda_\ell t/2} \tilde{u}_\ell(x) \tilde{u}_\ell(y),$$

where $\tilde{u}_\ell = \sqrt{\operatorname{vol}\mathcal{M}} u_\ell$. A similar normalized expansion holds for k_Ω^{bl}.

Note that Theorem 8 and Lemma 10 only give us bounds on the contants K_p and R_p in Assumption 4. For k_Ω^{bl}, this holds with $K_p = p(\Omega)$ (taking $\epsilon = 1/2$ in Theorem 8) and $R_p = 0$. Assumption 5 holds trivially with $\gamma = \gamma' = 0$.

For k_t^{h}, we can again take $K_p = p(\Omega)$ (again taking $\epsilon = 1/2$), and we get a bound on R_p such that $\gamma = \gamma' = 1$.

Finally, for both kernels, we take into account the fact that $\|\cdot\|_{L_2(\mathcal{M},\widetilde{V})} = \frac{\|\cdot\|_{L_2(\mathcal{M},V)}}{\sqrt{\operatorname{vol}\mathcal{M}}}$. With these considerations in mind, the results follow from Theorem 7. □

APPENDIX C

INTERPOLATION ANALYSIS

C.1 Notation

For convenience in reading, we collect all notation that is used for the proofs in Table C.1.

In addition, we will use many different norms. For a function $f\colon X \to \mathbf{R}$, $\|f\|_{L_p} \coloneqq$

Table C.1: Notation

Symbol(s)	Definition(s)	Description
k_x	$k_x = k(\cdot, x)$	Kernel function centered at x
$\mathcal{T}$	$\mathcal{T}(f) = \int f(x) k_x \, d\mu(x)$	Integral operator of kernel k
$\{(\lambda_\ell, v_\ell)\}_{\ell=1}^{\infty}$	$\mathcal{T}(f) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle f, v_\ell \rangle_{L_2} v_\ell, \ \lambda_1 \geq \lambda_2 \geq \cdots$	Eigenvalue decomposition of $\mathcal{T}$
$\mathcal{A}$	$\mathcal{A}(f) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$	Sampling operator from $\mathcal{H}$ to $\mathbf{R}^n$
$\mathcal{A}^*$	$\mathcal{A}^*(z) = \sum_{i=1}^{n} z_i k_{x_i}$	Adjoint of $\mathcal{A}$ w.r.t $\mathcal{H}$ and ℓ_2 inner products
$G, G^\perp$	$G = \operatorname{span}\{v_1, \ldots, v_p\}$	Span of first p eigenfunctions of $\mathcal{T}$ (and its complement)
$\mathcal{I}$ ($\mathcal{I}_G$)		Identity operator (restricted to G)
$\mathcal{T}_G, \mathcal{T}_{G^\perp}$	$\mathcal{T}_G = \mathcal{T}\mathcal{P}_G, \mathcal{T}_{G^\perp} = \mathcal{T}\mathcal{P}_{G^\perp}$	$\mathcal{T}$ restricted to G and $G^\perp$
$\mathcal{A}_G, \mathcal{R}$	$\mathcal{A}_G = \mathcal{A}\mathcal{P}_G, \mathcal{R} = \mathcal{A}\mathcal{P}_{G^\perp}$	Restrictions of sampling operator to G, $G^\perp$
$\mathcal{C}, \mathcal{C}^*$	$\mathcal{C} = \mathcal{A}_G, \mathcal{C}^* = \mathcal{T}_G^{-1}\mathcal{A}_G^*$	Sampling operator and its adjoint on G w.r.t. L_2 inner product on G
α		Explicit regularization parameter
α_L, α_U	$\alpha_L I_n \preceq \alpha I_n + \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$	Lower and upper bounds on explicit+implicit regularization
$\bar{\alpha}, \tilde{\alpha}$	$\bar{\alpha} = \frac{2\alpha_U\alpha_L}{\alpha_U+\alpha_L}, \tilde{\alpha} = \frac{\alpha_U+\alpha_L}{2}$	Harmonic and arithmetic means of α_U, α_L
$\mathcal{B}$	$\mathcal{B} = (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$	Bias operator on G
$\mathcal{S}$	$\mathcal{S} = \mathcal{I}_G - \mathcal{B}$	Kernel regression operator ("survival") on G
$\overline{\mathcal{B}}$	$\overline{\mathcal{B}} = \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$	Idealized approximation to bias $\mathcal{B}$
$\overline{\mathcal{S}}$	$\overline{\mathcal{S}} = \mathcal{I}_G - \overline{\mathcal{B}} = \frac{n}{\bar{\alpha}}\mathcal{T}_G\left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$	Idealized approximation to survival $\mathcal{S}$

$\left((\mathbf{E}_{x\sim\mu}|f(x)|^p)^{1/p}\right)$. For $f \in \mathcal{H}$, $\|f\|_{\mathcal{H}} = \|\mathcal{T}^{-1/2}f\|_{L_2}$ is the RKHS norm. For $u \in \mathbf{R}^n$, $\|u\|_{\ell_2}$ is the standard Euclidean norm. We denote the L_2, $\mathcal{H}$, and ℓ_2 inner products by $\langle\cdot,\cdot\rangle_{L_2}$, $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, and $\langle\cdot,\cdot\rangle_{\ell_2}$, respectively.

$\|\cdot\|_{L_2}$, $\|\cdot\|_{\mathcal{H}}$, and $\|\cdot\|_{\ell_2}$ also denote operator norms when applied to operators from the corresponding Hilbert space to itself. We will write the operator norm of an operator $T\colon H_1 \to H_2$ (for any Hilbert spaces H_1 and H_2) with respect to the H_1 and H_2 norms as $\|T\|_{H_1\to H_2}$. Similarly, $\|T\|_{HS,H_1\to H_2}$ refers to the Hilbert-Schmidt norm of T with respect to the H_1 and H_2 inner products.

C.2 Proofs of deterministic-sample results

We begin with the proofs of the deterministic-sample results (Theorems 11 and 12).

In this section, we will often abbreviate scaled identity operators such as aI_n, $a\mathcal{I}$, $a\mathcal{I}_G$, $a\mathcal{I}_G{}^{\perp}$ by the number a. The meaning should be clear from context.

C.2.1 Bias

The main technical challenge for proving Theorem 11 is bounding the approximation error between the "ideal" bias operator $\overline{\mathcal{B}} = \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$ (discussed in Section 5.2.4) and the actual bias, which turns out to be $\mathcal{B} \coloneqq (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$ (derived in the proof of Theorem 11 below). The following result quantifies this error.

Lemma 31. *Under the conditions of Theorem 11,*

$$\|\mathcal{B} - \overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \le \frac{c}{1-c}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2},$$

where $c < 1$ is an upper bound on the quantity $\frac{\alpha_U-\alpha_L}{\alpha_U+\alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2}$ (as defined in Theorem 11).

Proof. Recall that we have assumed

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \leq c < 1.$$

A standard perturbation argument (e.g., [196, p. 335]) gives

$$\begin{aligned}\mathcal{B} - \overline{\mathcal{B}} &= \sum_{i=1}^{\infty}(-1)^i\Big[\overline{\mathcal{B}}\Big(\mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G - \frac{n}{\bar{\alpha}}\mathcal{T}_G\Big)\Big]^k\overline{\mathcal{B}} \\ &= \left(\sum_{i=1}^{\infty}(-1)^i\left[\Big(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\Big)^{-1}\Big(\mathcal{C}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\Big)\right]^k\right)\overline{\mathcal{B}}\end{aligned}$$

as long as the operator norm (in any space) of the bracketed operator

$$\Big(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\Big)^{-1}\Big(\mathcal{C}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\Big)$$

is strictly less than 1.

We now show that this is the case. We have

$$\begin{aligned}\Big\|\mathcal{C}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\Big\|_{L_2} &\leq \left\|\mathcal{C}^*\left((\alpha + \mathcal{R}\mathcal{R}^*)^{-1} - \frac{1}{\bar{\alpha}}\right)\mathcal{C}\right\|_{L_2} + \frac{1}{\bar{\alpha}}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \\ &\leq \|\mathcal{C}\|_{L_2\to\ell_2}^2 \max\left\{\left|\frac{1}{\alpha_L} - \frac{1}{\bar{\alpha}}\right|, \left|\frac{1}{\alpha_U} - \frac{1}{\bar{\alpha}}\right|\right\} \\ &\quad + \frac{1}{\bar{\alpha}}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \\ &\leq \frac{\alpha_U - \alpha_L}{2\alpha_U\alpha_L}\big(n + \|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2}\big) + \frac{1}{\bar{\alpha}}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2},\end{aligned}$$

where the first and third inequalities use the triangle inequality, and the second inequality uses $\alpha_L \preceq \alpha + \mathcal{R}\mathcal{R}^* \preceq \alpha_U$. Then, since

$$\left\|\Big(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\Big)^{-1}\right\|_{L_2} \leq \frac{\bar{\alpha}}{n},$$

we have

$$\begin{aligned}
&\left\| \left(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}} \mathcal{I}_G \right)^{-1} \left(\mathcal{C}^* (\alpha + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{C} - \frac{n}{\bar{\alpha}} \mathcal{I}_G \right) \right\|_{L_2} \\
&\quad \leq \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \left(1 + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} \right) \cdot \frac{1}{n} \| \mathcal{C}^* \mathcal{C} - n \mathcal{I}_G \|_{L_2} \\
&\quad \leq \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n} \| \mathcal{C}^* \mathcal{C} - n \mathcal{I}_G \|_{L_2} \\
&\quad \leq c.
\end{aligned}$$

Since $c < 1$, the rest of the bound follows via the expression for the infinite sum of a geometric series. □

We are now ready to prove our main deterministic bias result (Theorem 11).

