Algebraic and Geometric Structures in Machine Learning

Jack Kendrick

A dissertation
submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Dmitriy Drusvyatskiy, Chair

Rekha Thomas, Chair

Sara Billey

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

Algebraic and Geometric Structures in Machine Learning

Jack Kendrick

Co-Chairs of the Supervisory Committee:

Dmitriy Drusvyatskiy
Halicioglu Data Science Institute, University of California San Diego

Rekha Thomas
Department of Mathematics

We explore four examples of algebraic and geometric structures motivated by problems in machine learning. First, we consider the problem of learning under symmetry with kernel methods. In particular, we demonstrate that invariances have a pronounced effect on the *rank of kernel matrices*, with this rank becoming independent of the data's dimension in key examples. Relating this to the phenomenon of *representation stability*, we show that a simple regression procedure is minimax optimal for learning invariant functions that are well-defined across different dimensions.

Next, we consider the problem of density estimation on non-Euclidean spaces. Motivated by results for kernel density estimation on manifolds, we consider density estimation on $d-$*rectifiable sets*, which include smooth manifolds and algebraic varieties as special cases. In this context, we give a suitably modified kernel density estimator and prove convergence rates that are analogous to previous results.

We then consider two families of ideals, *multiview* and *universal multiview* ideals, that arise in the field of computer vision and show that a well-known family of polynomials form a universal Gröbner basis for any ideal in either family. This recovers a known result and proves a conjecture from the literature. While we only discuss multiview and universal multiview ideals, our proof strategy, which relies on a recent criterion from the commutative algebra literature, symmetry reduction, and induction, is general and may be applied to other infinite families of ideals.

Finally, we discuss tree-SNE, a data visualization technique based on the popular t-distributed stochastic neighbor embedding (t-SNE) method. A tree-SNE embedding consists of many layers, each of which is a t-SNE embedding corresponding to a different In particular, we show that tree-SNE embeddings have a continuous structure and illustrate that the embeddings provide meaningful insights through three examples.

TABLE OF CONTENTS

Page

Chapter 1: Introduction 1
1.1 Invariant Kernels 1
1.2 Density Estimation on Rectifiable Sets 4
1.3 Universal Gröbner Bases in Computer Vision 5
1.4 Visualizing Data with Tree-SNE 7

Chapter 2: Invariant Kernels 10
2.1 Introduction 10
2.2 Polynomial kernels and their rank 22
2.3 The four examples 25
2.4 Proof of Theorem 2.1.5 and Consequences 28
2.5 Representation stability and free kernels 35
2.6 Generalization of polynomial regression across dimensions 41
2.7 Computing Free Bases 44
2.8 Numerical Experiments 56
2.9 Conclusion 61

Chapter 3: Density Estimation on Rectifiable Sets 62
3.1 Introduction 62
3.2 Preliminaries 66
3.3 Proof of Theorem 3.1.1 67
3.4 Proof of Theorem 3.1.2 70
3.5 Example: Density Estimation on Sparse Data 74
3.6 Conclusion 75

Chapter 4: Universal Gröbner Bases of (Universal) Multiview Varieties 76
4.1 Introduction 76

4.2 Background . . . 77
4.3 A proof of Theorem 4.2.3 in the irreducible case . . . 80
4.4 Multiview Ideals . . . 85
4.5 The universal multiview ideal . . . 92
4.6 Conclusion . . . 97

Chapter 5: Visualizing Data with Tree-SNE . . . 98
5.1 Introduction and Results . . . 98
5.2 Preliminaries . . . 102
5.3 Proof of Continuous Structure . . . 105
5.4 Examples . . . 110
5.5 Conclusion . . . 113

Appendix A: Appenndix to Chapter 2 . . . 114
A.1 Proofs from Section 2.2 . . . 114
A.2 Proofs from Section 2.6 . . . 115

ACKNOWLEDGMENTS

There are numerous people without whom this thesis would not exist, and I am grateful to all of them. I owe particular thanks to my advisors, Rekha Thomas and Dima Drusvyatskiy, for their guidance and support throughout my time in graduate school. I would also like to thank Sara Billey and Bamdad Hosseini for serving on my committee. I am very grateful to Mateo Díaz, Tim Duff, and Stefan Steinerberger; had it not been for them, the work in this thesis would never have been completed. Special thanks go to Patricia Cahn and Julianna Tymoczko at Smith College, who were the first to introduce me to mathematical research. I feel lucky to have met many brilliant people along this journey, and I am particularly thankful for my friends Elise Catania, Garrett Mulcahy, and Julie Curtis. Finally, I thank my parents for their unwavering trust and support.

Chapter 1

INTRODUCTION

As large, high-dimensional datasets become increasingly prevalent across a wide range of applied and domain sciences, a central question to consider is how researchers can learn effectively and reliably from these datasets. An important first step is to understand the internal structures of data and identify structure that may be leveraged for computational advantage. We consider the role of algebraic, geometric, and combinatorial structures in four problems inspired by data science and machine learning. More specifically, we will see that symmetries, invariances, low intrinsic dimension, and specific algebraic structures impact how well we can learn from data in different contexts. This thesis consists of four chapters, each of which is the content of an individual paper:

1. Invariant Kernels [26]: a machine learning architecture for learning under symmetries with connections to the phenomenon of representation stability.

2. Density Estimation on Rectifiable Sets [45]: an extension of results from density estimation on manifolds to a much richer class of spaces.

3. Universal Gröbner Bases of (Universal) Multiview Ideals [28]: a universal Gröbner basis for two families of ideals from the theory of computer vision and multiview geometry.

4. Visualizing Data with Tree-SNE [46]: a hierarchical variant of t-SNE, a popular nonlinear dimensionality reduction technique used for data visualization and clustering.

1.1 Invariant Kernels

Many problems in contemporary applications involve datasets that are invariant under the action of some group. For example, properties of graphs are invariant under relabelling of vertices, point clouds as they appear in computer vision are invariant under orthogonal transformations, and unlabelled sets are invariant under permutations. Understanding the benefits of incorporating these symmetries into learning algorithms is an active area of research. This chapter, which is joint work with Díaz, Drusvyatskiy, and Thomas, explores

this question by examining *invariant polynomial kernels*, a class of symmetric bivariate functions. In particular, we show that symmetry has a dramatic effect on the *rank* of these kernels, becoming independent of the data's ambient dimension in key examples. We show that this stabilization is a consequence of a phenomenon from algebraic topology known as *representation stability*. In this setting, we construct a minimax optimal regression procedure for estimating invariant polynomials when data is sampled across dimensions.

1.1.1 Background

A *positive semi-definite (PSD) kernel on a real vector space* V is a symmetric bivariate function $K : V \times V \to \mathbb{R}$ that we may think of as a measure of similarity between elements of V due to a classical theory of Aronszajn [5]. In particular, $K(x, y)$ can be written as an inner product in some Hilbert space $\mathcal{H}$. The *rank* of K is the minimum value $r = \dim \mathcal{H}$ of such a Hilbert space. The rank of K may be infinite, however low rank kernels are desirable for both statistical and computational reasons. Namely, the classical generalization bound for ordinary least squares linear regression is known to scale linearly in $\operatorname{rank} K$, and low rank kernels enable cheap matrix-vector multiplication for faster linear algebraic computations in applications. Motivated by the prominence of datasets that are invariant under the action of some group, we define *$G-$invariant kernels*. Suppose G acts linearly on V. Then, the kernel K is $G-$invariant if for all pairs of points $x, y \in V$ and group elements $g, h \in G$, the equality $K(x, y) = K(gx, hy)$ holds. Note that this is much stronger than being *$G-$translation invariant*, where the equality need only hold when $g = h$.

1.1.2 Results

Our main result, Theorem 1.1.1, bounds the rank of an invariant kernel by the dimension of the space of invariant polynomials of a given degree.

Theorem 1.1.1: (Invariant Kernels Have Low Rank)

If K is a $G-$invariant degree m polynomial kernel on V, then

$$\operatorname{rank} K \leq \dim(\mathbb{R}[V]_m^G),$$

where $\mathbb{R}[V]_m^G$ is the space of $G-$invariant polynomials on V with degree at most m.

In concrete circumstances, combinatorial arguments show that the dimension $\dim(\mathbb{R}[V]_m^G)$ *stabilizes* for large values of $\dim V$ and thus *the rank of K is independent of the ambient di-*

mension of the data. Three important examples are permutation invariant kernels, kernels on unordered sets, and kernels on graphs. We frame this as a consequence of a phenomenon known as *representation stability*, first studied in the works [20] and [19]. In short, representation stability asserts that in certain sequences of group representations, the subspaces of group invariants become isomorphic to each other. Importantly, these sequences of representations have *free bases of invariants*: collections of sequences of invariant elements that form bases for each invariant subspace and map to each other via the aforementioned isomoprhism. We show that free bases of invariants provide a convenient method for evaluating invariant kernels on high-dimensional data.

When the dimension $\dim(\mathbb{R}[V]_m^G)$ stabilizes, a free basis of invariants must exist but finding a basis that is easy to compute in high dimensions is non-trivial. For example, a free basis of invariants for the space of degree m symmetric polynomials on $V = \mathbb{R}^d$ may be formed by taking a scaled $\mathfrak{S}_d-$average of each monomial in $\mathbb{R}[V]_m$. However, when $d \geq 10$ this is not feasible since finding the $\mathfrak{S}_d-$average requires at least $d$operations. It is generally unclear how to form free basis of invariants that scale well to high dimensions, but we provide such bases for both permutation invariant and set-permutation invariant polynomials and show that these free bases of invariants can be used to efficiently compute invariant kernels in high dimensions.

In practice, many invariant functions are well-defined regardless of the dimension of the input data. For instance, set classification tasks are well-defined regardless of the number of points within each set. Thus, instead of learning one function that accepts inputs in a specific dimension, we would like to learn a sequence of functions that accept input from all dimensions. Free bases of invariants provide a framework for studying polynomial regression where the target is a sequence of invariant functions defined across all dimensions. We show that a least-squares type estimator is minimax optimal for learning sequences of invariant polynomials when data is sampled from different dimensions. In particular, this allows us to learn from data in low dimensions to make inferences about data in much higher dimensions.

The results in this chapter are validated with numerical experiments. First, we illustrate the efficacy of invariant kernels for learning set classification tasks. We compare the performance of a standard, non-invariant kernel to the performance of an invariant kernel in several regimes and see that the invariant kernel far outperforms the non-invariant kernel in each setting. Then, we use our invariant polynomial regression estimator to learn sequences of symmetric polynomials. After training the estimator in a fixed dimension, the learned sequence of functions can be used in higher dimensions with small mean percentage error.

1.2 Density Estimation on Rectifiable Sets

Given samples $X_1, \ldots, X_n \in \mathbb{R}^D$ from some probability distribution P, a natural question is whether we can estimate the underlying distribution. In particular, given some data, we would like to know the probability of observing some new, perhaps previously unobserved, outcome. A popular tool for this is *kernel density estimation*, a non-parametric density estimation technique. It is well known that, given appropriate choices of parameters, these kernel density estimators do converge to the true underlying probability distribution. However, classical kernel density estimators suffer from the 'curse of dimensionality': convergence is slow in high dimensional spaces. Moreover, when the distribution is supported on low-dimensional subspaces, the kernel density estimator fails to define a meaningful distribution. This chapter builds on previous work on density estimation on manifolds to study density estimation on $d-$*rectifiable sets*, a rich class of spaces that includes smooth manifolds, algebraic varieties, and semi-algebraic sets as special cases. In particular, we give convergence guarantees for a modified kernel density estimator and illustrate these theoretical convergence rates with numerical experiments.

1.2.1 Background

Kernel density estimation is a standard non-parametric technique for estimating the density of an unknown probability distribution on $\mathbb{R}^D$. It is powerful since it does not assume that the underlying distribution has any specific shape, as in parametric methods, and generates a smooth distribution, which is not possible using techniques such as a histogram. A kernel density estimator has the form

$$\hat{p}_n(x) = \frac{1}{h_n^D} \sum_{i=1}^{n} K\left(\frac{\|X_i - x\|}{h_n}\right)$$

where K is some kernel function and h_n is a user-defined bandwidth parameter. Classical results of Parzen [79] and Cacoullos [15] show that a kernel density estimator converges to the true probability density function, but this convergence quickly slows down in high dimensions. A common assumption when working with high-dimensional data is the 'manifold hypothesis': although the data lives in a high-dimensional ambient space $\mathbb{R}^D$, its intrinsic dimension d is much lower. In this setting, the classical kernel density estimator does not define a meaningful probability distribution since it is necessarily supported on the whole ambient space. Nevertheless, we expect that the convergence guarantees for an appropri-

ately modified kernel density estimator should depend on its intrinsic dimension rather the ambient dimension of the space it lives in.

Previous work has considered the case where data is assumed to live on a d-dimensional Riemannian submanifold $\mathbb{R}^D$ and provided convergence guarantees for a modified kernel density estimator that mirror the classical convergence rates. However, the assumption of a Riemannian submanifold is quite strong and does not cover important examples such as sparse vectors, low rank matrices, or positive semi-definite matrices. This motivates the study of density estimation on general low-dimensional spaces.

1.2.2 Results

We analyze a modified kernel density estimator on $d-$rectifiable sets, which can be thought of as a natural measure-theoretic generalization of $d-$dimensional submanifolds. In particular, rectifiable sets have approximate tangent spaces almost everywhere and include algebraic varieties and semi-algebraic sets as special cases. Our main result establishes convergence guarantees for this modified kernel density estimator that mirror classical convergence rates.

Theorem 1.2.1: (Density Estimation on Rectifiable Sets)

Let $\Omega \subset \mathbb{R}^D$ be a $d-$rectifiable set. When bandwidth parameters h_n are chosen appropriately, the equality

$$\text{MSE}\left[\hat{p}_n(x)\right] = O\left(\frac{1}{n^{2m/(d+2m)}}\right). \tag{1.1}$$

holds with probability 1, where m is a measure of how well Ω is approximated by its approximate tangent spaces.

In particular, we see that the convergence of this modified kernel density estimator is independent of the ambient dimension D and instead depends on the intrinsic dimension d and how well the space is approximated by its approximate tangent spaces. Numerical experiments on sparse data confirm these theoretical predictions and demonstrate the effectiveness of the estimator in high-dimensional settings with low intrinsic dimension.

1.3 Universal Gröbner Bases in Computer Vision

A central problem in computer vision is recovering three-dimensional structure from multiple two-dimensional images. In this setting, multiple cameras take images of an object in the real world and produce a set of points which satisfy algebraic relationships based on the configurations of cameras. Understanding these relationships is at the heart of *multiview*

geometry. When the cameras are given, these relationships form the *multiview ideal* and when the cameras are unknown, the relationships are captured by the *universal multiview ideal*. These ideals encode all of the polynomial constraints that images must satisfy and while the polynomials themselves may change with different choices of cameras, their overarching structure is determined only by the number of cameras. In this chapter, which is joint work with Duff and Thomas [28], we show that a well-known family of polynomials, that have intrinsic geometric meaning, form a universal Gröbner basis of both the multiview and universal multiview ideals associated to one point and any number of cameras. This proof relies on a recent combinatorial criterion, symmetry reduction, and induction on the number of cameras, demonstrating how combinatorial and algebraic structure can aid in the understanding of increasingly high-dimensional problems.