Proof of Theorem 11. Since $f^* \in G$, the full expression for the noiseless regression estimate is

$$\hat{f}_0 = \begin{bmatrix} \mathcal{A}_G^* \\ \mathcal{R}^* \end{bmatrix} (\alpha + \mathcal{A}_G \mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{A}_G f^*.$$

The pushthrough identity gives

$$\begin{aligned}
(\alpha + \mathcal{A}_G \mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{A}_G &= (\alpha + \mathcal{R}\mathcal{R}^*)^{-1} (I_n + \mathcal{A}_G \mathcal{A}_G^* (\alpha + \mathcal{R}\mathcal{R}^*)^{-1})^{-1} \mathcal{A}_G \\
&= (\alpha + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{A}_G (\mathcal{I}_G + \mathcal{A}_G^* (\alpha + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{A}_G)^{-1}.
\end{aligned}$$

This gives

$$\begin{aligned}
\mathcal{P}_G(\hat{f}_0) &= \mathcal{A}_G^* (\alpha + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{A}_G (\mathcal{I}_G + \mathcal{A}_G^* (\alpha + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{A}_G)^{-1} f^* \\
&= (\mathcal{I}_G - (\mathcal{I}_G + \mathcal{A}_G^* (\alpha + \mathcal{R}\mathcal{R}^*)^{-1} \mathcal{A}_G)^{-1}) f^*
\end{aligned}$$

and

$$\mathcal{P}_{G^\perp}(\hat{f}_0) = \mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G(\mathcal{I}_G + \mathcal{A}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}^*)^{-1}f^*.$$

We denote the actual bias and survival operators on G as

$$\begin{aligned}\mathcal{B} &= (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}, \text{ and} \\ \mathcal{S} &= \mathcal{I}_G - \mathcal{B}.\end{aligned}$$

We then have

$$\mathcal{P}_G(\hat{f}_0) = \mathcal{S}f^*,$$

and

$$\mathcal{P}_{G^\perp}(\hat{f}_0) = \mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\mathcal{B}f^*.$$

Clearly, $\|f^* - \mathcal{P}_G(\hat{f}_0)\|_{L_2} = \|\mathcal{B}f^*\|_{L_2} \leq \|\mathcal{B}\|_{\mathcal{H}\to L_2}\|f^*\|_{\mathcal{H}}$. To bound $\|\mathcal{P}_{G^\perp}(\hat{f}_0)\|_{L_2}$, note that (recalling $\mathcal{C} = \mathcal{A}_G$)

$$\left\|\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\right\|_{L_2} \leq \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2} \cdot \|\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\|_{\ell_2\to\mathcal{H}} \cdot \|\mathcal{C}\|_{L_2\to\ell_2}.$$

Note that $\|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2} = \|\mathcal{T}_{G^\perp}^{1/2}\|_{\mathcal{H}} = \sqrt{\lambda_{p+1}}$, and

$$\|\mathcal{C}\|_{L_2\to\ell_2}^2 = \|\mathcal{C}^*\mathcal{C}\|_{L_2} \approx \|n\mathcal{I}_G\|_{L_2} = n.$$

Furthermore, note that the singular values (from ℓ_2 to $\mathcal{H}$) of the operator $\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}$ are

$$\frac{\sqrt{\lambda_k(\mathcal{R}\mathcal{R}^*)}}{\alpha + \lambda_k(\mathcal{R}\mathcal{R}^*)} \leq \frac{1}{\sqrt{\alpha + \lambda_k(\mathcal{R}\mathcal{R}^*)}} \leq \frac{1}{\sqrt{\alpha_L}},\ k = 1, \ldots, n,$$

where $\lambda_k(S)$ denotes the kth eigenvalue of a symmetric matrix S. Therefore,

$$\left\|\mathcal{R}^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\right\|_{L_2} \lesssim \sqrt{\lambda_{p+1}} \cdot \frac{1}{\sqrt{\alpha_L}} \cdot \sqrt{n} = \sqrt{\frac{n\lambda_{p+1}}{\alpha_L}}.$$

Noting that $\bar{\alpha} \leq 2\alpha_L$, we have

$$\|\hat{f}_0 - f^*\|_{L_2} \lesssim \left(1 + \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\right)\|\mathcal{B}\|_{\mathcal{H}\to L_2}\|f^*\|_{\mathcal{H}}.$$

Lemma 31 gives

$$\|\mathcal{B}\|_{\mathcal{H}\to L_2} \leq \|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} + \|\mathcal{B} - \overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \leq \frac{1}{1-c}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2}.$$

Also, one can also easily check that $\|B\|_{\mathcal{H}} \leq 1$, and therefore $\|\mathcal{B}\|_{\mathcal{H}\to L_2} \leq \sqrt{\lambda_1}$. Thus

$$\|\mathcal{B}\|_{\mathcal{H}\to L_2} \leq \min\left\{\sqrt{\lambda_1}, \frac{1}{1-c}\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2}\right\}$$

Using the fact that $\|\overline{\mathcal{B}}\|_{\mathcal{H}\to L_2} \lesssim \min\left\{\sqrt{\frac{\bar{\alpha}}{n}}, \frac{\bar{\alpha}}{n\sqrt{\lambda_p}}\right\}$ completes the proof.

□

With the proof of Theorem 11 complete, recall that we introduced a more refined expression for the estimation error due to bias in Lemma 16 for the purpose of bounding classification error. Note that the proof of Lemma 16 is a very simple modification of the preceding proof. The error in $G^\perp$ is bounded the same way. For the error in G, we bound the norm of $(\mathcal{S} - \overline{\mathcal{S}})f^* = (\overline{\mathcal{B}} - \mathcal{B})f^*$ instead of $f^* - \mathcal{S}f^* = \mathcal{B}f^*$, and therefore we replace $\|\mathcal{B}\|_{\mathcal{H}\to L_2}$ by $\|\overline{\mathcal{B}} - \mathcal{B}\|_{\mathcal{H}\to L_2}$ in the bound.

C.2.2 Variance

Recall that $\alpha_L \preceq \alpha + \mathcal{R}\mathcal{R}^* \preceq \alpha_U$ and $\tilde{\alpha} = \frac{\alpha_U + \alpha_L}{2}$. Also recall the formula

$$\epsilon = \mathcal{A}^*(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\xi.$$

To allow us to replace $\alpha + \mathcal{A}\mathcal{A}^*$ with $\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*$, we need the following result:

Lemma 32.

$$\|(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{\ell_2} \leq \frac{1}{2}\left(\frac{\alpha_U}{\alpha_L} + 1\right).$$

Proof. Since $(\alpha + \mathcal{A}\mathcal{A}^*) - (\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*) = \alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha}$, another perturbation expansion (see Section C.2.1) gives

$$(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1} - (\alpha + \mathcal{A}\mathcal{A}^*)^{-1} = \sum_{k=1}^{\infty}(-1)^{k+1}(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\left[(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right]^k,$$

which is valid since $\alpha_L \preceq \alpha + \mathcal{R}\mathcal{R}^* \preceq \alpha_U$ implies

$$\left\|(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right\|_{\ell_2} \leq \frac{1}{\tilde{\alpha}} \cdot \frac{\alpha_U - \alpha_L}{2} = \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} < 1.$$

Then

$$I_n - (\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1} = \sum_{k=1}^{\infty}(-1)^{k+1}\left[(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right]^k.$$

We apply the triangle inequality to get

$$\begin{aligned}\|(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{\ell_2} &\leq \|I_n\|_{\ell_2} + \sum_{k=1}^{\infty}\left\|(\alpha + \mathcal{R}\mathcal{R}^* - \tilde{\alpha})(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\right\|_{\ell_2}^k \\ &\leq 1 + \sum_{i=1}^{\infty}\left(\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\right)^i \\ &= \frac{1}{2}\left(\frac{\alpha_U}{\alpha_L} + 1\right).\end{aligned}$$

□

We can now prove the main "variance" error bound:

Proof of Theorem 12. Since $\mathrm{var}(\xi_i) \leq \sigma^2$ for each i, we have

$$\begin{aligned}
\mathbf{E}_\xi \|\epsilon\|_{L_2}^2 &\leq \sigma^2 \|\mathcal{A}^*(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{HS,\ell_2\to L_2}^2 \\
&= \sigma^2 \|\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\|_{HS,\ell_2\to L_2}^2 \\
&\leq \frac{\sigma^2}{4}\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2 \|\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2,
\end{aligned}$$

where the last inequality substitutes Lemma 32. Furthermore, we have

$$\begin{aligned}
\|\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2 &= \|\mathcal{A}_G^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2 + \|\mathcal{R}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|_{HS,\ell_2\to L_2}^2 \\
&\leq \|(\tilde{\alpha} + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{A}_G^*\|_{HS,\ell_2\to L_2}^2 + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2} \\
&= \|(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-1}\mathcal{C}^*\|_{HS,\ell_2\to L_2}^2 + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2} \\
&\lesssim \frac{p}{n} + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2},
\end{aligned}$$

where the last inequality is due to the fact that $\mathcal{C}$ is an $n \times p$-dimensional operator, all of whose singular values are close to $\sqrt{n}$.