1.3.1 Background

In multiview geometry, the standard model of a pinhole (projective) camera is a surjective linear map $\mathbb{P}^3 \dashrightarrow \mathbb{P}^2$ which is represented up to scale by a 3×4 matrix A of full rank. The camera maps a *world point* $q \in \mathbb{P}^3$ to its *image point* $Aq \in \mathbb{P}^2$. Given a generic configuration of n cameras, represented by the matrices $A_1, \dots, A_n$, the *multiview variety* of $(A_1, \dots, A_n)$ is the closure in $(\mathbb{P}^2)^n$ of the image of the rational map $q \mapsto (A_1 q, \dots, A_n q)$. This is an irreducible variety containing the n-tuples of images of world points $q \in \mathbb{P}^3$. The *multiview ideal* associated to cameras $(A_1, \dots, A_n)$ is the vanishing ideal of the multiview variety. A natural extension of the multiview ideal is the *universal* multiview ideal, where the cameras themselves are treated as indeterminates. Although the multiview ideal varies depending on the choice of cameras, the underlying combinatorial and algebraic structure is governed only by the number of cameras, thus we view multiview and universal multiview ideals as a nested sequence of ideals indexed by n.

1.3.2 Results

The main result of this chapter is that the set of *bi-, tri-, and quadrifocals*, a well known family of polynomials that are maximal minors of certain matrices, is a universal Gröbner basis for both the multiview and universal multiview ideals associated to one point and any number of pinhole cameras. This recovers a known result for multiview ideals from [2] and answers a previously open question from [1].

Theorem 1.3.1: (Universal Gröbner Bases of Multiview Ideals)

The set of bi-, tri-, and quadrifocals form a universal Gröbner basis for the multiview and universal multiview ideals associated to n pinhole cameras.

The main tool used to prove this is a recent result from [43] that presents a criterion for a family of squarefree polynomials being a universal Gröbner basis of a fixed ideal. This relies on the construction of an abstract simplicial complex whose nonfaces are generated by the supports of the candidate polynomials. In order to apply the criterion to our infinite sequences of multiview and universal multiview ideals, we rely on combinatorics, symmetry reduction, and induction. In particular, we directly apply the criterion in the base case where $n = 4$ and induct on n to prove our main result.

1.4 Visualizing Data with Tree-SNE

Large sets of high-dimensional data are becoming increasingly prominent across the applied sciences and so it is necessary to develop tools that can be used for their analysis and visualisation. The visualisation problem poses a unique challenge: even if a high-dimensional dataset is intrinsically close to, for example, a 5-dimensional submanifold (the manifold hypothesis), embedding five dimensions into 2- or even 3-dimensional space poses a considerable challenge. One popular dimensionality reduction technique is t-distributed stochastic neighbour embedding (t-SNE), which aims to group the data into clusters. A typical use case for t-SNE is the creation of a 2-dimensional embedding to aid with the identification of clusters in the data. This chapter explores a hierarchical variant of t-SNE known as tree-SNE that was introduced in [83] and proves that the visualizations generated by tree-SNE have continuous structure.

1.4.1 Background

First introduced in [62], t-SNE is a non-linear dimensionality reduction technique that aims to preserve local structures in high-dimensional datasets. To create a low-dimensional embedding, the t-SNE method assigns a probability distribution to high-dimensional datapoints and a separate distribution to a set of points in a low-dimensional space. The t-SNE embedding is the set of points that minimizes the KL-divergence of these two distributions. As part of a broader class of attraction-repulsion methods, low-dimensional points that represent similar high-dimensional datapoints are attracted to each other while simultaneously repelling every other point. As a result, clusters appear in the low-dimensional embedding.

Although often treated as a static problem, clustering should be viewed as a dynamic problem with a hierarchical structure. Thus, in order for a clustering to be meaningful, an appropriate scale or granularity must be chosen. However, the granularity of clusters that appear in t-SNE embeddings is fixed. Previous work demonstrated that a one-parameter family of kernels can be exploited to produce t-SNE-type embeddings with varying granularity [52], but it is not clear how to choose an appropriate kernel a priori. Motivated by this, tree-SNE was introduced in [83]. The central idea is to produce a series of t-SNE embeddings that reveal clusters with finer and finer granularity. We begin with generating an ordinary t-SNE embedding of the data. Then, additional layers are generated by changing the parameter in the kernel and optimizing a certain loss function. Each embedding forms a layer of the tree-SNE embedding and the points in each layer can be interpolated to plot smooth trajectories of each of the data points. This process is illustrated in Figure 1.1.

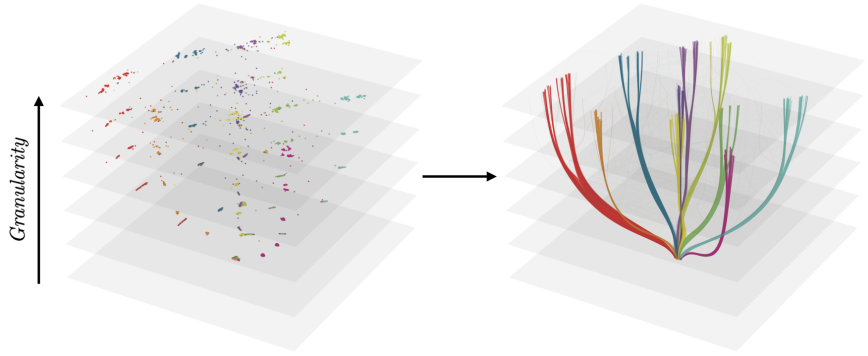


Figure 1.1: A tree-SNE embedding is generated from a series of t-SNE embeddings that reveal clusters with increasing granularity. These embeddings are then interpolated.

The resulting tree-SNE embedding has a continuous, tree-like structure. At the bottom of the tree there is a small number of large clusters, whereas at the top of the tree there is a much larger number of smaller clusters. The idea is that the branching we see within the embedding demonstrates the hierarchical structure of clusterings with different granularities. The large clusters at the bottom of the tree form the 'main branches' of the embedding that we hope correspond to macro-clusters within the data. The smaller clusters that appear higher up the tree form 'subbranches' that we hope correspond to meaningful subclusters within the data.

1.4.2 Results

The main result of this chapter concerns the continuous structure of tree-SNE embeddings. In particular, we show that continuous variation of the kernel parameter leads to continuous deformation of a t-SNE embedding. This means that when an appropriate sequence of parameters is chosen, the trajectories of points across different layers of a tree-SNE embedding are smooth. In particular, since previous work has observed that decreasing degrees of freedom increases granularity in clusters, this implies that the tree-SNE method will yield an embedding with a smooth tree-like structure.

Theorem 1.4.1: (The Tree-SNE Tree Exists)

Suppose $(y_1, \dots, y_n)$ *is a* t-SNE *embedding with parameter* α*, Under mild conditions, for each* α' *in some neighbourhood of* α *there is a* t-SNE *embedding* $(y'_1, \dots, y'_n)$ *in a neighbourhood of* $(y_1, \dots, y_n)$*.*

The 'mild conditions' referred to in this theorem relate to the smoothness of a certain function and the rank of its Hessian. These conditions are necessary, but are not restrictive. Indeed, these conditions are satisfied by generic datasets and initializations with reasonably chosen parameters.

We also demonstrate that tree-SNE embeddings capture meaningful structures in data by applying the method to the MNIST dataset and two natural language datasets from the HathuTrust library. In each example, we produce a tree-SNE embedding and analyse the large clusters that appear at the root of the tree and their trajectories as the layers of the embedding increase. We find that meaningful hierarchical structures are indeed revealed within the data, which suggests that tree-SNE is a useful tool for visually exploring both macro- and micro-level structures within high-dimensional datasets.

Chapter 2

INVARIANT KERNELS

This chapter, which is the content of the paper [26], is joint work with Díaz, Drusvyatskiy, and Thomas, and concerns the problem of learning under symmetries. Symmetry arises often when learning from high dimensional data. For example, data sets consisting of point clouds, graphs, and unordered sets appear routinely in contemporary applications, and exhibit rich underlying symmetries. Understanding the benefits of symmetry on the statistical and numerical efficiency of learning algorithms is an active area of research. In this chapter, we show that symmetry has a pronounced impact on the *rank of kernel matrices*. Specifically, we compute the rank of a polynomial kernel of fixed degree that is invariant under various groups acting independently on its two arguments. In concrete circumstances, including the three aforementioned examples, symmetry dramatically decreases the rank making it *independent of the data dimension*. We show that a simple regression procedure is minimax optimal for estimating an invariant polynomial from finitely many samples drawn across different dimensions. We conclude with numerical experiments that support our findings.

2.1 Introduction

Contemporary applications in the domain sciences and engineering often involve data sets that are invariant under some problem-specific group action. For example, graph properties are invariant under relabeling of vertices, point clouds are invariant under orthogonal transformations, unordered sets are invariant under permutations, and so forth. Estimation and inference with such data sets can benefit from taking invariance into account. In practice, one often explicitly hard-codes invariances into the architecture of the function class. High-impact examples of invariant architectures include DeepSets [108] for sets, Convolutional Neural Networks for images [53], PointNet [81, 82] and tensor field Neural Networks [101] for point clouds, and Graph Neural Networks [89] for graphs. A key question in this line of work therefore is to understand how group invariance of the data benefits both computation and sample efficiency of learning. A number of recent papers [100, 12, 68] have focused on kernel methods, showing that group invariance improves the standard sample complexity bounds by effectively scaling the dimension of the problem's data by the size of

the group. Our work is largely complementary to this effort. We show that incorporating invariance can significantly decrease the rank of polynomial kernels, which is beneficial both for computation and inference in practice.

Setting the stage, a *kernel* $K\colon V \times V \to \mathbb{R}$ *on a set* V is a symmetric bivariate function. Throughout, we will assume that V is a finite-dimensional vector space; e.g. $V = \mathbb{R}^d$ or $V = \mathbb{R}^{d\times d}$. The primary use of kernels in applications is to measure pairwise similarity between points $x_1, \dots, x_n \in V$ via the matrix $\mathbf{K} = [K(x_i, x_j)]_{i,j=1,\dots,n}$. A kernel K admits a *rank-r factorization* if it can be written as

$$K(x, x') = \langle \varphi(x), \psi(x') \rangle \qquad \forall x, x' \in V, \tag{2.1}$$

for some maps $\varphi, \psi\colon V \to \mathbb{R}^r$. The *rank* of K, denoted $\operatorname{rank} K$, is the minimal such r.

Low-rank kernels are desirable both for computational and statistical reasons. A low rank-factorization $\mathbf{K} = AB^\top$ enables cheap matrix vector products and therefore faster linear algebraic computations. Moreover, state-of-the-art kernel system solvers are based on forming low-rank approximations of kernel matrices to use as preconditioners; e.g. RPCholesky [17, 27], random Fourier Features [7, 25], EigenPro [60, 61], and Falcon [67, 66]. From a statistical perspective, the classical generalization bound for the ordinary least squares estimator for linear regression $y = \langle \varphi(x), \beta^* \rangle + \varepsilon$ scales as $O(\sigma^2 \operatorname{rank}(K)/n)$ (e.g. [104, Exercise 13.2]); therefore $\operatorname{rank} K$ serves as the intrinsic dimension of the problem.

In this chapter, we focus on polynomial kernels of a fixed degree m. By this we mean that $K(x, y)$ is a polynomial of degree m in each of the arguments x and y. We emphasize that we will regard m as being a fixed constant throughout. The reader should think of m as being relatively small while the dimension of the data d and the number of data points n are large. A baseline example to keep in mind is the standard polynomial kernel $K(x, y) = (1 + x^\top y)^m$ on $\mathbb{R}^d$ which can be realized by the feature maps $\varphi = \psi$ that evaluate all monomials x^α with $|\alpha| \le m$. In particular, K has rank $\sum_{j=0}^{m} \binom{d+j-1}{j} \asymp d^m$ (ignoring constants in m).

We investigate the impact of symmetry on the kernel rank. Namely, given a group G acting on V, we say that K is G-*invariant* if the following equality holds:[1]

$$K(gx, g'x') = K(x, x') \qquad \forall x, x' \in V,\ g, g' \in G. \tag{2.2}$$

[1] $G-$invariance should be contrasted with the much weaker property of $G-$translation invariance, which stipulates that (2.2) holds only with $g = g'$. In contrast to $G-$invariant kernels, the rank of $G-$translation invariant kernels is usually high.

In particular, if G is a finite group then any kernel induces a G-invariant kernel by averaging; see Section 2.2.1 for details. We investigate the rank of G-invariant kernels of degree m on a d-dimensional vector space V, which should always be compared to the rank d^m of the standard polynomial kernel. Our contributions are as follows:

We will see that in a number of important settings the rank of G-invariant kernels is independent of the ambient dimension d and derive consequences for sample efficiency of learning G-invariant polynomials when data is sampled across different dimensions.

2.1.1 Summary of results

Much of our computations stem from the following elementary observation. The rank of any G-invariant degree m kernel K on a vector space V can be upper-bounded as

$$\operatorname{rank} K \leq \dim(\mathbb{R}[V]_m^G) \tag{2.3}$$

where $\mathbb{R}[V]_m$ denotes the vector space of polynomials on V of degree at most m and $\mathbb{R}[V]_m^G$ the G-invariant subspace. So, bounding the rank of K is reduced to bounding the dimension of $\mathbb{R}[V]_m^G$—a well-studied object in combinatorics [97] and representation theory [92].

We next describe some concrete examples where group invariance yields a rank reduction. The explicit bounds rely on so-called integer partitions. A *partition of an integer j*, denoted $\alpha \vdash j$, is a sequence $\alpha_1 \geq \alpha_2 \geq \ldots$ of strictly positive integers (called *parts*) summing to j. The symbol $p(j)$ will denote the number of integer partitions of j.

Permutation Invariant Functions For our first example, we focus on functions on $V = \mathbb{R}^d$ that are invariant under the symmetric group $\mathfrak{S}_d$ acting via coordinate permutations. Common examples are coordinate sums, products, and functions of symmetric polynomials. It follows from well-known results—see, for example, [34]—that the dimension $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d}$ of the space of $\mathfrak{S}_d$-invariant polynomials is given by a quantity that only depends on the degree m of the polynomials and not on the dimension d of V.

Example 2.1.1: (Permutations)

In the setting $d \geq m$, the quantity $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d})$ is independent of $\dim V$ and satisfies:

$$\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d}) = \sum_{j=0}^{m} p(j). \tag{2.4}$$

We stress that the bound (2.4) depends only on m and is at most $\exp(\pi\sqrt{2m/3})$ by the Hardy-Ramanujan asymptotic formula [40]. Concretely, for $m = 1, \dots, 10$, we may compute $\sum_{j=0}^{m} p(j)$ yielding the bounds $2, 4, 7, 12, 19, 30, 45, 67, 97$, and 139, respectively.

Set-permutation Invariant Functions. The next example considers invariances that are typically exhibited when each data point is itself a set of vectors. Such problems appear routinely in practice; see for example the influential work [108]. Specifically, let $S = \{x_1, \dots, x_d\}$ be a set of d points $x_i \in \mathbb{R}^k$. In applications, x_i may denote the rows of a table S; as such, the number of rows d is large, while each row x_i lies in a low-dimensional space $\mathbb{R}^k$. Clearly, S can be represented by a matrix $X \in \mathbb{R}^{d\times k}$ having x_i as its rows. For any permutation matrix $P \in \mathbb{R}^{d\times d}$, the matrices PX and X have the same rows in a different order and therefore, the two matrices represent the same set. Thus we will be interested in kernels on $V = \mathbb{R}^{d\times k}$ that are invariant under the action of $\mathfrak{S}_d$ by permuting rows. Existing results, such as those in [32, 96, 105], imply that $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d})$ is bounded by a quantity independent of the dimension. We provide a simple combinatorial argument to explicitly calculate the value of this dimension, summarized in the following example.