Therefore,

$$\mathbf{E}_\xi \|\epsilon\|_{L_2}^2 \lesssim \sigma^2\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2\left(\frac{p}{n} + \frac{\mathrm{tr}_{L_2}(\mathcal{R}^*\mathcal{R})}{\tilde{\alpha}^2}\right).$$

□

High-probability Noise Bounds

If the ξ_i's are sub-Gaussian, we could use the Hanson-Wright inequality for sub-Gaussian random vectors (see, e.g., [170]) to get a high-probability bound in Theorem 12,

Note that we can write

$$\|\epsilon\|_{L_2}^2 = \langle Z\xi, \xi\rangle_{\ell_2},$$

where

$$Z = (\alpha + \mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\mathcal{T}\mathcal{A}^*(\alpha + \mathcal{A}\mathcal{A}^*)^{-1}.$$

We have already calculated an upper bound on the expectation of this quadratic form. To use the Hanson-Wright inequality to bound the upper tail, we need to bound both $\|Z\|_{\ell_2}$ and $\|Z\|_{\mathrm{HS}}$ (where $\|Z\|_{\mathrm{HS}}$ is the Hilbert-Schmidt norm with respect to the Euclidean inner product, also known as the Frobenius norm). By a similar argument as before, we have

$$\|Z\|_{\ell_2} \leq \frac{1}{4}\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2 \|\widetilde{Z}\|_{\ell_2},$$

and

$$\|Z\|_{\mathrm{HS}} \leq \frac{1}{4}\left(\frac{\alpha_U}{\alpha_L} + 1\right)^2 \|\widetilde{Z}\|_{\mathrm{HS}},$$

where

$$\begin{aligned}\widetilde{Z} &= (\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}\mathcal{T}\mathcal{A}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1} \\ &= \underbrace{(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G\mathcal{T}_G\mathcal{A}_G^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}}_{\widetilde{Z}_G} + \underbrace{(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{R}\mathcal{T}_{G^\perp}\mathcal{R}^*(\tilde{\alpha} + \mathcal{A}_G\mathcal{A}_G^*)^{-1}}_{\widetilde{Z}_{G^\perp}}.\end{aligned}$$

Note that

$$\widetilde{Z}_G = \mathcal{A}_G(\tilde{\alpha} + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{T}_G(\tilde{\alpha} + \mathcal{A}_G^*\mathcal{A}_G)^{-1}\mathcal{A}_G^* = \mathcal{C}(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-2}\mathcal{C}^*.$$

By a similar argument as before (in which we were effectively calculating the trace of $\widetilde{Z}_G$), we have $\|\widetilde{Z}_G\|_{\ell_2} \lesssim \frac{1}{n}$ and $\|\widetilde{Z}_G\|_{\mathrm{HS}} \lesssim \frac{\sqrt{p}}{n}$.

Similarly, $\|\widetilde{Z}_{G^\perp}\|_{\ell_2} \leq \frac{1}{\tilde{\alpha}^2}\|\mathcal{R}\mathcal{T}_{G^\perp}\mathcal{R}^*\|_{\ell_2}$, and $\|\widetilde{Z}_{G^\perp}\|_{\mathrm{HS}} \leq \frac{1}{\tilde{\alpha}^2}\|\mathcal{R}\mathcal{T}_{G^\perp}\mathcal{R}^*\|_{\mathrm{HS}}$.

C.3 Proofs of operator concentration results

Proof of Lemma 11. Let $\mathrm{diag}(Z)$ denote the projection of Z onto the space of diagonal matrices, and let $\mathrm{diag}^{\perp}(Z)$ denote the orthogonal projection (i.e., onto the space of matrices with zero diagonal). Note that

$$\begin{aligned}\|\mathcal{R}\mathcal{R}^* - (\operatorname{tr}\mathcal{T}_{G^\perp})I_n\| &\leq \|\mathrm{diag}^{\perp}(\mathcal{R}\mathcal{R}^*)\| + \|\mathrm{diag}(\mathcal{R}\mathcal{R}^*) - (\operatorname{tr}\mathcal{T}_{G^\perp})I_n\| \\ &\leq \|\mathrm{diag}^{\perp}(\mathcal{R}\mathcal{R}^*)\|_{\mathrm{HS}} + \max_i \ |k^R(x_i,x_i) - \operatorname{tr}\mathcal{T}_{G^\perp}| \\ &\leq \sqrt{\sum_{i\neq j}(k^R(x_i,x_j))^2} + \|k^R(\cdot,\cdot) - \operatorname{tr}\mathcal{T}_{G^\perp}\|_\infty.\end{aligned}$$

Squaring, taking an expectation, and noting that

$$\mathbf{E}_{x,y\overset{\text{i.i.d.}}{\sim}\mu}(k^R(x,y))^2 = \operatorname{tr}(\mathcal{T}_{G^\perp}^2)$$

completes the proof. □

Proof of Lemma 12. We have

$$\begin{aligned}\operatorname{tr}_{L_2}(\mathcal{R}^*\mathcal{R}) &= \sum_{i=1}^n \operatorname{tr}_{L_2}(k^R_{x_i}\otimes k^R_{x_i}) \\ &= \sum_{i=1}^n \|k^R_{x_i}\|^2_{L_2} \\ &= \sum_{i=1}^n \sum_{\ell>p} \lambda_\ell^2 v_\ell^2(x_i).\end{aligned}$$

Taking an expectation completes the proof. □

Proof of Lemma 13. We can write the operator $\mathcal{C}^*\mathcal{C}$ as a sum of independent random opera-

tors:

$$\mathcal{C}^*\mathcal{C} = \sum_{i=1}^{n} z(x_i) \otimes z(x_i),$$

where

$$z(x) := \sum_{\ell=1}^{p} v_\ell(x) v_\ell.$$

Note that the BOS condition implies $\|z(x)\|_{L_2}^2 \le Cp$ almost surely in x. We also have $\mathbf{E}\, z(x) \otimes z(x) = \mathcal{I}_G$ for $x \sim \mu$.

We use a matrix Bernstein inequality [49, Theorem 6.6.1] to analyze the zero-mean sum

$$\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G = \sum_{i=1}^{p} (z(x_i) \otimes z(x_i) - \mathbf{E}\, z(x_i) \otimes z(x_i)).$$

Writing $X_i = z(x_i) \otimes z(x_i) - \mathbf{E}\, z(x_i) \otimes z(x_i)$, we have $\|X_i\|_{L_2} \le Cp$ almost surely, and

$$\mathbf{E}\, X_i^2 \preceq \mathbf{E}(z(x_i) \otimes z(x_i))^2 = \mathbf{E}\|z(x_i)\|_{L_2}^2 z(x_i) \otimes z(x_i) \preceq Cp\, \mathbf{E}\, z(x_i) \otimes z(x_i) = Cp\mathcal{I}_G.$$

The Bernstein inequality then gives that for any $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} = \left\| \sum_{i=1}^{n} X_i \right\|_{L_2} \lesssim \sqrt{Cpn(t + \log p)} + Cp(t + \log p).$$

□

Proof of Lemma 14. For $z \in \mathbf{R}^n$, we have

$$\langle \mathcal{R}\mathcal{R}^* z, z \rangle = \sum_{\ell > p} \lambda_\ell \langle w_\ell, z \rangle^2.$$

By our assumptions, this is the sum of independent random variables.

If $\|z\|_{\ell_2} = 1$, then, for each ℓ, $\langle w_\ell, z \rangle^2$ is sub-exponential (as the square of a sub-Gaussian variable; since the sub-Gaussian norm is bounded, so is the sub-exponential norm), $\mathbf{E}\langle w_\ell, z \rangle^2 = 1$, and $\mathbf{E}\langle w_\ell, z \rangle^4 \lesssim 1$.

Note that in this case, $\mathbf{E}\langle \mathcal{R}\mathcal{R}^* z, z\rangle = \operatorname{tr} \mathcal{T}_{G^\perp} = \langle (\operatorname{tr} \mathcal{T}_{G^\perp}) I_n z, z\rangle$, and

$$\begin{aligned}\mathbf{E}(\langle \mathcal{R}\mathcal{R}^* z, z\rangle - \mathbf{E}\langle \mathcal{R}\mathcal{R}^* z, z\rangle)^2 &= \sum_{\ell > p} \lambda_\ell^2 \, \mathbf{E}(\langle w_\ell, z\rangle^2 - \mathbf{E}\langle w_\ell, z\rangle^2)^2 \\ &\lesssim \sum_{\ell > p} \lambda_\ell^2.\end{aligned}$$

A Bernstein inequality then implies that for $t > 0$, with probability at least $1 - e^{-t}$, we have

$$|\langle \mathcal{R}\mathcal{R}^* z, z\rangle - \operatorname{tr} \mathcal{T}_{G^\perp}| \lesssim \sqrt{\left(\sum_{\ell > p} \lambda_\ell^2\right) t} + \lambda_{p+1} t.$$

By a standard covering argument (e.g., [53, Exercise 4.4.3]), we then obtain, with probability at least $1 - e^{-t}$,

$$\max_{z \in S^{n-1}} |\langle \mathcal{R}\mathcal{R}^* z, z\rangle - \operatorname{tr} \mathcal{T}_{G^\perp}| \lesssim \sqrt{\left(\sum_{\ell > p} \lambda_\ell^2\right)(n+t)} + \lambda_{p+1}(n+t),$$

where S^{n-1} is the unit sphere in $\mathbf{R}^n$. □

C.4 Tightness of general feature results

With no independence assumptions on the features $\{v_\ell(x)\}_\ell$, our general results require $d \gtrsim n^2$ in order to upper and lower bound the residual Gram matrix $\mathcal{R}\mathcal{R}^*$ by constant multiples of the identity. The following theorem shows that for Fourier features, $d \gtrsim n^2$ is in fact a necessary condition, i.e. if $d = o(n^2)$, then the condition number of $\mathcal{R}\mathcal{R}^*$ grows as $n \to \infty$.