Example 2.1.2: (Set-permutations)

When $d \geq m$, the quantity $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d})$ is independent of $\dim V$ and satisfies:

$$\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d}) = \sum_{j=0}^{m} \sum_{\alpha\vdash j} \prod_{i=1}^{j} \binom{\binom{i+k-1}{i} + \mu_i(\alpha) - 1}{\mu_i(\alpha)},$$

where $\mu(\alpha)$ is the vector of the multiplicities of parts of α.

Although the bound might look cumbersome, it can be easily computed for small values of m and k, and most importantly, this bound is independent of d. Using the Hardy-Ramanujan asymptotic formula , the bound in Theorem 2.1.2 grows at most as $\exp(\pi\sqrt{2m/3})\cdot(k-1)^{mk}$.

Functions on Graphs. Many problems arising in applications involve data in the form of weighted graphs. Recall that all weighted graphs on d vertices can be uniquely represented by its $d \times d$ adjacency matrix which is invariant up to simultaneous permutation of rows and columns since the labeling of a graph's vertices is arbitrary. It follows that any function of graphs that takes adjacency matrices as input should be invariant under the action of $\mathfrak{S}_d$ by conjugation. Let V consist of all symmetric $d \times d$ matrices. Clearly, all adjacency matrices of simple graphs on d vertices are contained in V. Existing results, such as those contained

in the OEIS sequence [31], imply that the dimension $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d})$ is bounded by a quantity independent of d. We prove an alternative upper bound for $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d})$ and show that the dimension $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d})$ stabilizes for all $d \geq 2m$.

Example 2.1.3: (Graphs)

In the setting $d \geq 2m$, the quantity $\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d})$ is independent of $\dim V$ and satisfies:

$$\dim(\mathbb{R}[V]_m^{\mathfrak{S}_d}) \leq \sum_{j=0}^{m} \sum_{\alpha \vdash 2j} q_\alpha,$$

where q_α is the number of non-negative integer symmetric matrices with row sum α.

We are not aware of any procedure to compute q_α exactly. It can be upper bounded, however, by relaxing the symmetry assumption and stipulating that the row sums and columns sums are α and using the recursive procedure in [72]. A straightforward upper bound obtained by relaxing the symmetry assumption is $2m \cdot \exp(\pi\sqrt{m/12}) \cdot (m-1)^{2mk}(2m)^{m(m-1)}$.

Functions on Point Clouds Problems in fields such as computer vision and robotics often involve data in the form of *point clouds*, i.e. sets of vectors $S = \{x_1, \ldots, x_d\} \subset \mathbb{R}^k$. Clearly, we may store a point cloud as a matrix $X \in \mathbb{R}^{d \times k}$ having x_i as its rows. Point clouds are invariant under the relabeling of points in the cloud, as well as simultaneous orthogonal transformations. Let G be the group of these symmetries acting on matrices $X \in \mathbb{R}^{d \times k}$. The following result gives a dimension-independent upper bound on $\dim(\mathbb{R}[V]_m^G)$.

Example 2.1.4: (Point clouds)

In the regime $d \geq 2m$, the quantity $\dim(\mathbb{R}[V]_m^G)$ is independent of $\dim V$ and satisfies:

$$\dim(\mathbb{R}[V]_m^G) \leq \sum_{j=0}^{m/2} \sum_{\alpha \vdash 2j} q_\alpha,$$

where q_α is the number of non-negative integer symmetric matrices with row sum α.

A Monte Carlo rank estimator Computing the dimension of invariants, as in the aforementioned examples, typically reduces to tedious combinatorial arguments. We now describe an alternative approach, which allows one to estimate the dimension of the invariants simply by sampling elements of the group. The statement of the theorem requires some further no-

tation, which we outline now; see Section 2.3.1 for details. When G is a finite group acting on V, the symbol $\mathbb{E}_g[\cdot]$ will mean expectation with respect to the uniform measure on G. For any integer k, the symbol $C_d(k)$ denotes the set of all weak compositions of k into d parts and $\eta(\alpha)$ denotes the product of the factorials of the multiplicities of non-negative entries in α. The symbols $\lambda_i(g)$ denote the eigenvalues of $g \in G$ represented as a matrix, while $g[\alpha]$ and $\operatorname{per} g[\alpha]$ denote the submatrix of g indexed by α and its permanent.

Theorem 2.1.5: (Calculating the rank)

Let G be a finite group acting linearly on a d-dimensional vector space V. Then,

$$\dim(\mathbb{R}[V]_m^G) = \mathbb{E}_g\left[\sum_{j=0}^{m}\sum_{\alpha\in C_d(j)}\prod_{i=1}^{d}\lambda_i(g)^{\alpha_i}\right] = \mathbb{E}_g\left[\sum_{j=0}^{m}\sum_{\alpha\in C_d(j)}\frac{\operatorname{per} g[\alpha]}{\eta(\alpha)}\right]. \tag{2.5}$$

Although the first equation in (2.5) looks complicated, it is not very difficult to compute the terms inside the expectation since the degree m is assumed to be small. Consequently, even when the group G is very large one may estimate $\dim(\mathbb{R}[V]_m^G)$ by drawing iid samples from G and replacing $\mathbb{E}_g$ by an imperical average. The downside of the expression is that one needs to compute eigenvalues $\lambda_i(g)$ of a $d \times d$ matrix. The second equality in (2.5) provides an alternative, wherein one instead needs to only compute the permanent of small submatrices of g whose columns and rows are indexed by the weak composition α. Since α is a composition of at most m, each submatrix is at most $m \times m$. Again, even when the group G is very large one may estimate $\dim(\mathbb{R}[V]_m^G)$ by using an empirical average.

Connection to representation stability The stabilization of the dimension of fixed-degree invariant polynomials is not merely coincidental, but rather a consequence of a broader phenomenon known as *representation stability* [20, 19]. To elucidate this connection, consider a sequence of nested groups $\{G_d\}_{d\in\mathbb{N}}$ (e.g., $G_d = \mathfrak{S}_d$) acting on a sequence of vector spaces $\{W_d\}_{d\in\mathbb{N}}$ of increasing dimension that embed isometrically into each other $W_d \hookrightarrow W_{d+1}$ (e.g., $W_d = \mathbb{R}[V_d]_m$). Representation stability asserts that if these sequences satisfy certain algebraic relationships, then the subspaces of invariants $W_d^{G_d} = \{w \in W_d \mid g \cdot w = w \ \forall g \in G_d\}$ become isomorphic to each other. The following result, which can also be deduced from the more general categorical framework in [56], is presented here via direct and elementary arguments.

Theorem 2.1.6: (Polynomial basis across dimensions)

Let $\{G_d\}_{d\in\mathbb{N}}$ and $\{V_d\}_{d\in\mathbb{N}}$ be any of the sequences of groups and vector spaces considered in Examples 2.1.1-2.1.4 and define $W_d := \mathbb{R}[V_d]_m$. Then, for each example, there exist $r \in \mathbb{N}$ and invariant polynomials $\left\{f_1^{(d)}, \dots, f_r^{(d)} \in W_d^{G_d}\right\}_{d\in\mathbb{N}}$ such that

1. *(**Basis**) For all large d, the polynomials $f_1^{(d)}, \dots, f_r^{(d)}$ form a basis of $W_d^{G_d}$.*
2. *(**Consistency**) The basis elements project onto each other:*

$$\mathrm{proj}_{W_d}\left(f_i^{(d+1)}\right) = f_i^{(d)} \quad \text{for all } i \in [r] \text{ and } d \in \mathbb{N}.$$

A sequence of polynomials satisfying these two properties is called a *free basis*. The existence of a free basis enables us to encode a sequence of "consistent" invariant polynomials

$$\left\{p_d(x) = \sum_{i=1}^{r} \alpha_i f_i^{(d)}(x)\right\}_d$$

using only finitely many parameters $\alpha_1, \dots, \alpha_r$. Here, consistency means that polynomials with different numbers of variables yield matching outputs when evaluated on low-dimensional inputs. We refer to such dimension-independent polynomial sequences as *free polynomials*. Common examples of free polynomials include the ℓ_2-norm squared of a vector and the trace of a symmetric matrix, with many more appearing in optimization, signal processing, and particle systems [56]. The finite parameterizability of free polynomials makes them a powerful tool for computation and statistical inference across dimensions.

Computing invariant bases and kernels Although the bounds in Examples 2.1.1-2.1.4 ensure that the rank of an invariant kernel is independent of its data dimension, it is not always the case that invariant kernels can be efficiently computed in high dimensions. Typical kernels used in applications, such as the Gaussian and standard polynomial kernels, are compactly expressible using inner products or vector norms and are therefore efficient to compute. Our next contribution is to show in two important cases how invariant kernels and free polynomial bases can be computed efficiently.

Free basis of invariant polynomials allows us to develop a procedure for evaluating invariant kernels. Let $\{V_d\}_{d\in\mathbb{N}}$ a sequence of vector spaces acted on by a sequence of groups $\{G_d\}_{d\in\mathbb{N}}$. Suppose that $K_d : V_d \times V_d \to \mathbb{R}$ is a G_d—invariant kernel of degree m and $\{f_1^{(d)}, \dots, f_r^{(d)}\}_{d\in\mathbb{N}}$ form a free basis of invariants for $\{\mathbb{R}[V_d]_m^{G_d}\}_{d\in\mathbb{N}}$. We begin by showing

that the invariant kernel K_d can be written as a symmetric bilinear form:

$$K(x, y) = F(x)^\top C F(y)$$

for all $x, y \in V_d$, where $F(x) = \begin{bmatrix} f_1^{(d)}(x) & \dots & f_r^{(d)}(x) \end{bmatrix}^\top$. The matrix C is unique and is independent of the dimension d for all $d \geq m$.

For each sequence of vector spaces in Examples 2.1.1-2.1.4, a free basis of invariants can be formed by calculating the $\mathfrak{S}_d$−average of each monomial in $\mathbb{R}[V_d]_m$. However, the number of operations required to evaluate each average is at least $|\mathfrak{S}_d| = d!$, which is not feasible for large values of d. Generally, it is unclear how to form free bases that can be computed with a numerical effort that is polynomial in dimension d. With this in mind, we show that there are efficiently computable free bases for permutation and set-permutation invariant polynomials. We show that *elementary symmetric polynomials indexed by partitions* and *polarized elementary symmetric polynomials indexed by vector partitions* form free bases in these two cases, respectively. With these bases, we obtain bounds on the complexity of evaluating these invariant kernels that scale linearly in the dimension d.

Theorem 2.1.7: (Computing $\mathfrak{S}_d$-invariant kernels for vectors and sets)

1. *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a permutation invariant polynomial kernel of degree m. Assume that one has access to a matrix C such that $K(x, y) = E_m(x)^\top C E_m(y)$ for all choices of $x, y \in \mathbb{R}^d$, where $E_m(x)$ is the vector of elementary symmetric polynomials indexed by partitions of degree at most m. Then, $K(x, y)$ can be evaluated using*

$$O\left(md + m \exp\left(\sqrt{2m/3}\right)\right)$$

floating point operations.

2. *Let $K : \mathbb{R}^{d\times k} \times \mathbb{R}^{d\times k} \to \mathbb{R}$ be a set-permutation invariant polynomial kernel of degree m. Assume that one has access to a matrix C such that $K(x, y) = P_m(x)^\top C P_m(y)$ for all choices of $x, y \in \mathbb{R}^{d\times k}$, where $P_m(x)$ is the vector of polarized elementary symmetric polynomials indexed by partitions of degree at most m. Then, $K(x, y)$ can be evaluated using*

$$O\left(m \exp\left(2\sqrt{2m/3}\right) \cdot k^{2mk} + \frac{dk^m}{m!}\right)$$

floating point operations.

The proof of Theorem 2.1.7 amounts to determining the complexity of evaluating the relevant basis of invariant polynomials (see Propositions 2.7.4 and 2.7.10), and so similar statements could be made for other groups given a set of efficiently computable free bases. However, although efficient computation of invariant kernels is intertwined with the efficient computation of invariant bases, the two problems are distinct. So, while invariant polynomial bases provide one method of computing invariant kernels in high dimensions, invariant kernels may be computed without explicitly evaluating the relevant basis. For example, in Illustration 9, we give an example of set-permutation invariant kernel that can be evaluated efficiently without evaluating the basis of polarized elementary symmetric polynomials.

Generalization of invariant polynomial regression across dimensions As our final contribution, we leverage the existence of the basis prescribed by Theorem 2.1.6 to study polynomial regression problems where we aim to learn a sequence of fixed-degree invariant polynomials. Formally, we wish to recover a free polynomial[2] — parameterized by $\alpha^\star \in \mathbb{R}^k$ via $p^\star(x) = \sum_{i=1}^k \alpha_i^\star f_i(x)$ with $f_1, \dots, f_k$ a basis of $\mathbb{R}[V_d]_m^{G_d}$ — from a set of noisy input-output pairs (x_i, y_i) where $y_i = p^\star(x_i) + \varepsilon_i$ with the ε's drawn i.i.d. with mean-zero and bounded variance. Importantly, the feature vectors x might have different dimensions. We propose a least-squares-type estimator of the latent parameter via

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathbb{R}^k} \frac{1}{2n} \sum_{i=1}^n \left(\sum_{j=1}^k \alpha_i f_j(x_i) - y_i \right)^2, \tag{2.6}$$

note that the basis elements f_k are different depending on the dimension of their input.

Theorem 2.1.8: (Minimax optimality across dimensions - Informal)

Fix a sequence of polynomials $\{p_d^\star \in \mathbb{R}[V_d]^{G_d}\}$ parameterized by $\alpha^\star \in \mathbb{R}^k$, a training set of input-output pairs $S = \{(x_1, y_1), \dots, (x_1, y_n)\}$ and a test set $T = \{(x_1', y_1'), \dots (x_m', y_m')\}$ with the feature vectors in potentially different dimensions. Then, the estimator $\hat{\alpha}$ in (2.6), trained with S, yields a the sequence of invariant polynomials

$$\left\{ \hat{p}_d = \sum_{i=1}^k \hat{\alpha}_k f_k^{(d)} \in \mathbb{R}[V_d]^{G_d} \right\}$$

that is minimax optimal with respect to the excess risk defined by any test set T.

[2]We drop the index d as it can be inferred from the dimension of the input x.

Outline. The outline of the rest of the chapter is as follows. The remainder of this section overviews the related literature. Section 2.2 provides the relevant background on invariant polynomial kernels and proves our main Theorem 2.2.7. Section 2.3 verifies each of the Examples 2.1.1, 2.1.2, 2.1.3, and 2.1.4. Section 2.4 overviews basic representation and character theory and proves Theorem 2.1.5. The next three sections concern connections to representation stability and generalization across dimensions: Section 2.5 gives a background on representation stability, Section 2.6 discusses polynomial regression across dimensions and provides minimax optimal bounds for excess risk, and Section 2.7 provides efficient methods for computing permutation and set-permutation invariant kernels. Finally, Section 2.8 details three numerical experiments that support the theory in this chapter. Proofs from Section 2.2 and Section 2.6 are given in Appendix A.1 and Appendix A.2 respectively.

2.1.2 Related literature

A number of works have examined the gain achieved by learning with invariant kernels.