Theorem 13. *Consider the case of Fourier features with bi-level eigenvalues, i.e.* $v_\ell \in L_2([0,1])$ *for* $\ell = -d, \ldots, d$*, which are defined by* $v_\ell(x) = e^{\mathbf{j} 2\pi \ell x}$ *for* $x \in [0,1]$*, and* $\lambda_\ell = 1$ *for* $|\ell| \le p$*,* $\lambda_\ell = \gamma \in (0,1)$ *for* $p < |\ell| \le d$*. Then, for any constant* $\tau > 0$*, the residual*

Gram matrix $\mathcal{R}\mathcal{R}^*$ satisfies

$$\frac{\lambda_{max}(\mathcal{R}\mathcal{R}^*)}{\lambda_{min}(\mathcal{R}\mathcal{R}^*)} \gtrsim \frac{n^4}{\tau^2 d^2}$$

with probability at least $1 - e^{-\tau}$.

Intuitively, if there exist distinct indices $i, i' = 1, \ldots, n$ such that x_i and $x_{i'}$ are very close together, then the i-th and i'-th columns (and rows) of $\mathcal{R}\mathcal{R}^*$ are nearly identical, and thus, $\mathcal{R}\mathcal{R}^*$ is nearly rank-deficient. We now make this argument rigorous.

Proof. First, pick any two indices $i, i' \in \{1, \ldots, n\}$ with $i \neq i'$ and consider the 2×2 submatrix of $\mathcal{R}\mathcal{R}^*$ formed by the i-th and i'-th rows and columns, i.e,

$$(\mathcal{R}\mathcal{R}^*)_{\text{sub}} := \begin{bmatrix} k^R(x_i, x_i) & k^R(x_i, x_{i'}) \\ k^R(x_{i'}, x_i) & k^R(x_{i'}, x_{i'}) \end{bmatrix}.$$

The kernel restricted to $G^\perp$ is given by

$$\begin{aligned} k^R(x, y) &= \sum_{p < |\ell| \leq d} \lambda_\ell v_\ell(x) \overline{v_\ell(y)} \\ &= \sum_{p < |\ell| \leq d} \gamma e^{\mathrm{j} 2\pi \ell (x-y)} \\ &= \gamma \frac{\sin[(2d+1)\pi(x-y)] - \sin[(2p+1)\pi(x-y)]}{\sin[\pi(x-y)]}. \end{aligned}$$

Hence, $k^R(x_i, x_i) = k^R(x_{i'}, x_{i'}) = 2(d-p)\gamma$.

Furthermore, using the inequality $2\cos\theta \geq 2 - \theta^2$ for $\theta \in \mathbf{R}$, we have

$$\begin{aligned}
\frac{\sin[(2d+1)\pi t] - \sin[(2p+1)\pi t]}{\sin[\pi t]} &= \sum_{p<|\ell|\leq d} e^{\mathbf{j}2\pi\ell t} \\
&= \sum_{\ell=p+1}^{d} 2\cos(2\pi\ell t) \\
&\geq \sum_{\ell=p+1}^{d} \left[2 - (2\pi\ell t)^2\right] \\
&= 2(d-p) - 4\pi^2 \left(\sum_{\ell=p+1}^{d} \ell^2\right) t^2 \\
&\geq 2(d-p) - 4\pi^2 d^2 (d-p) t^2
\end{aligned}$$

for all $t \in \mathbf{R}$, and thus,

$$\begin{aligned}
k^R(x_i, x_{i'}) = k^R(x_{i'}, x_i) &= \gamma \frac{\sin[(2d+1)\pi(x_i - x_{i'})] - \sin[(2p+1)\pi(x_i - x_{i'})]}{\sin[\pi(x_i - x_{i'})]} \\
&\geq 2(d-p)\gamma - 4\pi^2 d^2 (d-p)\gamma (x_i - x_{i'})^2.
\end{aligned}$$

We can then bound the smallest eigenvalue of $\mathcal{R}\mathcal{R}^*$ by

$$\lambda_{\min}(\mathcal{R}\mathcal{R}^*) \leq \lambda_{\min}((\mathcal{R}\mathcal{R}^*)_{\text{sub}}) = k^R(x_i, x_i) - k^R(x_i, x_{i'}) \leq 4\pi^2 d^2 (d-p)\gamma (x_i - x_{i'})^2.$$

Then, by using the trivial bound $\lambda_{\max}(\mathcal{R}\mathcal{R}^*) \geq \frac{1}{n}\operatorname{tr}(\mathcal{R}\mathcal{R}^*) = \frac{1}{n} \cdot 2(d-p)\gamma n = 2(d-p)\gamma$, we have

$$\frac{\lambda_{\max}(\mathcal{R}\mathcal{R}^*)}{\lambda_{\min}(\mathcal{R}\mathcal{R}^*)} \geq \frac{2(d-p)\gamma}{4\pi^2 d^2 (d-p)\gamma (x_i - x_{i'})^2} = \frac{1}{2\pi^2 d^2 (x_i - x_{i'})^2}.$$

This bound holds for any distinct indices $i \neq i'$. A relatively straightforward calculation[1]

[1]Thanks to Hans's answer at https://mathoverflow.net/questions/1294

shows that if $x_1, \ldots, x_n$ are i.i.d. Uniform$[0,1]$, then for any $\delta \in (0, \frac{1}{n-1})$,

$$\begin{aligned}
\mathrm{P}\left\{|x_i - x_{i'}| \geq \delta \text{ for all } i \neq i'\right\} &= n!\,\mathrm{P}\left\{x_{i-1} + \delta \leq x_i \text{ for all } i = 2, \ldots, n\right\} \\
&= n! \int \cdots \int_{\{0 \leq x_1,\ x_{i-1}+\delta \leq x_i \text{ for } i=2,\ldots,n,\ x_n \leq 1\}} dx_1 \cdots dx_n \\
&= n! \int \cdots \int_{\{y_i \geq 0 \text{ for } i=1,\ldots,n,\ y_1+\cdots+y_n \leq 1-(n-1)\delta\}} dy_1 \cdots dy_n \\
&= n! \cdot \frac{1}{n!}(1-(n-1)\delta)^n \\
&= (1-(n-1)\delta)^n
\end{aligned}$$

where we made the change of variable $y_1 = x_1$ and $y_i = x_i - x_{i-1} - \delta$ for $i = 2, \ldots, n$, and we used the fact that the volume of the standard n-simplex is $\frac{1}{n!}$. Hence, if $0 < \delta < \frac{1}{n-1}$, the probability that $|x_i - x_{i'}| \leq \delta$ for some indices $i \neq i'$ is $1 - (1-(n-1)\delta)^n$.

If $0 < \tau < n$, we can apply this result for $\delta = \frac{\tau}{n(n-1)}$, to obtain that with probability $1 - (1 - \frac{\tau}{n})^n \geq 1 - e^{-\tau}$ there exist $i \neq i'$ such that $|x_i - x_{i'}| \leq \frac{\tau}{n(n-1)}$, and thus,

$$\frac{\lambda_{\max}(\mathcal{R}\mathcal{R}^*)}{\lambda_{\min}(\mathcal{R}\mathcal{R}^*)} \geq \frac{1}{2\pi^2 d^2 (x_i - x_{i'})^2} \geq \frac{n^2(n-1)^2}{2\pi^2 d^2 \tau^2} \gtrsim \frac{n^4}{\tau^2 d^2}.$$

If $\tau \geq n$, then it is guaranteed that there exist two indices $i \neq i'$ which satisfy $|x_i - x_{i'}| \leq \frac{1}{n-1} \leq \frac{\tau}{n(n-1)}$, and the same bound holds. □

C.5 Proof of bi-level ensemble asymptotic results

If $\beta > 2$ and $r < 1$, the concentration results Lemmas 11 and 13 will hold as n becomes large, since we will have $n \gg p \log p$ and $d - p \approx n^\beta \gg n^2$. We now apply Lemma 16 and Theorem 12 to the bi-level ensemble. Since we are in the interpolating regime, we take $\alpha = 0$; then, $\alpha_L = \lambda_{\min}(\mathcal{R}\mathcal{R}^*)$ and $\alpha_U = \lambda_{\max}(\mathcal{R}\mathcal{R}^*)$ are the smallest and largest eigenvalues of $\mathcal{R}\mathcal{R}^*$. As long as α_L and α_U are close together (which we will analyze next),

we will have

$$\tilde{\alpha} \approx \bar{\alpha} \approx \sum_{\ell > p} \lambda_\ell \approx n^\beta \cdot n^{-(\beta - r - q)} = n^{r+q}.$$

Furthermore,

$$\sum_{\ell > p} \lambda_\ell^2 \approx n^\beta n^{-2(\beta - q - r)} = n^{2q+2r-\beta}.$$