Approximation and generalization gain in invariant methods. Mei, Misiakiewicz, and Montanari [68] characterized the gain in both approximation and generalization error of certain invariant random feature and kernel methods. In particular, the assumed target functions are invariant under the action of some subgroup of the orthogonal group and that data is uniformly sampled from either a $d-$dimensional sphere of radius $\sqrt{d}$ or a $d-$dimensional hypercube. In both the under- and overparametrized regimes, they showed that the gain from learning under invariances with kernel ridge regression is d^{α} for some parameter α that is dependent on the group of invariances. Several groups, such as both one and two-dimensional cyclic groups, satsify $\alpha = 1$ yielding a linear gain in the dimension.

Sample complexity and generalization bounds under invariances. Both [100] and [12] discussed the gain in sample complexity and prove generalization bounds for kernel ridge regression when learning under invariances. The paper [12] studied invariances under isometric actions of finite permutation groups on the sphere and [100] presented similar results for arbitrary Lie group actions on arbitrary compact manifolds, a much more general class of problems. In both works the statistical properties of the KRR estimator are controlled by the decay of the eigenvalues of certain operators. In [100], the eigenvalues of the Laplace-Beltrami operator are controlled and in [12] the eigenvalues of the covariance operator are controlled. A key tool used in both of these works is counting the dimensions of various classes of invariant functions: in [12], the proofs rely on comparing the dimensions of invariant and non-invariant spherical harmonics at varying degrees, and in [100] the arguments

utilize a dimension counting theorem for functions on the quotient space $\mathcal{M}/G$ where $\mathcal{M}$ is a compact manifold and G a Lie group. Though we are interested in a different question than in [12, 100], we use a similar strategy of determining the dimensions of invariant function spaces. Our work focuses on spaces of polynomials and in many of our concrete examples, elementary combinatorial arguments can be used to explicitly count and bound the dimensions of invariant polynomial spaces without being restricted to finite group actions on the sphere as in [12] or having to calculate volumes and dimensions of quotient spaces as in [100]. In our examples, there is no dependence on the dimension of the data; instead, our results depend on the degree of the polynomials being considered.

Generalization gain of invariant predictors. The work [30] also discussed the benefit of invariances, though a different approach to enforcing invariance is taken. While our work, as well as the previously discussed works, consider kernels that are themselves assumed to be invariant under the action of some group, [30] considered predictors that arise via normal KRR and are then averaged over a group to form an invariant function. For a given RKHS $\mathcal{H}$, the action of a group G induces the decomposition $\mathcal{H} = \bar{\mathcal{H}} \oplus \mathcal{H}_\perp$ where $\bar{\mathcal{H}}$ is the subspace of $\mathcal{H}$ consisting of all $G-$invariant functions. The so-called *generalization gap* of the squared error loss between a predictor and its projection onto the invariant subspace $\bar{\mathcal{H}}$ scales with the effective dimension of $\mathcal{H}_\perp$, which is calculated in terms of the eigenvalues of the integral operator mapping the space of square-integrable functions to $\mathcal{H}$. From this result, it is clear that having strong upper bounds on the dimension of invariant function spaces is important for understanding the benefit of learning with invariances in the worst case. Unlike [68, 12, 100], the generalization benefits in [30] do not depend on properties of the underlying data distribution and are instead focused on the underlying structure of kernels, which is also the case for our work.

Kernels with finite rank. The previous works examine the role of invariance in decreasing the generalization error of kernel methods. This is different from our guiding question of how exploiting invariances can lead to a decrease in the rank of a kernel. The recent work [18] considers the generalization error of kernel ridge regression for kernels of finite rank. In particular, the authors proved sharp upper and lower bounds on the test error showed that variance is bounded linearly in the rank of the kernel. This suggests that low-rank kernels, such as those that arise from exploiting invariance, could be less prone to over-fitting data.

Permutation invariant kernels. The work [36] defined *argumentwise invariant kernels*, which coincides with our notion of an invariant kernel, and discussed the use of these kernels for modeling invariant functions via Kriging, also known as Gaussian process regression.

The authors showed that the use of argumentwise invariant kernels led to invariant paths in conditional simulations as well as the invariance of the Kriging mean and variances functions. This work was built upon in [35], where the authors dsicussed *argumentwise degeneracy* of kernels and studied pathwise invariances of random fields. The focus of our work is different: while our notion of invariant and argumentwise invariant kernels overlap, we are concerned with how invariance affects the rank of kernels and leads to stabilization. Furthermore, the authors of [36] compute invariant kernels by averaging kernels over a group and so their methods do not scale well to high dimensional data, whereas we provide methods for computing certain $\mathfrak{S}_d-$invariant kernels in high dimensions.

The paper [49] defined *antisymmetric* and *symmetric* kernels on $\mathbb{R}^d$, with their notion of symmetric kernels coinciding with our definition of $\mathfrak{S}_d-$invariant kernels. These antisymmetric and symmetric kernels were applied to problems in quantum chemistry and physics to exploit known symmetries in data. The authors showed that the degree of the feature space generated by an $\mathfrak{S}_d-$invariant polynomial kernel of degree m is bounded by the sum $\sum_{j=0}^{m} p(j)$, which is comparable to our Example 2.1.1. However, in our verification of Example 2.1.1, we show that this is a special case of our main theorem, Theorem 2.2.7. While there is some overlap between our results and the work in [49], our arguments generalize to a much wider class of groups. Moreover, the authors of [49] relied on either averaging kernels over the whole group $\mathfrak{S}_d$ and/or evaluating large number of hyperpermanents to compute the value of $\mathfrak{S}_d-$invariant kernels. This becomes infeasible for high dimensional problems since the size of the permutation group is exponential in the dimension. In contrast, we utilize the theory of symmetric polynomials to provide efficient methods of computing $\mathfrak{S}_d$-invariant kernels for large values of d.

Representation stability and generalization across dimensions. Originally introduced in algebraic topology [20, 19], representation stability has since been applied to study limits of several mathematical objects [87, 88, 86, 22, 3]. Closer to our setting, [56] used representation stability to parameterize sequences of convex sets of increasing dimension, and [57] employed it to design equivariant neural networks that accommodate inputs of arbitrary dimension. To our knowledge, no generalization bounds across dimensions have been established. However, in the context of Graph Neural Networks (GNNs) [37, 70, 14, 107], there has been noteworthy progress on the related notion of transferability. GNNs can handle graphs of arbitrary size and as these graphs increase in size, converging to a *graphon*, the question of whether GNN outputs also converge arises. Recent results [85, 64] show that they do, ensuring that the outputs on "related" graphs of different sizes remain close to one another.

2.2 Polynomial kernels and their rank

We begin by recording some background on kernels, following the standard texts on the subject [91, 93]. To this end, let V be a finite-dimensional vector space with a fixed basis. A *kernel* $K\colon V \times V \to \mathbb{R}$ is any symmetric function of its two arguments. The main use of kernels in applications is to measure pairwise similarity between data points $x_1, \dots, x_n \in V$ via the $n \times n$ symmetric matrix $\mathbf{K} = [K(x_i, x_j)]_{i,j=1}^n$. The kernel K is called *positive semidefinite (PSD)* if the matrix $\mathbf{K}$ is positive semidefinite for any data points. Typical examples of PSD kernels on $\mathbb{R}^d$ are the polynomial kernel $K(x, y) = (1 + x^\top y)^m$, the Taylor features kernel $K(x, y) = \exp(-(\|x\|^2 + \|y\|^2)/2\sigma^2) \cdot \sum_{j=1}^m \frac{\langle x, y\rangle^j}{\sigma^{2j} j!}$, and the Gaussian kernel $K(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$, where σ is some fixed bandwidth parameter.

A kernel K is said to admit a *rank r-factorization* if there exist maps $\phi,\ \psi\colon V \to \mathbb{R}^r$—called *feature maps*—satisfying

$$K(x, y) = \langle \phi(x), \psi(y)\rangle \qquad \forall x, y \in V.$$

The *rank of* K, denoted $\operatorname{rank} K$, is the minimal r such that K admits a rank r-factorization. Importantly, the rank of K upper-bounds the rank of the corresponding kernel matrix:

$$\operatorname{rank} \mathbf{K} \le \operatorname{rank} K.$$

2.2.1 Group invariance and symmetrization.

We will be interested in the interplay between kernel rank and group invariance. Namely, let G be a group acting on V. A kernel K on V is called *G-invariant* if the following equality holds:

$$K(gx, hy) = K(x, y) \qquad \forall x, y \in V,\ g, h \in G. \tag{2.7}$$

Notice that in (2.7), the group G acts independently on the two coordinates x and y. This should be contrasted with the much weaker property of *G-translation invariance*, which stipulates that (2.7) holds only with $h = g$. Importantly, G-invariant kernels are well-defined on the orbit space V/G, i.e., the set of orbits of V under the action of G.

Any kernel can be symmetrized with respect to any finite group G simply by averaging:

$$K^*(x, y) := \tfrac{1}{|G|^2} \sum_{g,h \in G} K(gx, hy).$$

The analogous construction for infinite groups requires averaging with respect to a translation invariant measure on G, called the Haar measure. The precise formalism is as follows.

Assumption 2.2.1 (Haar measure)**.** Suppose that G is a compact Hausdorff topological group, equipped with the Borel σ-algebra and the right Haar measure μ normalized to satisfy $\mu(G) = 1$. Suppose moreover that G acts continuously on V.

To shorten the notation, we will use the symbol $\mathbb{E}_g$ to mean expectation with respect to the Haar measure μ on G. A key property of the Haar measure is the invariance $\mathbb{E}_g \phi(gh) = \mathbb{E}_g \phi(h)$ for any measurable function ϕ and group element $h \in G$. In particular, any continuous function f yields the G-invariant symmetrization

$$f^*(x) := \mathop{\mathbb{E}}_{g} f(gx).$$

Similarly, any continuous kernel K on V yields the new G-invariant kernel by averaging

$$K^*(x, y) := \mathop{\mathbb{E}}_{g,h} K(gx, hy).$$

The following lemma is standard; we provide a short proof in Section A.1.1.

Lemma 2.2.2. *The kernel K^* is G-invariant. Moreover, if K is PSD then so is K^*.*

The main thrust of our work is to show that the rank of the symmetrized kernel K^* is often much smaller than the rank of K.

2.2.2 Invariant polynomials and kernel rank

In the rest of this chapter, we focus on kernels that are polynomials of a fixed degree. To be precise, let $\mathbb{R}[V]_m$ be the vector space of all polynomials over V with total degree at most m. Any linear action of a group G on V promotes to a linear action of G on the polynomial ring $\mathbb{R}[V]_m$ by

$$(g \cdot f)(x) = f(g^{-1}x).$$

The symbol $\mathbb{R}[V]_m^G$ will denote the set of G-invariant polynomials $f \in \mathbb{R}[V]_m$, meaning

$$f(gx) = f(x) \qquad \forall x \in V,\ g \in G.$$

The dimension of $\mathbb{R}[V]_m^G$, a linear subspace of $\mathbb{R}[V]_m$, will play a key role in this chapter.

Definition 2.2.3 (Polynomial kernel)**.** A kernel $K\colon V \times V \to \mathbb{R}$ is a *polynomial kernel* if $K(x, y)$ is a polynomial for all $x, y \in V$. The *degree* of $K(x, y)$ is its maximal degree in x (or equivalently in y, by symmetry).

Assumption 2.2.4. We fix the following setting.

A.1 The function K is a degree m polynomial kernel on V.

A.2 The group G acts on V satisfying Assumption 2.2.1.

A.3 The set of polynomials $\{f_1, \dots, f_N\}$ forms a basis of the vector space $\mathbb{R}[V]_m$.

It will be convenient to express K as a bilinear form with respect to the basis $\{f_i\}_{i=1}^N$—the content of the following standard lemma. A short proof can be found in Section A.1.2.

Lemma 2.2.5 (Bilinear form)**.** *Suppose that Assumption 2.2.4 holds. Then there exists a unique matrix $C \in \mathbb{R}^{N\times N}$, which is necessarily symmetric, such that for all $x, y \in V$, it holds:*

$$K(x, y) = F(x)^\top \cdot C \cdot F(y), \tag{2.8}$$

$$K^*(x, y) = F^*(x)^\top \cdot C \cdot F^*(y), \tag{2.9}$$

where we set $F(x) = [f_1(x), \dots, f_N(x)]^\top$ and $F^*(x) = [f_1^*(x), \dots, f_N^*(x)]$.

Conveniently, the rank of the matrix C in equation (2.8) coincides with the rank of K; the following lemma is proved in [4, Proposition 3.6] and [4, Corollary 3.8] in greater generality.

Lemma 2.2.6 (Exact rank formula)**.** *Suppose Assumption 2.2.4 holds. Fix a matrix C satisfying (2.8). Then $\operatorname{rank} K = \operatorname{rank} C$. Moreover, there exist polynomial maps $\varphi, \psi\colon V \to \mathbb{R}^{\operatorname{rank} K}$ satisfying $K(x, y) = \langle \varphi(x), \psi(y)\rangle$ for all $x, y \in V$. In addition, K is PSD if and only if C is PSD, and then we can set $\psi = \varphi$.*

The following is the main theorem of the section, and it verifies the estimate (2.3). It shows that we may upper bound the rank of a polynomial kernel by the dimension of the vector space $\mathbb{R}[V]_m^G$, consisting of G-invariant polynomials of degree at most m.

Theorem 2.2.7 (Rank and polynomial invariants)**.** *Suppose Assumption 2.2.4 is satisfied. Then, the following inequalities hold:*

$$\operatorname{rank} K \le \dim\left(\operatorname{span}\{f_1^*, \dots, f_N^*\}\right) = \dim(\mathbb{R}[V]_m^G). \tag{2.10}$$

Note that $\dim(\text{span}\{f_1^*, \dots, f_N^*\})$ is upper bounded by $|\{f_1^*, \dots, f_N^*\}|$, the number of distinct symmetrizations of the basis elements $f_1, \dots, f_N$. While it is not true in general that the number of $G-$symmetrizations of basis elements coincides with $\dim(\mathbb{R}[V]_m^G)$ (see Illustration 1), the two quantities do agree for permutation groups applied to the monomial basis, and so for our examples, it suffices to count the orbits of monomials under the relevant permutation actions. This is the content of the following lemma and illustration.

Example 1. Suppose $\mathbb{Z}/4\mathbb{Z}$ acts on $\mathbb{R}^2$ by 90° counterclockwise rotation, where the primitive generator sends $(a, b) \mapsto (-b, a)$. Consider the space of degree 2 polynomials $\mathbb{R}[x_1, x_2]_2$ with the basis $\{1, x_1, x_2, x_1^2, x_1^2 + x_2^2, x_1x_2\}$. For all $g \in \mathbb{Z}/4\mathbb{Z}$, the equality $g \cdot (x_1^2 + x_2^2) = x_1^2 + x_2^2$ holds and so $(x_1^2 + x_2^2)^* = x_1^2 + x_2^2$. Note that $(x_1^2)^* = \frac{1}{4}(x_1^2 + x_2^2 + x_1^2 + x_2^2) = \frac{1}{2}(x_1^2 + x_2^2)$. It follows that $(x_1^2)^*$ and $(x_1^2 + x_2^2)^*$ are not equal but are linearly dependent. So, the number of $G-$symmetrizations of the basis is not equal to the dimension of the span of symmetrizations.

Lemma 2.2.8. *Suppose $G \subset \mathfrak{S}_d$ acts on $V = \mathbb{R}^d$ by permutating coordinates and let $f_1, \dots, f_N$ be the monomial basis of $\mathbb{R}[V]_m$. Then,*

$$|\{f_1^*, \dots, f_N^*\}| = \dim(\text{span}\{f_1^*, \dots, f_N^*\}) = \dim(\mathbb{R}[V]_m^G).$$

Proof. Let $f_{i_1}^*, \dots, f_{i_r}^*$ be representatives of all $G-$orbits. As G acts via permutation and each f_j is a monomial, each $f_{i_j}^*$ is a linear combination of monomials, with each monomial in $\mathbb{R}[V]_m$ appearing in exactly one $f_{i_j}^*$. Since all monomials are linearly independent, it follows that any nontrivial linear combination $\sum_{j=1}^r \alpha_i f_{i_j}^*$ is not equal to zero. Thus, the set $\{f_{i_1}^*, \dots, f_{i_r}^*\}$ is linearly independent and the desired equalities hold. □

2.3 The four examples

Concrete applications of Theorem 2.2.7 require a few combinatorial constructions, which we now recall and use to prove the bounds in Examples 2.1.1-2.1.4.

2.3.1 Partitions and compositions.

Fix some positive integer $j \in \mathbb{N}$. A *composition* of j is an ordered sequence of strictly positive integers $(\alpha_1, \alpha_2, ...)$ that sum to j. If the values α_i are required to only be nonnegative, then the sequence is called a *weak composition.* If the number of terms in the sequence is d, then the sequence is called a *(weak) composition of j into d parts.* We let the set $C(j)$ consist of all compositions of j and we let $C_d(j)$ consist of all weak compositions of j into d parts.

The cardinality of these two sets will be written as

$$c(j) := \text{Card } C(j) \qquad \text{and} \qquad c_d(j) := \text{Card } C_d(j).$$

In particular, $c(j)$ and $c_d(j)$ can be explicitly written as

$$c(j) = 2^{j-1} \qquad \text{and} \qquad c_d(j) = \binom{j+d-1}{j}.$$

A *partition of* j is a composition of j with ordered elements $\alpha_1 \geq \alpha_2 \geq \dots$. A *partition of* j *into* d *parts* is defined analogously. We denote by $P(j)$ the set of all partitions of j and by $P_d(j)$ the set of all partitions of j with at most d parts. The cardinality of these two sets will be written as

$$p(j) := \text{Card } P(j) \qquad \text{and} \qquad p_d(j) := \text{Card } P_d(j).$$

We will use the shorthand $\alpha \vdash j$ to mean that α is a partition of j, while the number of parts of α will be denoted by $|\alpha|$. It is sometimes convenient to identify the partition $\alpha \vdash j$ as a multiset $\alpha = \{1^{\mu_1}, 2^{\mu_2}, \dots, j^{\mu_j}\}$. We denote the multipilicities of parts of α by the vector $\mu(\alpha) = (\mu_1(\alpha), \dots, \mu_j(\alpha))$ and the product of factorials of the multiplicities by $\eta(\alpha) = \prod_{i=1}^{j} \mu_i(\alpha)!$.

By convention, the number of compositions and partitions of zero is equal to 1.

Example 2. As a concrete example, 4 can be partitioned in 5 distinct ways:

$$4 \qquad 3+1 \qquad 2+2 \qquad 2+1+1 \qquad 1+1+1+1.$$

Consequently, we have $p_1(4) = 1$, $p_2(4) = 3$, $p_3(4) = 4$, and $p_4(4) = 5$. Arbitrarily permuting the elements within each partition yields a composition. Padding each composition with zeros yields a weak composition.

We will now give a simple direct proof of Theorem 2.1.1. The result also follows from facts about the invariant ring $\mathbb{R}[V]^{\mathfrak{S}_d}$ that will require introducing algebraic results that are not necessary in this chapter, see for example [34].

2.3.2 *Verification of Example 2.1.1*

We plan to apply Theorem 2.2.7 and Lemma 2.2.8. Assume that $V = \mathbb{R}^d$ and fix a monomial basis of $\mathbb{R}[V]_m$. Each monomial in $\mathbb{R}[V]_m$ can be written as $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}$ for some set of nonnegative integers $\alpha_1, \dots, \alpha_d$. The degree of the monomial is $j := \sum_{i=1}^d \alpha_i \leq m$. Since the action of $\mathfrak{S}_d$ on $\mathbb{R}^d$ is given by coordinate permutations, the action of $\mathfrak{S}_d$ of $\mathbb{R}[V]_m$ is given by a permutation of the monomials. Namely, for all $\sigma \in \mathfrak{S}_d$, the equality holds:

$$\sigma \cdot x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} = x_{\sigma(1)}^{\alpha_1} x_{\sigma(2)}^{\alpha_2} \dots x_{\sigma(d)}^{\alpha_d} = x_1^{\alpha_{\sigma(1)}} x_2^{\alpha_{\sigma(2)}} \dots x_d^{\alpha_{\sigma(d)}}.$$

Consequently, the $\mathfrak{S}_d-$orbit of a degree j monomial x^α can be labeled by a partition of j into at most d parts. Thus, the number of $\mathfrak{S}_d-$orbits of the monomial basis of $\mathbb{R}[V]_m$ is $\sum_{j=0}^m p_d(j)$. Noting $p_d(j) \leq p(j)$ when $d \leq j$ and $p_d(j) = p(j)$ for $d > j$ and applying Theorem 2.2.7 completes the proof.

2.3.3 *Verification of Example 2.1.2*

Recall that we are interested in the group of permutations $\mathfrak{S}_d$ acting on matrices $X \in \mathbb{R}^{d \times k} = V$ by permuting rows. Fix a monomial basis of $\mathbb{R}[V]_m$. It is straightforward to see that each monomial X^β can be labeled by a matrix of nonnegative integers $\beta \in \mathbb{R}^{d \times k}$ whose i'th row β_i is a weak composition of some integer j_i into k parts and we have $\sum_{i=1}^d j_i = j \leq m$. By permuting the rows, we may find at least one monomial in the $\mathfrak{S}_d$-orbit of X^β so that $j_1 \geq j_2 \geq \dots \geq j_d$. Note that when $d \geq m$, all nonzero entries of β will be in the first m rows. We identify the row sums $(j_1, \dots, j_d)$ with a partition of j into at most d parts. Identifying the partition $\alpha \vdash j$ with the multiset $\alpha = \{1^{\mu_1}, \dots, j^{\mu_j}\}$, each $\mathfrak{S}_d-$orbit representative with row sums α corresponds to a distinct choice of μ_i weak compositions of i for each $i \leq j$. It follows that the number of $\mathfrak{S}_d-$orbits is given by $\sum_{j=0}^m \sum_{\alpha \vdash j} \prod_{i=1}^j \binom{c_k(i)+\mu_i(\alpha)-1}{\mu_i(\alpha)}$ and is independent of d for all $d \geq m$, as claimed. Applying Theorem 2.2.7 completes the proof.

2.3.4 *Verification of Example 2.1.3*

Each monomial $X^\beta \in \mathbb{R}[V]_m$ corresponds to a matrix $\beta \in \mathbb{R}^{d \times d}$ with each entry β_{ij} being the exponent of x_{ij} in the monomial. Since every adjacency matrix is symmetric, in each monomial, x_{ij} can be replaced by x_{ji} for all pairs i, j. It follows that we need only consider symmetric matrices of exponents since we can replace the ij^{th} entry of the exponent matrix β with $\frac{1}{2}(\beta_{ij} + \beta_{ji})$, making some entries of the matrix odd multiples of $\frac{1}{2}$ and not integers.

Next, let j be the degree of X^β and let α be the vector of row sums of β, that is $\alpha_i = \sum_j \beta_{ij}$. By simultaneously permuting the rows and columns of X, equivalently those of β, we can find a monomial in the $\mathfrak{S}_d$—orbit of X^β such that α is ordered, meaning $\alpha_1 \geq \alpha_2 \ldots \geq \alpha_d$. When $d \geq 2m$, we must have that $\alpha_i = 0$ for all $i > 2m$ and so we need only consider $2m \times 2m$ matrices. Noting that 2α is a partition of $2j$, we deduce that the number of $\mathfrak{S}_d$—orbits is upper bounded by $\sum_{j=0}^{m} \sum_{\alpha \vdash 2j} q_\alpha$ where q_α is the number of symmetric non-negative integer matrices with row sums equal to α and is independent of d when $d \geq 2m$. An application of Theorem 2.2.7 completes the proof.

2.3.5 Verification of Example 2.1.4

Recall that the data points are matrices $X \in \mathbb{R}^{d \times k}$. The first fundamental theorem of invariant theory for the orthogonal group [105, Theorem 2.9.A] shows that any degree m polynomial in $\mathbb{R}[V]$ that is invariant under simultaneous orthogonal action on the rows x_i is a degree $m/2$ polynomial of the pairwise inner-products $\langle x_i, x_j \rangle$. That is, any such polynomial $f(x) = \tilde{f}(A(x))$ where $A(x) = [\langle x_i, x_j \rangle] \in \mathbb{R}^{d \times d}$. If the polynomial f is in addition invariant under relabeling of the rows, we can find a representing element for its G-orbit such that $\tilde{f}$ is invariant under conjugation by permutations. A completely analogous argument to the proof in Theorem 2.1.3 shows that the number of such invariants is bounded by $\sum_{j=0}^{m/2} \sum_{\alpha \vdash 2j} q_\alpha$. Invoking Theorem 2.2.7 completes the proof.

2.4 Proof of Theorem 2.1.5 and Consequences

In this section, we will prove Theorem 2.1.5. The arguments will be based on the theory of *characters* from representation theory. This is a systematic approach that applies to any finite group or, more generally, to any compact Lie group G acting continuously on an Euclidean space. See [33, 92, 94], for the representation theory facts used in this section.

Suppose that a group G acts linearly on a vector space V. This action can be summarized by a map (group homomorphism) $\rho \colon G \to \mathrm{GL}(V)$, called a *representation of* G. By choosing a basis for V, we may identify $\rho(g)$ with a matrix M_g so that for any $x \in V$

$$gx := \rho(g)(x) = M_g x.$$

Definition 2.4.1 (Character)**.** The *character* of ρ is the function $\chi : G \to \mathbb{R}$ defined as

$$\chi(g) := \mathrm{Trace}(M_g).$$

The main use of characters for us is to compute the dimension of the invariant space

$$V^G := \{x \in V : gx = x \quad \forall g \in G\}.$$

This is the content of the following lemma, an easy consequence of [92, Corollary 2.4.1].

Lemma 2.4.2 (Invariants and characters)**.** *Let G be a compact Lie group acting linearly and continuously on a vector space V. Then,*

$$\dim(V^G) = \mathop{\mathbb{E}}_{g \sim \mathrm{Unif}\, G} \chi(g), \tag{2.11}$$

where the expectation is taken with respect to the Haar probability measure.

In particular, one can in principle estimate $\dim(V^G)$ by sampling elements $g \in G$ according to the Haar measure. In the case when G is a finite group acting linearly on a vector space V, the expression (2.11) simply becomes

$$\dim(V^G) = \tfrac{1}{|G|} \sum_{g \in G} \chi(g) = \mathrm{Trace}\left(\tfrac{1}{|G|} \sum_{g \in G} M_g \right).$$

We next aim to apply Lemma 2.4.2 in order to compute $\dim(\mathbb{R}[V]_m^G)$. In order to do so, we need to describe how the representation ρ of G on V promotes to a representation ρ' of G on $\mathbb{R}[V]_m$. Let N denote the size of the (monomial) vector space basis of $\mathbb{R}[V]_m$. Since G acts linearly on $\mathbb{R}[V]_m$ we may write

$$g \cdot f = M'_g \cdot [f],$$

where $[f] \in \mathbb{R}^N$ is the vector of coefficients of f in the monomial basis and M'_g is an $N \times N$ matrix. Applying Lemma 2.4.2 therefore yields the expression

$$\dim(\mathbb{R}[V]_m^G) = \mathop{\mathbb{E}}_{g \sim \mathrm{Unif}\, G} \mathrm{Trace}(M'_g). \tag{2.12}$$

Let us look at a concrete example that illustrates the expression (2.12).

Example 3. Consider the permutation group $G = \mathfrak{S}_2 = \{e, g\}$ acting on $\mathbb{R}^2$, where e is the identity permutation and $g = (12)$. Then the representing matrices are

$$\rho(e) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad \rho(g) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Picking $m = 2$, the space $\mathbb{R}[x_1, x_2]_2 = \text{span}\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}$ has dimension $N = 6$. The representation ρ' of G on $\mathbb{R}[x_1, x_2]_2$ has the form

$$M'_e = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1, \end{bmatrix}, \quad \text{and} \quad M'_g = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The character χ' of ρ' is then given by $\chi'(e) = \text{Trace}(\rho'(e)) = 6$ and $\chi'(g) = \text{Trace}(\rho'(g)) = 2$. By (2.12), we expect $\dim(\mathbb{R}[x_1, x_2]_2^G) = 4$. Indeed, a basis for $\mathbb{R}[x_1, x_2]_2^G$ is given by $\{1, x_1 + x_2, x_1^2 + x_2^2, x_1x_2\}$.

More generally, in order to apply (2.12) we need to compute the trace of the representing matrix M'_g. This can be explicitly calculated directly from the action of G on V. As shown in [78, Equation 8] and [63, Section 2.12], the entries of M'_g are indexed by pairs of weak compositions $\alpha, \beta \in C_d(k)$ of all integers $k \leq m$ into d parts. In particular, the diagonal entries of M'_g are given by

$$(M'_g)_{\alpha\alpha} = \frac{\text{per}M_g[\alpha]}{\eta(\alpha)}. \tag{2.13}$$

Here, $M_g(\alpha)$ is the principle submatrix of $M_g \in \mathbb{R}^{d\times d}$ indexed by $\alpha \in \mathbb{R}^d$, the symbol $\text{per}M_g[\alpha]$ denotes the permanent of the matrix $M_g[\alpha]$, and $\eta(\alpha)$ is the product of the factorials of the multiplicities of entries in α. Summing the expressions (2.13) across α yields the second equality in (2.5).

Alternatively, we may express the trace of M'_g as the sum of its eigenvalues. To this end, let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of M_g. Then it is shown in [63, Section 2.15.14] and [78, Section 3] that the eigenvalues of M'_g are $\lambda'_\alpha = \prod_{i=1}^d \lambda_i^{\alpha_i}$ where $\alpha \in C_d(k)$ is some weak composition of $k \leq m$ into d parts. Summing eigenvalues yields the first equality in (2.5).

A consequence of Lemma 2.4.2 is a lower bound on the dimension $\dim(\mathbb{R}[V]_m^G)$ of invariant polynomials where V is $d-$dimensional and $G \subset \mathfrak{S}_d$ acts by permuting basis elements of V.