Applying these scalings to Theorem 12 gives us

$$\mathbf{E}_\xi \|\epsilon\|_{L_2}^2 \lesssim n^{r-1} + \frac{n}{n^{2(r+q)}} n^{2q+2r-\beta} = n^{r-1} + n^{1-\beta}.$$

To bound the bias, note that combining the above calculations with Lemma 11 gives

$$\begin{aligned}
\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} &\lesssim \frac{1}{\bar{\alpha}} \sqrt{n^2 \sum_{\ell > p} \lambda_\ell^2} \\
&\approx \frac{1}{n^{r+q}} \sqrt{n^2 n^{2q+2r-\beta}} \\
&= n^{1-\beta/2}.
\end{aligned}$$

Combining this with Lemma 13, the quantity c in Theorem 11 and Lemma 16 can be bounded as

$$c \lesssim n^{1-\beta/2} + n^{(r-1)/2} \sqrt{\log n}.$$

Then Lemma 16 gives

$$\begin{aligned}
&\frac{\|\hat{\eta}_0 - \overline{\mathcal{S}} \eta^*\|_{L_2}}{\|\eta^*\|_{\mathcal{H}}} \\
&\quad \lesssim \left(n^{1-\beta/2} + n^{(r-1)/2} \sqrt{\log n} + \sqrt{\frac{n^{-(\beta-r-q)} n}{n^{r+q}}} \right) \cdot \min\left\{1, n^{r+q-1}, n^{(r+q-1)/2}\right\} \\
&\quad \lesssim \left(n^{1-\beta/2} + n^{(r-1)/2} \sqrt{\log n} \right) \cdot \min\{1, n^{r+q-1}\}.
\end{aligned}$$

Recall from (Equation 5.3) that excess classification risk has upper bound $\mathcal{E} \leq \frac{\|\hat{\eta}_r\|_{L_2}}{s}$ for any decomposition $\hat{\eta}_0 = s\eta^* + \hat{\eta}_r$ with an $s > 0$ that we can choose. We will now

characterize the terms s and $\|\hat{\eta}_r\|_{L_2}$, beginning with the factor s. The ideal survival operator is given by

$$\overline{\mathcal{S}} = \mathcal{I}_G - \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1} = \frac{1}{1+\bar{\alpha}n}\mathcal{I}_G \approx \frac{1}{1+n^{r+q-1}}\mathcal{I}_G.$$

Then, we can decompose

$$\hat{\eta} = \overline{\mathcal{S}}\eta^* + \hat{\eta}_r \approx \frac{1}{1+n^{r+q-1}}\eta^* + \hat{\eta}_r,$$

where $\hat{\eta}_r = \epsilon + \hat{\eta}_0 - \overline{\mathcal{S}}\eta^*$. This gives us $s \approx \frac{1}{1+n^{r+q-1}}$.

Next, we bound $\|\hat{\eta}_r\|_{L_2}$. We have

$$\|\hat{\eta}_r\|_{L_2} \lesssim n^{(r-1)/2} + n^{(1-\beta)/2} + \left(n^{1-\beta/2} + n^{(r-1)/2}\sqrt{\log n}\right) \cdot \min\{1, n^{r+q-1}\}\|\eta^*\|_{L_2}.$$

Above, we used the fact that $\|\eta^*\|_{L_2} = \|\eta^*\|_{\mathcal{H}}$.

There are several cases to consider (recall that we are already assuming $\beta > 2$ and $r < 1$):

1. $q < 1 - r$: In this case, $s \approx \frac{1}{1+n^{r+q-1}} \to 1$, and $\|\hat{\eta}_r\|_{L_2} \to 0$. Thus both the excess regression and classification risk converge to 0 as $n \to \infty$.

2. If $q > 1 - r$, we have $s \to 0$ and $\|\hat{\eta}_r\|_{L_2} \to 0$, so $\|\hat{\eta}\|_{L_2} \to 0$. Therefore will will *not* get regression consistency (for nonzero η^*).

3. If $1 - r < q < \frac{3}{2}(1 - r)$ and $\beta > 2r + 2q$, then $s \to 0$, but $\frac{\|\hat{\eta}_r\|_{L_2}}{s} \approx \|\hat{\eta}_r\|_{L_2} \cdot (1 + n^{r+q-1}) \to 0$, so the excess classification risk converges to zero as $n \to \infty$ even though the regression risk does not.

4. If $1 - r < q < \frac{3}{2}(1 - r)$ but $\beta < 2r + 2q$ or if $q > \frac{3}{2}(1 - r)$, our analysis does not yield any convergence results. It is an interesting and important direction for future

work to characterize precisely what relations between the parameters q, r, β are both sufficient and necessary for classification risk to go to 0 as $n \to \infty$.

C.6 Distortion analysis

In this section, we analyze more carefully the regularization-induced distortion. In particular, we consider how different the (deterministic) ideal survival operator $\overline{\mathcal{S}}$ is from a multiple of the identity. Recall that

$$\overline{\mathcal{S}} = \mathcal{I}_G - \Big(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\Big)^{-1} = \frac{n}{\bar{\alpha}}\mathcal{T}_G\Big(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\Big)^{-1}.$$

We want to solve

$$\begin{aligned}\arg\min_{s>0} \ \big\|s\mathcal{I}_G - \overline{\mathcal{S}}\big\|_{\mathcal{H}\to L_2} &= \arg\min_{s>0} \ \Big\|s\mathcal{T}_G^{1/2} - \mathcal{T}_G^{1/2}\overline{\mathcal{S}}\Big\|_{L_2} \\ &= \arg\min_{s>0} \ \max_{1\le\ell\le p} \sqrt{\lambda_\ell}\left|s - \frac{\lambda_\ell}{\lambda_\ell + \frac{\bar{\alpha}}{n}}\right|.\end{aligned}$$

We abbreviate $b \coloneqq \frac{\bar{\alpha}}{n}$. The objective function in s is convex as the maximum of convex functions. Some convex analysis tell us that there must be (at least) two distinct $i, j \in \{1, \dots, p\}$ such that, for s at its optimal value s^*, both i and j achieve the maximum over ℓ, and the arguments to the absolute value have different signs. Assuming, without loss of generality, that $\lambda_j > \lambda_i$, this implies

$$\begin{aligned}\arg\min_{s>0} \ \max_{1\le\ell\le p} \sqrt{\lambda_\ell}\left|s - \frac{\lambda_\ell}{\lambda_\ell + \frac{\bar{\alpha}}{n}}\right| &= \sqrt{\lambda_i}\left(s^* - \frac{\lambda_i}{\lambda_i + b}\right) \\ &= \sqrt{\lambda_j}\left(\frac{\lambda_j}{\lambda_j + b} - s^*\right).\end{aligned}$$

Note that the last expression is increasing in λ_j, so we can take $j = 1$. Solving for s^* gives

$$s^* = \frac{\lambda_i\lambda_1 + b(\lambda_i + \lambda_1 - \sqrt{\lambda_i\lambda_1})}{(b + \lambda_i)(b + \lambda_1)}.$$

Plugging this into the objective function gives

$$\|s^*\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2} = \max_i \frac{b\sqrt{\lambda_i\lambda_1}(\sqrt{\lambda_1} - \sqrt{\lambda_i})}{(b+\lambda_1)(b+\lambda_i)}.$$

One can check that if $\lambda_p \geq \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$, this minimum is achieved for $i = p$. Otherwise, we can find an upper bound by optimizing over continuous λ:

$$\begin{aligned}\max_i \frac{b\sqrt{\lambda_i\lambda_1}(\sqrt{\lambda_1} - \sqrt{\lambda_i})}{(b+\lambda_1)(b+\lambda_i)} &\leq \max_{\lambda\geq 0} \frac{b\sqrt{\lambda\lambda_1}(\sqrt{\lambda_1} - \sqrt{\lambda})}{(b+\lambda_1)(b+\lambda)} \\ &= \frac{b\lambda_1^{3/2}}{2(b+\lambda_1)(b+\sqrt{b(b+\lambda_1)})},\end{aligned}$$

where the minimum is achieved at $\lambda = \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$.

Whatever value of λ we use, we then have, for the corresponding choice of s,

$$\frac{\|s\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s} = \frac{b\sqrt{\lambda_1\lambda}(\sqrt{\lambda_1} - \sqrt{\lambda})}{\lambda_1\lambda + b(\lambda_1 + \lambda - \sqrt{\lambda_1\lambda})}.$$

For $\lambda = \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$, we get

$$\frac{\|s\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s} \leq \frac{\sqrt{b\lambda_1(b+\lambda_1)}}{2b + 2\lambda_1 + \sqrt{b(b+\lambda_1)}}.$$

If $\frac{\bar{\alpha}}{n} = b \gtrsim \lambda_1$, then this last bound is approximately $\sqrt{\lambda_1} \approx \|\mathcal{I}_G\|_{\mathcal{H}\to L_2}$, so there appears to be little hope of getting small classification error from this bound.

Alternatively, if $\frac{\bar{\alpha}}{n} \ll \lambda_1$, we get

$$\frac{\|s\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s} \lesssim \sqrt{\frac{\bar{\alpha}}{n}}.$$

However, recall that the *regression* error is of the same order, so this analysis does not significantly improve our classification risk.