Corollary 2.4.3. *Suppose that the finite group $G \subset \mathfrak{S}_d$ (with identity element 1) acts on the $d-$dimensional vector space V by permuting basis elements. Then, the inequality holds:*

$$\dim \mathbb{R}[V]_m^G > \frac{\dim \mathbb{R}[V]_m}{|G|} > \frac{d^m}{m^m|G|}.$$

Proof. Note that since 1 is the identity element of G, the action of 1 on $\mathbb{R}[V]_m$ is represented by the identity matrix and thus $\chi'(1) = \dim \mathbb{R}[V]_m$. As V is $d-$dimensional, the dimension is equal to the number of monomials in d variables with degree at most m :

$$\chi'(1) = \dim \mathbb{R}[V]_m = \sum_{j=0}^{m} \binom{m+d-1}{d-1}.$$

Moreover, since G acts by permuting basis elements of V, G acts on $\mathbb{R}[V]_m$ by permuting monomials and thus each element $g \in G$ is represented by some permutation matrix. In particular, $\chi'(g)$ is equal to the number of monomials fixed by g. As all constant monomials are fixed by each element $g \in G$, the inequality $\chi'(g) \geq 1$ holds for each g. We compute

$$\begin{aligned}
\dim \mathbb{R}[V]_m^G &= \frac{1}{|G|} \sum_{g \in G} \chi'(g) \\
&> \frac{1}{|G|} \chi'(1) \\
&= \frac{1}{|G|} \dim \mathbb{R}[V]_m \\
&> \frac{1}{|G|} \binom{m+d-1}{d-1}.
\end{aligned}$$

Noting that $\binom{m+d-1}{d-1} > \frac{d^m}{m^m}$ completes the proof. □

It follows that in order for the dimension $\mathbb{R}[V]_m^G$ to stabilize, we must have that the size of the group $|G|$ scales at approximately the same rate as $\dim \mathbb{R}[V]_m$. As another illustration of Lemma 2.4.2, we now use it to compute $\dim \mathbb{R}[V]_m^G$ for a cyclic group $G \subset \mathfrak{S}_d$.

Lemma 2.4.4 (Invariants under cyclic permutations)**.** *Let $G = \mathbb{Z}/d\mathbb{Z}$ act on $V = \mathbb{R}^d$ by cyclic permutation. Then, assuming $m < d$, the following equality holds:*

$$\dim(\mathbb{R}[V]_m^G) = 1 + \frac{1}{d} \left(\sum_{r:\ s=\gcd(r,d)>1} \sum_{j=1}^{\lfloor ms/d \rfloor} \binom{j+s-1}{s-1} \right).$$

Proof. Let σ be a $d-$cycle generating $\mathbb{Z}/d\mathbb{Z}$. Then, the equality

$$\dim \mathbb{R}[V]_m^G = \frac{1}{d} \sum_{r=1}^{d} \chi'(\sigma^r)$$

holds by Lemma 2.4.2. So, we need to calculate the trace of each element σ^r acting on $\mathbb{R}[V]_m$.

Note that each element σ^r permutes the monomials in $\mathbb{R}[V]_m$ and so the trace $\chi'(\sigma^r)$ is equal to the number of monomials fixed by the permutation σ^r. As the monomial 1 is fixed by all permutations, we will now focus on monomials of degree at least 1.

Note that for the $d-$cycle σ, a monomial is fixed if and only if it is of the form $x_1^\alpha x_2^\alpha ... x_d^\alpha$ for some integer α. This monomial has total degree $d \cdot \alpha$. By assumption the dimension d is larger than the degree m of polynomials being considered and so no nontrivial monomials are fixed by $d-$cycles. Thus, $\chi'(\sigma) = 1$. Moreover, for every r with $\gcd(r, d) = 1$ we have that σ^r is also a $d-$cycle and so $\chi'(\sigma^r) = 1$.

If $\gcd(r, d) > 1$ the permutation $\sigma^r = \tau_1 ... \tau_s$ decomposes as a product of $s := \gcd(r, d)$ disjoint cycles of length $\ell := d/s$. In order for a monomial to be fixed by σ^r, exponents corresponding to the entries of each cycle τ_i must all be equal and so each monomial of degree j fixed by σ^r corresponds to a weak composition of j into s parts such that every part is divisible by ℓ. Let $C_s^\ell(j)$ be the set of weak compositions of j into s parts divisible by ℓ. Note that in order for $C_s^\ell(j)$ to be non-empty, we must have that j is divisible by ℓ. Using a simple stars and bars argument, we have $|C_s^\ell(j)| = |C_s(j/\ell)| = \binom{j/\ell+s-1}{s-1}$. It follows that the equality

$$\chi'(\sigma^r) = 1 + \sum_{j=1}^{m} |C_s^\ell(j)| = 1 + \sum_{j=1}^{\lfloor m/\ell \rfloor} \binom{j+s-1}{s-1},$$

holds. Thus, altogether we have that the equality

$$\dim \mathbb{R}[V]_m^G = \frac{1}{d} \sum_{j=1}^{d} \chi'(\sigma^j) = \frac{1}{d} \left(d + \sum_{r:\ s=\gcd(r,d)>1} \sum_{j=1}^{\lfloor ms/d \rfloor} \binom{j+s-1}{s-1} \right)$$

holds as desired. □

We now use Lemma 2.4.4 to calculate $\dim \mathbb{R}[V]_m^{\mathbb{Z}/d\mathbb{Z}\times\mathbb{Z}/d\mathbb{Z}}$ which bounds the rank of kernels on the space $V = \mathbb{R}^{d\times d\times k}$ of $d \times d$ images. Images are represented as 2D arrays $X = (x_{ij})_{ij=1}^d$ of pixels where each entry x_{ij} is a vector in $\mathbb{R}^k$ for some dimension k. For example, grayscale images can be represented as $d \times d$ matrices where the value of each entry determines how dark the corresponding pixel is ($k = 1$), and RGB images are represented as 2D arrays with entries in $\mathbb{R}^3$ determining the red, green, and blue values for each pixel ($k = 3$).

Setting $k = 1$, note that we can capture the effect of translating a $d \times d$ image in an infinite plane by the action of $\mathbb{Z}/d\mathbb{Z} \times \mathbb{Z}/d\mathbb{Z}$ on the rows and columns of the image so that the translated image stays in the original $d \times d$ frame. This is important since a kernel may

only take in a $d \times d$ dimensional input. Functions of images should be invariant under the set of translations since any translation of an image does not change its content. The group of translations is $\mathbb{Z}/d\mathbb{Z} \times \mathbb{Z}/d\mathbb{Z}$ with the 2D cyclic permutation $(\sigma, \tau) \in \mathbb{Z}/d\mathbb{Z} \times \mathbb{Z}/d\mathbb{Z}$ acting on $X = (x_{ij})_{i,j=1}^{d}$ by $(\sigma, \tau) \cdot X = \sigma X \tau^{\top}$. Note that while we concentrate on grayscale images, the same argument could be used to give an analogous bound for the rank of kernels on other spaces of images, such as RGB images.

Theorem 2.4.5 (2D-cyclic permutations)**.** *Let the group $G = \mathbb{Z}/d\mathbb{Z} \times \mathbb{Z}/d\mathbb{Z}$ act on matrices $V = \mathbb{R}^{d \times d}$ by conjugation, that is $(\sigma, \tau) \cdot X = \sigma X \tau^{\top}$. Then, the following equality holds:*

$$\dim \mathbb{R}[V]_m^G = \frac{1}{d^2} \sum_{r=1}^{d} \sum_{r'=1}^{d} \left(\sum_{j=0}^{\lfloor mss'/d^2 \rfloor} \sum_{\lambda \in C_s^{d^2/(ss')}(jss'/d^2)} \prod_{i=1}^{s} \binom{\lambda_i ss'/d^2 + s' - 1}{s' - 1} \right)$$

where for each r, r' we define $s := \gcd(r, d)$ and $s' := \gcd(r', d)$.

Proof. Let σ be a $d-$cycle generating $\mathbb{Z}/d\mathbb{Z}$. Then, the equality

$$\dim \mathbb{R}[V]_m^G = \frac{1}{d^2} \sum_{r=1}^{d} \sum_{r'=1}^{d} \chi'(\sigma^r, \sigma^{r'})$$

holds by Lemma 2.4.2. As in Lemma 2.4.4, the character $\chi'(\sigma^r, \sigma^{r'})$ is equal to the number of monomials in $\mathbb{R}[V]_m$ that are fixed by the pair of permutations $(\sigma^r, \sigma^{r'})$. To each monomial X^{α} in $\mathbb{R}[V]_m$ we associate the matrix of exponents α and X^{α} is fixed by $(\sigma^r, \sigma^{r'})$ if and only if $\alpha = M_{\sigma^r} \alpha (M_{\sigma^{r'}})^{\top}$. So, the number of monomials of degree j that are fixed by $(\sigma^r, \sigma^{r'})$ is equal to the number of non-negative matrices whose entries sum to j that are invariant under the action of $(\sigma^r, \sigma^{r'})$.

Let $s = \gcd(r, d)$ and $s' = \gcd(r', d)$. Then, $\sigma^r = \sigma_1 \dots \sigma_s$ decomposes as a product of s disjoint cycles of length d/s and $\sigma^{r'}$ decomposes as a product of $s' = \tau_1 \dots \tau_{s'}$ disjoint cycles of length d/s'. In order for a matrix α to be invariant under $(\sigma^r, \sigma^{r'})$, the rows corresponding to each σ_i must be equal and the columns corresponding to each τ_i must be equal. It follows that the row sums of α must correspond to a weak composition $\lambda = (\lambda_1, \dots, \lambda_s)$ of j into s parts divisible by d/s. Then, the rows corresponding to σ_i are all equal, have entries that sum to $\lambda_i/(d/s)$ and are invariant under column permutations by $\sigma^{r'}$. So, each row corresponding to σ_i is a weak composition of $\lambda_i s/d$ into s' parts divisible by d/s'. It follows that λ_i must be divisible by $d^2/(ss')$ and so λ is a weak composition of j into s parts divisible by $d^2/(ss')$.

Thus, the total number of degree j monomials fixed by $(\sigma^r, \sigma^{r'})$ is given by the sum

$$\sum_{\lambda \in C_s^{d^2/(ss')}(j)} \prod_{i=1}^{s} C_{s'}^{d/s'}(\lambda_i s/d) = \sum_{\lambda \in C_s^{d^2/ss'}(j)} \prod_{i=1}^{s} \binom{\lambda_i ss'/d^2 + s' - 1}{s' - 1}.$$

So, the equality holds:

$$\chi'(\sigma^r, \sigma^{r'}) = \sum_{j=0}^{m} \sum_{\lambda \in C_s^{d^2/(ss')}(j)} \prod_{i=1}^{s} \binom{\lambda_i ss'/d^2 + s' - 1}{s' - 1}.$$

In order for the set $C_s^{d^2/ss'}(j)$ to be non-empty we must have that j is divisible by d^2/ss' and so this is equivalent to:

$$\chi'(\sigma^r, \sigma^{r'}) = \sum_{j=0}^{\lfloor mss'/d^2 \rfloor} \sum_{\lambda \in C_s^{d^2/(ss')}(jss'/d^2)} \prod_{i=1}^{s} \binom{\lambda_i ss'/d^2 + s' - 1}{s' - 1}.$$

Then, by Lemma 2.4.2, we have

$$\dim \mathbb{R}[V]_m^G = \frac{1}{d^2} \sum_{r=1}^{d} \sum_{r'=1}^{d} \left(\sum_{j=0}^{\lfloor mss'/d^2 \rfloor} \sum_{\lambda \in C_s^{d^2/(ss')}(jss'/d^2)} \prod_{i=1}^{s} \binom{\lambda_i ss'/d^2 + s' - 1}{s' - 1} \right)$$

The proof is complete. □

We now use the equality in Theorem 2.4.5 to show that $\dim \mathbb{R}[V]_m^{\mathbb{Z}/d\mathbb{Z} \times \mathbb{Z}/d\mathbb{Z}}$ does not stabilize and is polynomial in d. The right hand side of the equality in Theorem 2.4.5 is minimized when d is a prime number and m is less than d. If d is prime and σ generates $\mathbb{Z}/d\mathbb{Z}$, then for any choice of $r, r' < d$, the permutations σ^r and $\sigma^{r'}$ are both $d-$cycles. In order for a monomial X^α of degree j to be fixed by $(\sigma^r, \sigma^{r'})$, all entries in the exponent matrix α must be equal. However, since we assume that $j < d^2$, this condition can only be satisfied when $j = 0$ and so $\chi'(\sigma^r, \sigma^{r'}) = 1$. If exactly one of r, r' is equal to d, the monomial X^α is fixed by $(\sigma^r, \sigma^{r'})$ is fixed if and only if all rows/columns in α are equal. Since the degree of the monomial is less than d, this is once again only true when all entries of α are zero and so $\chi'(\sigma^r, \sigma^{r'}) = 1$. Thus, $\chi'(\sigma^r, \sigma^{r'}) > 1$ if and only if $r = r' = d$ and so $(\sigma^r, \sigma^{r'})$ is the identity. Since every monomial in $\mathbb{R}[V]_m$ is fixed by the identity, it follows that the

equality $\chi(\sigma^d, \sigma^d) = \dim \mathbb{R}[V]_m = \sum_{j=0}^{m} \binom{j+d^2-1}{d^2-1}$ holds. Then, we compute the bound

$$\begin{aligned}\dim \mathbb{R}[V]_m^{\mathbb{Z}/d\mathbb{Z}\times\mathbb{Z}/d\mathbb{Z}} &= \frac{1}{d^2}\left(d^2 + \sum_{j=1}^{m}\binom{j+d^2-1}{d^2-1}\right) \\ &> \frac{d^{2m-2}}{m^m}.\end{aligned}$$

It follows that $\dim \mathbb{R}[V]_m^{\mathbb{Z}/d\mathbb{Z}\times\mathbb{Z}/d\mathbb{Z}}$ is polynomial in d and does not stabilize.

2.5 Representation stability and free kernels

We have seen in several interesting and relevant examples that the rank of G—invariant kernels becomes dimension-independent because $\dim(\mathbb{R}[V]_m^G)$ is independent of $\dim V$. This stabilization can be understood in the context of a broader phenomenon called *representation stability* [20, 19]. In this section, we introduce some of the key definitions from the literature on representation stability and show that our examples fit into this framework. Importantly, we will be able to parametrize invariant kernels in a way that enables comparing data from different dimensions of interest. As we will see in the next section, this is useful for learning problems where we have access to training data in multiple dimensions and would like to learn a function that can take inputs across dimensions.

2.5.1 Representation stability background.

The basic objects of study are a sequence of groups $\{G_d\}_{d\in\mathbb{N}}$ acting on a sequence of Euclidean spaces $\{V_d\}_{d\in\mathbb{N}}$, and satisfying certain compatibility conditions. In particular, we assume that G_d acts on V_d by orthogonal transformations. The following is the formal definition; see (2.14) for the setup.

$$\cdots \underset{\gamma_{d-1}^*}{\overset{\gamma_{d-1}}{\rightleftarrows}} \underset{\curvearrowright G_d}{V_d} \underset{\gamma_d^*}{\overset{\gamma_d}{\rightleftarrows}} \underset{\curvearrowright G_{d+1}}{V_{d+1}} \underset{\gamma_{d+1}^*}{\overset{\gamma_{d+1}}{\rightleftarrows}} \underset{\curvearrowright G_{d+2}}{V_{d+2}} \underset{\gamma_{d+2}^*}{\overset{\gamma_{d+2}}{\rightleftarrows}} \cdots \tag{2.14}$$

Definition 2.5.1 (Consistent sequence)**.** Let $G_1 \subset G_2 \subset \ldots$ be a nested sequence of compact groups acting on a sequence of Euclidean spaces $V_1, V_2, \ldots$. Suppose moreover, that there exist linear isometries $\gamma_d : V_d \to V_{d+1}$ that are G_d—equivariant, meaning

$$\gamma_d(gx) = g\gamma_d(x) \qquad \forall x \in V_d,\ g \in G_d.$$

Then the sequence of pairs $\mathcal{V} := \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ is called a $\{G_d\}$-*consistent sequence*.

We can identify V_d with $\gamma_d(V_d) \subseteq V_{d+1}$, in which case the orthogonal projection from V_{d+1} onto V_d coincides with the adjoint of γ_d [57, Section 1.1.1]:

$$\gamma_d^* = \mathrm{proj}_{V_d} \qquad \forall d \in \mathbb{N}.$$

We say that a sequence $\{x_d \in V_d\}$ is a *free element* of $\{V_d\}$ if the following holds:

$$x_d = \mathrm{proj}_{V_d}(x_{d+1}) \qquad \forall d \in \mathbb{N}.$$

Thus, points of a free element in higher dimensions project onto points in lower dimensions.

Importantly, when the groups G_d are in a sense "sufficiently large", the invariant spaces $V_d^{G_d}$ become isomorphic to each other and hence of the same dimension. Let us give a name to this situation.

Definition 2.5.2 (Stability degree)**.** Given a consistent sequence $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$, its *stability degree* is the smallest $D \in \mathbb{N}$ such that the maps proj_{V_d} restrict to an isomorphism between $V_{d+1}^{G_{d+1}}$ and $V_d^{G_d}$ for all $d \geq D$.

We will focus on one sufficient condition, from [57, Definition 2.2], which ensures that the stability degree is finite.

Definition 2.5.3 (Generation degree)**.** The consistent sequence $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ is *finitely generated* if there exists $D \in \mathbb{N}$ such that the equality

$$\mathbb{R}[G_d]V_D = V_d \qquad \text{holds for all } d \geq D.$$

The *generation degree* of $\mathcal{V}$ is the smallest such D.

Recall that $\mathbb{R}[G_d] = \{\sum_{i=1}^r \alpha_i g_i \mid r \in \mathbb{N}, \text{and } \forall i \in [r], \alpha_i \in \mathbb{R}, g_i \in G_d\}$ is the set of linear combinations of finitely many elements in G_d, while $\mathbb{R}[G_d]V_D = \{\rho \cdot v \mid \rho \in \mathbb{R}[G_d], v \in V_d\}$ is the set of all products between $\mathbb{R}[G_d]$ and V_D. Here is a small example.

Example 4. Take $G_d = \mathfrak{S}_d$ acting by permuting coordinates of $V_d = \mathbb{R}^d$ and let $\gamma_d \colon \mathbb{R}^d \to \mathbb{R}^{d+1}$ be the zero padding map that sends $x \mapsto (x, 0)$. Then, the generation degree of $\mathcal{V}$ is one, since $\mathbb{R}$ is identified with $\mathrm{span}\{e_1\} \subseteq \mathbb{R}^d$ and we can generate any vector $v \in \mathbb{R}^d$ via $v = \sum_{i=1}^d v_i g_i \cdot e_1$, where g_i permutes the first and ith coordinates only, i.e., g_i is the transposition $(1i)$. In this example, the stability degree is also one, since $(\mathbb{R}^d)^{\mathfrak{S}_d}$ is the line spanned by the all-ones vector for any d.

Importantly, finite generation ensures that the invariants stabilize [57, Proposition 2.3].

Proposition 2.5.4 (Finite generation implies stability)**.** *If $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ is a consistent sequence generated in degree D, the orthogonal projections $\mathrm{proj}_{V_d} : V_{d+1}^{G_{d+1}} \to V_d^{G_d}$ are injective for all $d \geq D$ and isomorphisms for large d.*

We refer the reader to [57, 56, Appendix B] for a discussion of a possible gap between the stability and generation degrees, although in our main examples they coincide.

A direct corollary of Proposition 2.5.4 is the existence of a *free basis* of invariants.

Lemma 2.5.5 (Free basis)**.** *Let $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ be a consistent sequence with stability degree D. Then there exists $k \in \mathbb{N}$ and a collection of free elements $\{b_i^{(d)}, \dots, b_k^{(d)}\}_{d\in\mathbb{N}}$ such that $b_1^{(d)}, \dots b_k^{(d)}$ forms a basis of $V_d^{G_d}$ for all $d \geq D$.*

In particular, we can parameterize any free invariant element $\{x_d \in V_d^{G_d}\}$ using a finite amount of parameters $\alpha \in \mathbb{R}^k$ by the expression

$$x_d = \sum_i^k \alpha_i b_i^{(d)} \qquad \forall d \geq D,$$

where D is the stability degree. We emphasize that the coefficients α_i fully describe the free-element $\{x_d \in V_d^{G_d}\}$ and are in particular dimension-independent. We will see that this is the underlying reason why polynomial kernels have constant rank in our examples.

2.5.2 Representation stabiliy of polynomial invariants and free kernels.

We now turn to representation stability of polynomial invariants—the main setting of this chapter. To this end, it is elementary to see that any consistent sequence of $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ induces a consistent sequence of polynomial spaces $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}$ upon setting $\gamma_d(f) := f \circ \mathrm{proj}_{V_d}$, meaning that if $x \in \mathbb{R}[V_d]_m$, then $\gamma_d(f)((x, x_{d+1})) = f(x)$. The sequence (2.14) induces the consistent sequence depicted in (2.15).

$$\cdots \underset{\gamma_{d-1}^*}{\overset{\gamma_{d-1}}{\rightleftarrows}} \mathbb{R}[V_d]_m \overset{G_d}{\circlearrowright} \underset{\gamma_d^*}{\overset{\gamma_d}{\rightleftarrows}} \mathbb{R}[V_{d+1}]_m \overset{G_{d+1}}{\circlearrowright} \underset{\gamma_{d+1}^*}{\overset{\gamma_{d+1}}{\rightleftarrows}} \mathbb{R}[V_{d+2}]_m \overset{G_{d+2}}{\circlearrowright} \underset{\gamma_{d+2}^*}{\overset{\gamma_{d+2}}{\rightleftarrows}} \cdots \tag{2.15}$$

In particular, When $\mathcal{F}_m$ has stability degree D, for all $d \geq D$, there exists a free basis $\{f_1^{(d)}, \dots f_k^{(d)} \in \mathbb{R}[V_d]_m^{G_d}\}$ of invariant polynomials. In this case, a free element of $\mathcal{F}_m$ is any

sequence $\{f_d \in \mathbb{R}[V_d]_m\}$ satisfying $f_d(x) = f_{d+1}(\gamma_d(x))$ for all $x \in V_d$ and $d \geq D$. The following simple example illustrates this setup.

Example 5. Recall the setting of Illustration 4 where $G_d = \mathfrak{S}_d$ and $\{V_d = \mathbb{R}^d\}$. The zero padding map $\gamma_d(x) = (x, 0)$ makes $\{(\mathbb{R}^d, \gamma_d)\}$ a consistent sequence. Then, the polynomial vector spaces $\mathbb{R}[V_d]_m$ also form a consistent sequence when paired with the transition map $\gamma_d(f)(x_1, \ldots, x_d, x_{d+1}) = f(x_1, \ldots, x_d)$. In the linear case $m = 1$, the stability degree of the consistent sequence of polynomial spaces is $D = 1$ and has the free basis

$$\left\{ f_1^{(d)}(x) = 1, f_2^{(d)}(x) = \sum_{i=1}^{d} x_i \right\}.$$

We now return to the settings in Examples 2.1.1, 2.1.2, 2.1.3, and 2.1.4 and show that the corresponding sequences $(\mathbb{R}[V_d]_m, \gamma_d)$ are consistent and have finite stability degree. The following result can be derived as a corollary of [56, Theorem A.13]; however, its statement and proof require additional category theory notions. We include a direct proof for completeness.

Theorem 2.5.6 (Stability degree in the running examples)**.** *Let V_d and G_d be the vector spaces and groups appearing in Examples 2.1.1, 2.1.2, 2.1.3, and 2.1.4. Then, for the zero-padding map γ_d, the sequences $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ are consistent and the stability degree of $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}_{d\in\mathbb{N}}$ is bounded by $D = m$, $D = m$, $D = 2m$, and $D = 2m$, respectively.*

Proof. We first prove that each sequence $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ is consistent. By definition, $G_1 \subset G_2 \subset \ldots$ are sequences of compact groups acting linearly on the Euclidean spaces $\{V_d\}$. The zero-padding map γ_d is a linear isometry, and thus we need only check that γ_d is G_d−equivariant. In each of the sequences, the zero-padding map γ_d is equal to $\gamma_d = \mathrm{id}_d \oplus 0$ where id_d is the identity map on V_d. Moreover, as the group G_d is a subgroup of G_{d+1}, the action of G_d on V_{d+1} is given by $g \cdot x = (g \oplus \mathrm{id}_1)(x)$ for all $g \in G_d$ and $x \in V_{d+1}$. Then, for any $x \in V_d$, the equalities $\gamma_d(g \cdot x) = g \cdot x \oplus 0 = (g \oplus \mathrm{id}_1)(x \oplus 0) = g \cdot \gamma_d(x)$. It follows that γ_d is G_d−equivariant.

Now, we show that each consistent sequence $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}_{d\in\mathbb{N}}$ is generated in dimension D, i.e. for all $d \geq D$, the equality $\mathbb{R}[V_d]_m = \mathbb{R}[\mathfrak{S}_d]\mathbb{R}[V_D]_m$. Note that if any consistent sequence is generated in dimension D, it is also generated in dimension $d \geq D$ and thus we need only show that $\mathbb{R}[V_{D+1}]_m = \mathbb{R}[\mathfrak{S}_{D+1}]\mathbb{R}[V_D]_m$. Since any polynomial is a linear combination of monomials, it suffices to show that for each monomial $f_{D+1} \in \mathbb{R}[V_{D+1}]_m$ there exists a group element $g \in G_{D+1}$ such that $f_{D+1} = g \cdot f_D$ for some monomial $f_D \in \mathbb{R}[V_D]_m$.

First, consider the sequence $\{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ appearing in Example 2.1.1. Let $f_{m+1} = \prod_{i=1}^{m+1} x_i^{a_i}$ be a monomial in $\mathbb{R}[V_{m+1}]_m$. If $a_{m+1} = 0$, then $f_{m+1} \in \mathbb{R}[V_m]_m$ so assume that $a_{m+1} > 0$. If $a_{m+1} > 0$, there exists $j \leq m$ such that $a_j = 0$. Define the monomial $f_m = \left(\prod_{i=1}^{m} x_i^{a_i}\right) \cdot x_j^{a_{m+1}} \in \mathbb{R}[V_m]_m$ and let $\sigma \in \mathfrak{S}_{m+1}$ be the transposition defined by $\sigma(j) = m+1$ and $\sigma(m+1) = j$. Then, $f_{m+1} = \sigma \cdot f_m$.

Now, consider the sequence $\{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ appearing in Example 2.1.2. Recall that data points in V_d are matrices $X \in \mathbb{R}^{d\times k}$. Let $f_{m+1} = \prod_{i=1}^{m+1} X_i^{A_i}$ be a monomial in $\mathbb{R}[V_{m+1}]_m$, where each X_i is the vector $[x_{i1}, \ldots, x_{ik}]$ and A_i is the vector $[a_{i1}, \ldots, a_{ik}]$ of exponents. Then $\sum a_{ij} = m$. As in the previous example, if A_{m+1} is the zero vector, $f_{m+1} \in \mathbb{R}[V_m]_m$. Suppose that A_{m+1} is not the zero vector. Then, there exists some $j \leq m$ such that A_j is the zero vector. Define the monomial $f_m = \left(\prod_{i=1}^{m} X_i^{A_i}\right) \cdot X_j^{A_{m+1}} \in \mathbb{R}[V_m]_m$ and let $\sigma \in \mathfrak{S}_{m+1}$ be the transposition defined by $\sigma(j) = m+1$ and $\sigma(m+1) = j$. Then, we have that the equality $f_{m+1} = \sigma \cdot f_m$ holds.

Consider the sequence $\{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ appearing in Example 2.1.3 and choose a monomial $f_{2m+1} = \prod_{i=1}^{2m+1} X_i^{A_i}$ in $\mathbb{R}[V_{2m+1}]_m$. Since elements of V_d are symmetric matrices, we assume without loss of generality that the matrix A with rows A_i is also symmetric. If A_{2m+1} is the zero vector, we have that $f_{m+1} \in \mathbb{R}[V_{2m}]_m$. Suppose that A_{2m+1} is not the zero vector. Then, we have that there exists $j \leq 2m$ such that A_j is the zero vector. As in the previous arguments, define the monomial $f_{2m} = \left(\prod_{i=1}^{2m} X_i^{A_i}\right) \cdot X_j^{A_{2m+1}} \in \mathbb{R}[V_{2m}]_m$ and let $\sigma \in \mathfrak{S}_{2m+1}$ be the transposition defined by $\sigma(j) = 2m+1$ and $\sigma(2m+1) = j$. Then, the equality $f_{2m+1} = \sigma \cdot f_{2m}$ holds.

Finally, consider the sequence $\{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ appearing in Example 2.1.4. Recall that the data points are matrices $X \in \mathbb{R}^{d\times k}$ and the group G_d consists of both simultaneous orthogonal transformations on the rows of X and permutations of the rows of X. Let $f_{2m+1} = \prod_{i=1}^{2m+1} X_i^{A_i}$ be a monomial in $\mathbb{R}[V_{2m+1}]_m$, where each X_i is the vector $[x_{i1}, \ldots, x_{ik}]$ and A_i is the vector $[a_{i1}, \ldots, a_{ik}]$ of exponents. As in the previous example, if A_{2m+1} is the zero vector, $f_{2m+1} \in \mathbb{R}[V_{2m}]_m$. Suppose that A_{2m+1} is not the zero vector. Then, there exists some $j \leq 2m$ such that A_j is the zero vector. Define the monomial $f_{2m} = \left(\prod_{i=1}^{2m} X_i^{A_i}\right) \cdot X_j^{A_{2m+1}} \in \mathbb{R}[V_{2m}]_m$ and let $\sigma \in G_{m+1}$ be the transposition defined by $\sigma(j) = m+1$ and $\sigma(m+1) = j$. Then, the equality $f_{2m+1} = \sigma \cdot f_{2m}$. holds.

Now, since each of the sequences are generated in dimension $D = m, D = m, D = 2m$, and $D = 2m$ respectively, by Proposition 2.5.4, the orthogonal projections $\text{proj}_{\mathbb{R}[V_d]_m} : \mathbb{R}[V_{d+1}]_m^{G_{d+1}} \to \mathbb{R}[V_d]_m^{G_d}$ are injective for all $d \geq D$. However, by Theorems 2.1.1, 2.1.2, 2.1.3, and 2.1.4, we have that the dimensions $\dim \mathbb{R}[V_d]_m^{G_d}$ are constant for $d \geq D$ and so the

orthogonal projections $\text{proj}_{\mathbb{R}[V_d]_m}$ are isomorphisms. It follows that the stability degree of each sequence $\mathcal{F}_m$ is at most D as desired. □

Typical kernels, including those that are polynomials in the inner product, can be used to compare data across different dimensions. For example, the polynomial kernel $K(x,y) = (1+x^\top y)^m$ can be used to compare data points $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$ with $d_1 \leq d_2$ by simply zero-padding x into a vector in $\mathbb{R}^{d_2}$. Formally, we define a *free kernel* as follows.

Definition 2.5.7 (Free Kernel)**.** Let $\mathcal{K} = \{K_d : V_d \times V_d \to \mathbb{R}\}_{d\in\mathbb{N}}$ be a sequence of kernels and let $\mathcal{V} = \{(V_d, \gamma_d)\}_{d\in\mathbb{N}}$ be a consistent sequence. We say that $\mathcal{K}$ is a *free kernel* if for any dimension d and any choice of $x, y \in V_d$, the equality $K_d(x,y) = K_{d+1}(\gamma_d(x), \gamma_d(y))$ holds. Moreover, $\mathcal{K}$ is a free $G-$invariant kernel if in addition, each K_d is G_d-invariant.