Therefore, the only regime in which we gain anything over the regression analysis is when $\lambda_p > \frac{\lambda_1}{\left(1+\sqrt{1+\frac{\lambda_1}{b}}\right)^2}$.

If $b \gtrsim \lambda_1$, then this constraint implies that λ_p/λ_1 is not very small. Furthermore,

$$\frac{\|s^*\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s^*} \approx \sqrt{\frac{\lambda_p}{\lambda_1}}(\sqrt{\lambda_1} - \sqrt{\lambda_p}).$$

Since λ_p is not too small, this ratio is only small when λ_1 and λ_p are very close together.

If $b \lesssim \lambda_1$, the constraint implies $\lambda_p \gtrsim b$. Then

$$\frac{\|s^*\mathcal{I}_G - \overline{\mathcal{S}}\|_{\mathcal{H}\to L_2}}{s^*} \approx \frac{b}{\sqrt{\lambda_1\lambda_p}}(\sqrt{\lambda_1} - \sqrt{\lambda_p}).$$

This is better than the previous case when b is small, and it improves over the regression error bound when λ_1 and λ_p are close. However, note that in this case we get $c \gtrsim 1$, so unless λ_p is very close to λ_1, there is no significant improvement over regression error.

REFERENCES

[1] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York: Birkhäuser, 2013.

[2] A. D. McRae and M. A. Davenport, "Low-rank matrix completion and denoising under Poisson noise," *Inform. Inference.*, vol. 10, no. 2, pp. 697–720, 2021. arXiv: 1907.05325 [stat.ML].

[3] S. Negahban and M. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, pp. 1665–1697, 2012.

[4] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Stat.*, vol. 39, no. 5, pp. 2302–2329, 2011.

[5] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[6] X. Zhou, C. Yang, H. Zhao, and W. Yu, "Low-rank modeling and its applications in image analysis," *ACM Comput. Surv.*, vol. 47, no. 2, 36:1–36:33, Dec. 2014.

[7] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. ACM SIGIR Conf. (SIGIR)*, Berkeley, CA, Aug. 1999.

[8] D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[9] J. Canny, "GaP: A factor model for discrete data," in *Proc. ACM SIGIR Conf. (SIGIR)*, Sheffield, United Kingdom, Jul. 2004.

[10] A. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, 2009.

[11] A. Mnih and R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2007.

[12] H. Ma, C. Liu, I. King, and M. Lyu, "Probabilistic factor models for web site recommendation," in *Proc. ACM SIGIR Conf. (SIGIR)*, Beijing, China, Jul. 2011.

[13] P. Gopalan, J. Hofman, and D. Blei, "Scalable recommendation with hierarchical poisson factorization," in *Proc. Conf. Uncert. in Artif. Intell. (UAI)*, Amsterdam, Netherlands, Jul. 2015.

[14] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[15] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution," *Found. Comput. Math.*, vol. 20, pp. 451–632, 2020.

[16] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan, "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization," *SIAM J. Optim.*, vol. 30, no. 4, pp. 3098–3121, 2020.

[17] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, 2010.

[18] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, 2010.

[19] Y. Cao and Y. Xie, "Poisson matrix recovery and completion," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1609–1620, Mar. 2016.

[20] A. Soni, S. Jain, J. Haupt, and S. Gonella, "Noisy matrix completion under sparse factor models," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3636–3661, Jun. 2016.

[21] A. Soni and J. Haupt, "Estimation error guarantees for Poisson denoising with sparse and structured dictionary models," in *Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, Honolulu, HI, Jun.–Jul. 2014.

[22] O. Klopp, "Matrix completion by singular value thresholding: Sharp bounds," *Electron. J. Stat.*, vol. 9, pp. 2348–2369, 2015.

[23] ——, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[24] S. Gunasekar, A. Banerjee, and J. Ghosh, "Unified view of matrix completion under general structural constraints," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, Canada, Dec. 2015.

[25] S. Gaïffas and G. Lecué, "Sharp oracle inequalities for high-dimensional matrix prediction," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6942–6957, Oct. 2011.

[26] O. Klopp, "Rank penalized estimators for high-dimensional matrices," *Electron. J. Stat.*, vol. 5, pp. 1161–1183, 2011.

[27] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *Ann. Stat.*, vol. 43, no. 1, pp. 177–214, 2015.

[28] S. Gaïffas and O. Klopp, "High dimensional matrix estimation with unknown variance of the noise," *Stat. Sin.*, vol. 27, pp. 115–145, 2017.

[29] J. Lafond, "Low rank matrix completion with exponential family noise," in *Proc. Conf. Learn. Theory (COLT)*, Paris, France, Jul. 2015.

[30] S. Gunasekar, P. Ravikumar, and J. Ghosh, "Exponential family matrix completion under structural constraints," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Beijing, China, Jun. 2014.

[31] G. Robin, H.-T. Wai, J. Josse, O. Klopp, and É. Moulines, "Low-rank interaction with sparse additive effects model for large data frames," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, Dec. 2018.

[32] M. Collins, S. Dasgupta, and R. Schapire, "A generalization of principal component analysis to the exponential family," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2001.

[33] J. Chiquet, M. Mariadassou, and S. Robin, "Variational inference for probabilistic Poisson PCA," *Ann. Appl. Stat.*, vol. 12, no. 4, pp. 2674–2698, 2018.

[34] L. Liu, E. Dobriban, and A. Singer, "ePCA: High dimensional exponential family PCA," *Ann. Appl. Stat.*, vol. 12, no. 4, pp. 2121–2150, 2018.

[35] T. Kenney, T. Huang, and H. Gu, "Poisson pca: Poisson measurement error corrected pca, with application to microbiome data," Apr. 26, 2019. arXiv: 1904.11745 [stat.ME].

[36] E. Dobriban, W. Leeb, and A. Singer, "Optimal prediction in the linearly transformed spiked model," *Ann. Stat.*, vol. 48, no. 1, pp. 491–513, 2020.

[37] X. Chen and J. Storey, "Consistent estimation of low-dimensional latent structure in high-dimensional data," Oct. 13, 2015. arXiv: 1510.03497 [stat.ML].

[38] G. Raskutti, M. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.

[39] A. S. Bandeira and R. van Handel, "Sharp nonasymptotic bounds on the norm of random matrices with independent entries," *Ann. Probab.*, vol. 44, no. 4, pp. 2479–2506, 2016.

[40] A. V. Sambasivan and J. D. Haupt, "Minimax lower bounds for noisy matrix completion under sparse factor models," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3274–3285, 2018.

[41] Q. Huang, S. Kakade, W. Kong, and G. Valiant, "Recovering structured probability matrices," in *Proc. Innovations Theor. Comput. Sci. (ITCS)*, Cambridge, MA, Jan. 2018.

[42] S. Arora *et al.*, "A practical algorithm for topic modeling with provable guarantees," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, Jun. 2013.

[43] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, pp. 2773–2832, 2014.

[44] T. Bansal, C. Bhattacharyya, and R. Kannan, "A provable SVD-based algorithm for learning topics in dominant admixture corpus," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, Montréal, Canada, Dec. 2014.

[45] Z. T. Ke and M. Wang, "A new svd approach to optimal topic estimation," 2017. arXiv: 1704.07016 [stat.ME].

[46] X. Bing, F. Bunea, and M. Wegkamp, "A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics," 2018. arXiv: 1805.06837 [stat.ML].

[47] R. Latała, R. van Handel, and P. Youssef, "The dimension-free structure of nonhomogeneous random matrices," *Invent. Math.*, vol. 214, pp. 1031–1080, 2018.

[48] D. Pollard, "MiniEmpirical," 2016.

[49] J. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, no. 1-2, pp. 1–230, 2015.

[50] C. Huber, "Lower bounds for function estimation," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. L. Yang, Eds. New York: Springer, 1997, ch. 15, pp. 245–258.

[51] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. L. Yang, Eds. New York: Springer, 1997, ch. 29, pp. 423–435.

[52] A. D. McRae, J. Romberg, and M. A. Davenport, "Optimal convex lifted sparse phase retrieval and PCA with an atomic matrix norm regularizer," 2021. arXiv: 2111.04652 [math.ST].

[53] R. Vershynin, *High-Dimensional Probability, An Introduction with Applications in Data Science*. Cambridge, 2018, 296 pp., ISBN: 1108415199.

[54] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2012.

[55] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[56] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 14, no. 5, pp. 1017–1026, 2013.

[57] C. Thrampoulidis and A. S. Rawat, "Lifting high-dimensional non-linear models with Gaussian regressors," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Naha, Okinawa, Japan, Apr. 2019, pp. 3206–3215.

[58] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Lake Tahoe, Utah, 2013.

[59] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *Ann. Stat.*, vol. 44, no. 5, pp. 2221–2251, 2016.

[60] G. Wang, L. Zhang, G. B. Giannakis, M. Akcakaya, and J. Chen, "Sparse phase retrieval via truncated amplitude flow," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 479–491, 2018.

[61] Z. Yuan, H. Wang, and Q. Wang, "Phase retrieval via sparse Wirtinger flow," *J. Comput. Appl. Math.*, vol. 355, pp. 162–173, 2019.

[62] Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov, "Misspecified nonconvex statistical optimization for sparse phase retrieval," *Math. Program.*, vol. 176, no. 1-2, pp. 545–571, 2019.

[63] G. Jagatap and C. Hegde, "Sample-efficient algorithms for recovering structured signals from magnitude-only measurements," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4434–4456, 2019.

[64] M. Soltanolkotabi, "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2374–2400, 2019.

[65] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1043–1055, 2015.

[66] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM J. Math. Anal.*, vol. 45, no. 5, pp. 3019–3033, 2013.

[67] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," in *Proc. IFAC Symp. System Identif.*, vol. 16, Brussels, Belgium, Jul. 2012, pp. 89–94.

[68] V. Koltchinskii and K. Lounici, "Concentration inequalities and moment bounds for sample covariance operators," *Bernoulli*, vol. 23, no. 1, pp. 110–133, 2017.

[69] V. Q. Vu and J. Lei, "Minimax rates of estimation for sparse PCA in high dimensions," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, La Palma, Canary Islands, Apr. 2012.

[70] H. Zou and L. Xue, "A selective overview of sparse principal component analysis," *Proc. IEEE*, vol. 106, no. 8, pp. 1311–1320, 2018.

[71] T. T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," *Ann. Stat.*, vol. 41, no. 6, pp. 3074–3110, 2013.

[72] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul, "Minimax bounds for sparse PCA with noisy high-dimensional data," *Ann. Stat.*, vol. 41, no. 3, pp. 1055–1084, 2013.

[73] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Proc. Conf. Learn. Theory (COLT)*, Princeton, NJ, United States, Jun. 2013.

[74] T. Wang, Q. Berthet, and R. J. Samworth, "Statistical and computational trade-offs in estimation of sparse principal components," *Ann. Stat.*, vol. 44, no. 5, pp. 1896–1930, 2016.

[75] C. Gao, Z. Ma, and H. H. Zhou, "Sparse CCA: Adaptive estimation and computational barriers," *Ann. Stat.*, vol. 45, no. 5, pp. 2074–2101, 2017.

[76] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

[77] M. Kliesch, S. J. Szarek, and P. Jung, "Simultaneous structures in convex signal recovery—revisiting the convex combination of norms," *Front. Appl. Math. Stat.*, vol. 5, 2019.

[78] J. Diestel, J. Fourie, and J. Swart, "The metric theory of tensor products (Grothendieck's *Résumé* revisited) part 1: Tensor norms," *Quaest. Math.*, vol. 25, pp. 37–72, 2002.

[79] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, pp. 805–849, 2012.

[80] B. D. Haeffele and R. Vidal, "Structured low-rank matrix factorization: Global optimality, algorithms, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1468–1482, 2020.

[81] E. Richard, G. R. Obozinski, and J.-P. Vert, "Tight convex relaxations for sparse matrix factorization," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, Montréal, Canada, Dec. 2014, pp. 3284–3292.

[82] T. Zhang, "On the dual formulation of regularized linear systems with convex risks," *Mach. Learn.*, vol. 46, pp. 91–129, 2002.

[83] A. D. McRae, J. Romberg, and M. A. Davenport, "Sample complexity and effective dimension for regression on manifolds," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual conference, Dec. 2020. arXiv: 2006.07642 [stat.ML].

[84] D. Donoho and C. Grimes, "Image manifolds which are isometric to Euclidean space," *J. Math. Imaging Vis.*, vol. 23, no. 1, pp. 5–24, 2005.

[85] G. Peyré, "Manifold models for signals and images," *Comput. Vis. Image Underst.*, vol. 113, pp. 249–260, 2009.

[86] B. Zhu, J. Liu, S. Cauley, B. Rosen, and M. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, pp. 487–492, 2018.

[87] S. Ganguli and H. Sompolinsky, "Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis," *Annu. Rev. Neurosci.*, vol. 35, pp. 485–508, 2012.

[88] A. Cohen, M. A. Davenport, and D. Leviatan, "On the stability and accuracy of least squares approximations," *Found. Comput. Math.*, vol. 13, pp. 819–834, 2013.

[89] I. Steinwart and C. Scovel, "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs," *Constr. Approx.*, vol. 35, pp. 363–417, 2012.

[90] I. Chavel, *Eigenvalues in Riemannian Geometry*. Academic Press, 1984.

[91] E. Hsu, *Stochastic Analysis on Manifolds*. Providence, RI: American Mathematical Society, 2002, ISBN: 978-0-8218-0802-3.

[92] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[93] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[94] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, 2003.

[95] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 5–30, 2006.

[96] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[97] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. Bronstein, "Geometeric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. Conf. Comput. Vis. Pattern Recog. (CVPR)*, Honolulu, Hawaii, United States, Jul. 2017, pp. 5425–5434.

[98] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst, "Geodesic convolutional neural networks on riemannian manifolds," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshop*, Santiago, Chile, Dec. 2015, pp. 832–840.

[99] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Barcelona, Spain, Dec. 2016.

[100] H. Shao, A. Kumar, and P. T. Fletcher, "The Riemannian geometry of deep generative models," in *Proc. Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, Salt Lake City, Utah, United States, Jun. 2018.

[101] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[102] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[103] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, pp. 717–772, 2009.

[104] R. Baraniuk and M. Wakin, "Random projections of smooth manifolds," *Found. Comput. Math.*, vol. 9, pp. 51–77, 2009.

[105] G. Ongie, R. Willett, R. Nowak, and L. Balzano, "Algebraic variety models for high-rank matrix completion," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Sydney, Australia, Aug. 2017, pp. 2691–2700.

[106] H. Hendriks, "Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions," *Ann. Stat.*, vol. 18, no. 2, pp. 832–849, 1990.

[107] B. Pelletier, "Kernel density estimation on riemannian manifolds," *Stat. Probab. Lett.*, vol. 73, no. 3, pp. 297–304, 2005.

[108] P. J. Bickel and B. Li, "Local polynomial regression on unknown manifolds," in *Complex Datasets and Inverse Problems*. Beachwood, OH: Institute of Mathematical Statistics, 2007, pp. 177–186.

[109] A. Aswani, P. Bickel, and C. Tomlin, "Regression on manifolds: Estimation of the exterior derivative," *Ann. Stat.*, vol. 39, no. 1, pp. 48–81, 2011.

[110] T. Hamm and I. Steinwart, "Adaptive learning rates for support vector machines working on data with low intrinsic dimension," 2020. arXiv: 2003.06202 [math.ST].

[111] M. Chen, H. Jiang, W. Liao, and T. Zhao, "Nonparametric regression on low-dimensional manifolds using deep ReLU networks," 2019. arXiv: 1908.01842 [cs.LG].

[112] R. Guhaniyogi and D. B. Dunson, "Compressed Gaussian process for manifold regression," *J. Mach. Learn. Res.*, vol. 17, 2016.

[113] R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth, "Manifold Gaussian processes for regression," in *Proc. Joint Int. Conf. Neural Netw. (ICJNN)*, Vancouver, BC, Canada, Jul. 2016.

[114] N. Dyn, F. J. Narcowich, and J. D. Ward, "Variational principles and Sobolev-type estimates for generalized interpolation on a riemannian manifold," *Constr. Approx.*, vol. 15, no. 2, pp. 175–208, 1999.

[115] H. Wendland, *Scattered Data Approximation*. Cambridge, 2004.

[116] T. Hangelbroek, F. J. Narcowich, and J. D. Ward, "Kernel approximation on manifolds I: Bounding the Lebesgue constant," *SIAM J. Math. Anal.*, vol. 42, no. 4, pp. 1732–1760, 2010.

[117] T. Hangelbroek, F. J. Narcowich, X. Sun, and J. D. Ward, "Kernel approximation on manifolds II: The L_∞ norm of the L_2 projector," *SIAM J. Math. Anal.*, vol. 43, no. 2, pp. 662–684, 2011.

[118] T. Hangelbroek, F. J. Narcowich, and J. D. Ward, "Polyharmonic and related kernels on manifolds: Interpolation and approximation," *Found. Comput. Math.*, vol. 12, no. 5, pp. 625–670, 2012.

[119] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Found. Comput. Math.*, vol. 7, no. 3, pp. 331–368, 2007.

[120] I. Steinwart, D. Hush, and C. Scovel, "Optimal rates for regularized least squares regression," in *Proc. Conf. Learn. Theory (COLT)*, Montreal, Canada, Jun. 2009.

[121] S. Mendelson and J. Neeman, "Regularization in kernel learning," *Ann. Stat.*, vol. 38, no. 1, pp. 526–565, 2010.

[122] G. Blanchard and N. Mücke, "Optimal rates for regularization of statistical inverse learning problems," *Found. Comput. Math.*, vol. 18, pp. 971–1013, 2018.

[123] ——, "Kernel regression, minimax rates, and effective dimensionality: Beyond the regular case," *Anal. Appl.*, 2020, in press.

[124] J. Lin, A. Rudi, L. Rosasco, and V. Cevher, "Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 3, pp. 868–890, 2020, in press.

[125] S. Fischer and I. Steinwart, "Sobolev norm learning rates for regularized least-squares algorithms," 2017. arXiv: 1702.07254 [stat.ML].

[126] F. Bauer, S. Pereverzev, and L. Rosasco, "On regularization algorithms in learning theory," *J. Complexity*, vol. 23, pp. 52–72, 2007.

[127] G. Blanchard and N. Krämer, "Optimal learning rates for kernel conjugate gradient regression," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 23, Vancouver, Canada, 2010.

[128] A. Dieuleveut, N. Flammarion, and F. Bach, "Harder, better, faster, stronger convergence rates for least-squares regression," *J. Mach. Learn. Res.*, vol. 18, 2017.

[129] A. Dieuleveut and F. Bach, "Nonparamatric stochastic approximation with large step-sizes," *Ann. Stat.*, vol. 44, no. 4, pp. 1363–1399, 2016.

[130] Y. Zhang, J. Duchi, and M. Wainwright, "Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates," *J. Mach. Learn. Res.*, vol. 16, pp. 3299–3340, 2015.

[131] S.-B. Lin, X. Guo, and D.-X. Zhou, "Distributed learning with regularized least squares," *J. Mach. Learn. Res.*, vol. 18, 2017.

[132] Z.-C. Guo, S.-B. Lin, and D.-X. Zhou, "Learning theory of distributed spectral algorithms," *Inverse Probl.*, vol. 33, 2017.

[133] T. Zhang, "Learning bounds for kernel regression using effective data dimensionality," *Neural Comput.*, vol. 17, pp. 2077–2098, 2005.

[134] D. Hsu, S. M. Kakade, and T. Zhang, "Random design analysis of ridge regression," *Found. Comput. Math.*, vol. 14, pp. 569–600, 2014.

[135] L. H. Dicker, D. P. Foster, and D. Hsu, "Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators," *Electron. J. Stat.*, vol. 11, pp. 1022–1047, 2017.

[136] ——, "Kernel methods and regularization techniques for nonparametric regression: Minimax optimality and adaptation," Technical report, Rutgers University, Tech. Rep., 2015.

[137] F. Bach, "On the equivalence between kernel quadrature rules and random feature expansions," *J. Mach. Learn. Res.*, vol. 18, 2017.

[138] Y. Canzani and B. Hanin, "Scaling limit for the kernel of the spectral projector and remainder estimates in the pointwise Weyl law," *Anal. PDE*, vol. 8, no. 7, pp. 1707–1731, 2015.

[139] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, Canada, Dec. 2018.

[140] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang, "On exact computation with an infinitely wide neural net," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[141] R. M. Neal, "Bayesian learning for neural networks," Ph.D. dissertation, University of Toronto, 1995.

[142] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, "Deep neural networks as Gaussian processes," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vancouver, Canada, Apr.–May 2018.

[143] A. D. McRae, S. Karnik, M. A. Davenport, and V. Muthukumar, "Harmless interpolation in regression and classification with structured features," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Virtual conference, Mar. 2022. arXiv: 2111.05198 [stat.ML], forthcoming.

[144] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Toulon, France, 2017.

[145] M. Belkin, S. Ma, and S. Mandal, "To understand deep learning we need to understand kernel learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 35, Stockholm, Sweden, Jul. 2018.

[146] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias-variance trade-off," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[147] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 67–83, 2020.

[148] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai, "Classification vs. regression in overparameterized regimes: Does the loss function matter?" *J. Mach. Learn. Res.*, 2021. arXiv: 2005.08054, forthcoming.

[149] N. S. Chatterji and P. M. Long, "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime," *J. Mach. Learn. Res.*, vol. 22, 2021.

[150] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[151] A. Rakhlin and X. Zhai, "Consistency of interpolation with laplace kernels is a high-dimensional phenomenon," in *Proc. Conf. Learn. Theory (COLT)*, Phoenix, Arizona, 2019, pp. 2595–2623.

[152] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000, ISBN: 0387987800.

[153] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, 2002.

[154] J. H. Friedman, "On bias, variance, 0/1-loss, and the curse-of-dimensionality," *Data Min. Knowl. Discov.*, vol. 1, pp. 55–77, 1997.

[155] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996, ISBN: 0387946187.

[156] V. Koltchinskii and O. Beznosova, "Exponential convergence rates in classification," in *Proc. Conf. Learn. Theory (COLT)*, Bertinoro, Italy: Springer, Jun. 2005, pp. 295–307.

[157] J.-Y. Audibert and A. B. Tsybakov, "Fast learning rates for plug-in classifiers," *Ann. Stat.*, vol. 35, no. 2, pp. 608–633, Apr. 2007.

[158] K. Wang and C. Thrampoulidis, "Benign overfitting in binary classification of gaussian mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021.

[159] Y. Cao, Q. Gu, and M. Belkin, "Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures," 2021. arXiv: 2104.13628 [cs.LG].

[160] B. Schölkopf and A. J. Smola, *Learning with Kernels, Support Vector machines, Regularization, Optimization, and Beyond*. Cambridge, Massachusetts: MIT Press, 2002, ISBN: 9780262194754.

[161] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, 2019, 568 pp., ISBN: 1108498027.

[162] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," 2020. arXiv: 2009.14286 [math.ST].

[163] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "Linearized two-layers neural networks in high dimension," *Ann. Stat.*, vol. 49, no. 2, pp. 1029–1054, 2021.

[164] K. Donhauser, M. Wu, and F. Yang, "How rotational invariance of common kernels prevents generalization in high dimensions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual conference, Jul. 2021, pp. 2804–2814.

[165] M. Belkin, "Approximation beats concentration? an approximation view on inference with smooth radial kernels," in *Proc. Conf. Learn. Theory (COLT)*, Stockholm, Sweden, Jul. 2018.

[166] L. Hui and M. Belkin, "Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Virtual conference, May 2021.

[167] D. Hsu, V. Muthukumar, and J. Xu, "On the proliferation of support vectors in high dimensions," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Virtual conference, Apr. 2021.

[168] E. Rio, "About the constants in the Fuk-Nagaev inequalities," *Electron. Commun. Probab.*, vol. 22, 2017.

[169] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[170] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, 2013.

[171] S. Mendelson, "Learning without concentration," *J. ACM*, vol. 62, no. 3, 2015.

[172] J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," in *Sampling Theory, a Renaissance, Compressive Sensing and Other Developments*, G. E. Pfander, Ed., Springer, 2015, pp. 67–101.

[173] R. Adamczak, "A tail inequality for suprema of unbounded empirical processes with applications to Markov chains," *Electron. J. Probab.*, vol. 13, pp. 1000–1034, 2008.

[174] J. M. Lee, *Introduction to Riemannian Manifolds*, 2nd ed. Springer, 2018, ISBN: 3319917544.

[175] P. Petersen, *Riemannian Geometry*, 3rd ed. Springer, 2016.

[176] X. Cheng and T.-H. Wang, "Bessel bridge representation for the heat kernel in hyperbolic space," *Proc. Amer. Math. Soc.*, vol. 146, no. 4, pp. 1781–1792, 2018.

[177] T. Liang and A. Rakhlin, "Just interpolate: Kernel "ridgeless" regression can generalize," *Ann. Stat.*, vol. 48, no. 3, pp. 1329–1347, 2020.

[178] T. Liang, A. Rakhlin, and X. Zhai, "On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels," in *Proc. Conf. Learn. Theory (COLT)*, Virtual conference, Jul. 2020, pp. 2683–2711.

[179] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM J. Math. Data Sci.*, vol. 2, no. 4, pp. 1167–1180, 2020.

[180] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," Mar. 19, 2019. arXiv: 1903.08560v4 [math.ST].

[181] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and the double descent curve," *Commun. Pure Appl. Math.*, 2021, pre-published.

[182] S. Mei, T. Misiakiewicz, and A. Montanari, "Generalization error of random features and kernel methods: Hypercontractivity and kernel matrix concentration," 2021. arXiv: 2101.10588 [math.ST].

[183] B. Adlam and J. Pennington, "Understanding double descent requires a fine-grained bias-variance decomposition," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2020.

[184] J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang, "Generalization of two-layer neural networks: An asymptotic viewpoint," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Apr. 2020.

[185] O. Dhifallah and Y. M. Lu, "A precise performance analysis of learning with random features," 2020. arXiv: 2008.11904 [cs.IT].

[186] S. D'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala, "Double trouble in double descent: Bias and variance(s) in the lazy regime," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual Conference, 2020.

[187] F. Gerace, B. Loureiro, F. Krzakala, M. Mezard, and L. Zdeborova, "Generalisation error in learning with random features and the hidden manifold model," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual Conference, 2020.

[188] H. Hu and Y. M. Lu, "Universality laws for high-dimensional learning with random features," 2020. arXiv: 2009.07669 [cs.IT].

[189] Z. Li, Z.-H. Zhou, and A. Gretton, "Towards an understanding of benign overfitting in neural networks," 2021. arXiv: 2106.03212 [stat.ML].

[190] Z. Liao, R. Couillet, and M. W. Mahoney, "A random matrix analysis of random Fourier features: Beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2020.

[191] L. Lin and E. Dobriban, "What causes the test error? going beyond bias-variance via ANOVA," *J. Mach. Learn. Res.*, vol. 22, 2021.

[192] M. Belkin, "Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation," *Acta Numer.*, vol. 30, pp. 203–248, 2021.

[193] P. L. Bartlett, A. Montanari, and A. Rakhlin, "Deep learning: A statistical viewpoint," *Acta Numer.*, vol. 30, pp. 87–201, 2021.

[194] Y. Dar, V. Muthukumar, and R. G. Baraniuk, "A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning," 2021. arXiv: 2109.02355 [stat.ML].

[195] G. Chinot and M. Lerasle, "On the robustness of the minimum ℓ_2 interpolator," 2020. arXiv: 2003.05838 [math.ST].

[196] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, 1985.