Let $\mathcal{K}$ be a free kernel. Recall that Lemma 2.2.5 showed that for each d, we may write

$$K_d(x,y) = F_d(x)^\top C_d F_d(y),$$

where $F_d = (f_1^{(d)}, \dots, f_k^{(d)})^\top$ stacks some basis for $\mathbb{R}[V_d]_m$, the integer k is bounded by the rank of K_d, and $C_d \in \mathbb{R}^{k\times k}$ is a symmetric matrix. Now if $\{(\mathbb{R}[V_d]_m, \gamma_d)\}_{d\in\mathbb{N}}$ is a consistent sequence with finite stability degree D, then by switching basis, we may ensure that $f_1^{(d)}, \dots, f_k^{(d)}$ is a free basis of $\mathbb{R}[V_d]_m$. The following theorem shows that upon this change of basis the matrix C_d becomes independent of d. In words, this means that free kernels admit a free description.

Lemma 2.5.8 (Free kernels admit a free parametrization)**.** *Suppose that $\mathcal{V} = \{(V_d, \gamma_d)\}$ is a consistent sequence and that the induced consistent sequence $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}$ has stability degree D. Let $\mathcal{K} = \{K_d\}_{d\in\mathbb{N}}$ be any free $G-$invariant polynomial kernel of degree at most m. Then, for any free basis $F_d = (f_1^{(d)}, \dots, f_k^{(d)})^\top$ for $\mathcal{F}_m$, there exists a symmetric matrix $C \in \mathbb{R}^{k\times k}$ such that for all $d \geq D$, we can write*

$$K_d(x,y) = F_d(x)^\top C F_d(y) \quad \textit{for all } x, y \in V_d. \tag{2.16}$$

Proof. By Lemma 2.2.5, for each d there exist some symmetric matrices $C_d \in \mathbb{R}^{k\times k}$ such that $K_d(x,y) = F_d(x)^\top C_d F_d(y)$ for all $x, y \in V_d$, and $F_d = (f_1^{(d)}, \dots, f_k^{(d)})^\top$ stacks a basis for $\mathbb{R}[V_d]_m$. Upon switching basis, we may assume that $f_1^{(d)}, \dots, f_k^{(d)}$ is a free basis. Next we show that the matrices C_d are equal for all $d \geq D$. Since $\mathcal{K}$ is a free kernel, for all choices

of x, y in V_d the equality $K_d(x,y) = K_{d+1}(\gamma_d(x), \gamma_d(y))$ holds. Thus, we conclude

$$F_d(x)^\top C_d F_d(y) = K_d(x,y) = F_{d+1}(\gamma_d(x))^\top C_{d+1} F_{d+1}(\gamma_d(y)).$$

Moreover, due to the freeness of the polynomial basis $\{F_d\}_{d\in\mathbb{N}}$, $F_{d+1}(\gamma_d(x)) = F_d(x)$ for all $x \in V_d$, and therefore

$$F_{d+1}(\gamma_d(x))^\top C_{d+1} F_{d+1}(\gamma_d(y)) = F_d(x)^\top C_{d+1} F_d(y).$$

It follows that $C_{d+1} = C_d = ... = C_D$ for all $d \geq D$, as claimed. □

2.6 Generalization of polynomial regression across dimensions

Standard regression problems seek to determine a function p^* on a vector space V from finitely many noisy observations of the form $y_i = p^*(x_i) + \varepsilon_i$. In this section, we consider the problem of learning invariant polynomials from data generated *across different dimensions*. Setting the stage, consider a sequence of groups $\mathcal{G} = \{G_d\}_{d\in\mathbb{N}}$ acting on a sequence of vector spaces $\{V_d\}_{d\in\mathbb{N}}$. Suppose that there exists a sequence of (latent) invariant polynomials

$$p_d^\star \in \mathbb{R}[V_d]_m^{G_d} \qquad \forall d \geq 1,$$

which generate the observation model

$$y = p_d^*(x) + \varepsilon \qquad \text{for } x \in \mathbb{R}^d.$$

We are provided a training sample $S = \bigcup_{d=1}^K S_d$ from this model, where each set $S_d = \{(x_i, y_i)\} \subseteq V_d \times \mathbb{R}$ consists of feature/label pairs in a fixed dimension and the noise values ε_i are independent and identically distributed with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$. We will focus on the fixed design setting wherein the feature vectors x_i are fixed, i.e. not random. We let $n = |S|$ be the total number of training examples and $X = \{x\}_{(x,y)\in S}$ be the collection of feature vectors (in potentially different dimensions) of the training set. Our goal is to predict $p_{\bar{d}}^*$ from S for some arbitrary dimension $\bar{d}$. Clearly, without making further assumptions, this goal is untenable since the invariants $p_d^\star$ may bear no resemblance to $p_{\bar{d}}^\star$ for any $d \neq \bar{d}$.

In order to make the problem tractable, we assume that $\mathcal{V} = \{(V_d, \gamma_d)\}$ is a consistent sequence relative to $\mathcal{G}$ for some transition maps γ_d and that the sequence of invariant spaces $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m^{G_d}, \gamma_d)\}_{d\in\mathbb{N}}$ has a finite stability degree $D < \infty$. Without loss of generality,

we will assume $D = 1$, since we can always truncate the sequence to begin at $d = D$. In particular, $\mathcal{F}_m$ admits a free basis $\{\varphi_1^{(d)}, \dots \varphi_k^{(d)} \in \mathbb{R}[V_d]_m^{G_d}\}_{d\in\mathbb{N}}$ of invariant polynomials. Thus for all $d \geq 1$ the following two key properties hold:

1. (**Basis**) The collection $\varphi_1^{(d)}, \dots, \varphi_k^{(d)}$ forms a basis for $\mathcal{F}_m^d = \mathbb{R}[V_d]_m^{G_d}$.
2. (**Consistency**) Equality holds:

$$\varphi_i^{(d)}(x) = \varphi_i^{(d+1)}(\gamma_d(x)) \qquad \forall x \in V_d,\ i \in \{1, \dots, k\}.$$

In order to link the invariants p_d^* across dimensions, we will assume that $\{p_d^*\}_{d\in\mathbb{N}}$ is a free element of $\mathcal{F}_m$. Thus, we may write

$$p_d^* = \sum_{i=1}^{k} \alpha_i^* \varphi_i^{(d)} \qquad \forall d \geq 1$$

for some vector $\alpha^* \in \mathbb{R}^k$. Henceforth, with a slight abuse of notation, we will drop the dimension d index and simply write $\varphi(x) = (\varphi_1(x), \dots, \varphi_k(x))^\top$ as the dimension index can be inferred from the dimension of x. Define now the matrix

$$\Phi(X) = (\varphi(x_1), \dots, \varphi(x_n)) \in \mathbb{R}^{n\times k}. \tag{2.17}$$

We assume throughout that $\Phi(X)$ has rank k, which in particular entails $n \geq k$.

With this notation, we may write $p_d^*(x) = \langle \varphi_d(x), \alpha^* \rangle$ for each $d \geq 1$ and $x \in V_d$. Therefore, we aim to estimate the free element $\{p_d^*\}_{d\in\mathbb{N}}$ by solving the regression problem:

$$\min_{\alpha\in\mathbb{R}^k} \frac{1}{2n} \sum_{i=1}^{n} \left(\langle \varphi(x_i), \alpha \rangle - y_i\right)^2. \tag{2.18}$$

Note that the solution of (2.18) is simply:

$$\hat{\alpha} = \left(\Phi(X)^\top \Phi(X)\right)^{-1} \Phi(X)^\top y. \tag{2.19}$$

The choice $\hat{\alpha}$ of coefficients induces a sequence of invariant polynomials

$$\hat{p}_d(x) := \langle \varphi_d(x), \alpha \rangle \qquad \forall\ d \in \mathbb{N}.$$

Once more, we will drop the dimension d index for simplicity. We evaluate the performance of this sequence of polynomials using finite test data in a potentially different dimension

$T := \{(x_i, y_i)\} \subseteq \bigcup_{d=1}^{\infty} V_d \times \mathbb{R}$ where once more $y_i = p^\star(x_i) + \varepsilon_i$ with i.i.d. noise ε_i. Given any estimator $\hat{\alpha}$, define the risk

$$R_T^{(\alpha^\star)}(\hat{\alpha}) := \frac{1}{|T|} \sum_{(x,y)\in T} \mathbb{E}\left(\hat{p}(x) - y\right)^2,$$

where the index $\alpha^\star$ indicates the latent sequence of polynomial generating the labels y. Our goal is to establish a bound on the excess risk, $R_T(\hat{\alpha}) - \min_{\alpha\in\mathbb{R}^k} R_T(\alpha)$. To this end, let us introduce a notion of Gram matrices that will play a crucial role: for a finite set $\Delta \subseteq \bigcup_{d=1}^{\infty} V_d$, define the Gram matrix

$$\Sigma_\Delta = \frac{1}{|\Delta|} \sum_{(x,y)\in\Delta} \varphi(x)\varphi(x)^\top.$$

All proofs of results in this section appear in Appendix A.2.

Theorem 2.6.1 (Upper bound). *The estimator $\hat{\alpha}$ in (2.19) is unbiased with excess risk:*

$$R_T^{(\alpha^\star)}(\hat{\alpha}) - \min_{\alpha\in\mathbb{R}^k} R_T^{(\alpha^\star)}(\alpha) = \frac{\sigma^2}{n}\mathrm{Trace}\left(\Sigma_S^{-1}\Sigma_T\right). \tag{2.20}$$

Example 6. Recall the setting of Illustration 5. In this case, the basis size is two. For concreteness, assume that the training data $S \subseteq \mathbb{R}^d \times \mathbb{R}$ lie in some dimension d, while the test set $T \subseteq \mathbb{R}^{\bar{d}} \times \mathbb{R}$ is independent of S and lives in a different dimension $\bar{d}$. Suppose that the feature vectors $\{x\}_{(x,y)\in S}$ in S are drawn i.i.d. from a standard Gaussian distribution $N(0, I_d)$ and similarly the feature vectors in T are i.i.d. with distribution $N(0, I_{\bar{d}})$. A quick computation reveals that

$$\mathbb{E}\Sigma_S = \begin{pmatrix} 1 & 0 \\ 0 & d \end{pmatrix} \quad \text{and} \quad \mathbb{E}\Sigma_T = \begin{pmatrix} 1 & 0 \\ 0 & \bar{d} \end{pmatrix}.$$

Then, for any sufficiently small $\Delta > 0$ the excess risk is bounded by

$$\frac{\sigma^2}{n}\mathrm{Trace}\left(\Sigma_S^{-1}\Sigma_T\right) \le (1+\Delta)\cdot\frac{\sigma^2}{n}\cdot\left(1+\frac{\bar{d}}{d}\right)$$

with probability at least $1 - 4\left(\exp(-58\Delta|S|) + \exp(-58\Delta|T|)\right)$ We defer the proof of this bound to Appendix A.2.2. When $\bar{d} = d$ the upper bound scales like the intrinsic dimension of the parameter space over the number of data points, which is standard for regression

in a fixed dimension [8]. On the other hand, when $\bar{d} > d$, the bound reveals a graceful degradation of the excess risk.

In general, the bound in Theorem 2.6.1 appears difficult to compute explicitly. Nonetheless, we now show that this bound is optimal among any estimation procedure in a minimax sense. Namely, as before, suppose we have a fixed samples of feature vectors for training $X_S \subseteq \cup_{d=1}^{\infty} V_d$ and for testing $X_T \subseteq \cup_{d=1}^{\infty} V_d$, both of which might have vectors in different dimensions. Given any parameter $\alpha \in \mathbb{R}^k$ we consider the dataset $S = \{(x, y) \mid x \in X_S\}$ where for each x the label is given by $y = \langle \varphi(x), \alpha \rangle + \varepsilon$, here the noise ε is i.i.d. with distribution $N(0, \sigma^2)$. We define the test set T in a completely analogous fashion.

Theorem 2.6.2 (**Lower bound**). *Assume that $\hat{\alpha}$ is any estimator having access to training data S and $\alpha \in \mathbb{R}^k$ is arbitrary. Then, the minimax excess risk on test set T is exactly*

$$\inf_{\widehat{\alpha}} \sup_{\alpha} \left\{ R_T^{(\alpha)}(\hat{\alpha}) - \min_{\alpha' \in \mathbb{R}^k} R_T^{(\alpha)}(\alpha') \right\} = \frac{\sigma^2}{|S|} \mathrm{Trace}\left(\Sigma_S^{-1} \Sigma_T\right). \tag{2.21}$$

2.7 Computing Free Bases

This section studies how to compute a free basis of invariant polynomials and provides computationally efficient bases for permutation and set-permutation invariant polynomials.

Setting the stage, let $\{(V_d, \gamma_d)\}_{d \in \mathbb{N}}$ be a consistent sequence of $\{G_d\}_{d \in \mathbb{N}}$−representations with $\mathbb{R}[V_d]_m$ having finite stability degree. Monomials yield a natural choice for a basis of the polynomial spaces $\mathbb{R}[V_d]_m$; the next result shows that if γ_d is the zero-padding map, and $G_d \subset \mathfrak{S}_d$ acts by permutation on the basis of V_d, then the set of polynomials obtained by summing all monomials in a G_d-orbit yields a free basis of invariants.

Proposition 2.7.1 (Free basis via symmetrization). *Consider $\{(V_d, \gamma_d)\}_{d \in \mathbb{N}}$, a consistent sequence of $\{G_d\}_{d \in \mathbb{N}}$−representations, where γ_d is the zero-padding map, $G_d \subset \mathfrak{S}_d$ acts by permuting coordinates of V_d, and suppose $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}_{d \in \mathbb{N}}$ has stability degree D. Then, if $f_1^{(d)}, \ldots, f_{N_d}^{(d)}$ is the monomial basis of $\mathbb{R}[V_d]_m$, the set of monic polynomials*

$$B_d = \left\{ \frac{1}{c_i} \sum_{g \in G_d} g \cdot f_i^{(d)} : i = 1, \ldots, N_d \right\}_{d \in \mathbb{N}}$$

forms a free basis of invariants for $\mathcal{F}_m$, where each c_i is a normalizing constant. [3]

[3] The normalizing constants c_i can be explicitly computed by taking the $\mathfrak{S}_d$-sum of a monomial and dividing so that the result is monic.

Proof. By Lemma 2.2.8, we have that B_d forms a basis of $\mathbb{R}[V_d]_m^{\mathfrak{S}_d}$ for each dimension $d \geq D$ and so it remains to check that each G_d−summed monomial is a free element of $\{\mathbb{R}[V_d]_m\}_{d\in\mathbb{N}}$. Picking a monomial $f_i^{(d+1)}$, each monomial $g\cdot f_i^{(d+1)}$ appears c_i times in the sum $\sum_{g\in G_{d+1}} g\cdot f_i^{(d+1)}$. By dividing by the normalizing constant c_i, we ensure that the coefficient on each monomial in the orbit is one. Since γ_d is the zero-padding map, the projection γ_d^* sets the value of variables that are in $\mathbb{R}[V_{d+1}]_m$ but not $\mathbb{R}[V_d]_m$ equal to zero. It follows that $\gamma_d^*\left(\frac{1}{c_i}\sum_{g\in G_{d+1}} g\cdot f_i^{(d+1)}\right)$ is the sum of all monomials in the orbit of $f_i^{(d+1)}$ that are also contained in $\mathbb{R}[V_d]_m$. If two monomials $f_j^{(d+1)}, f_{j'}^{(d+1)}$ are in the orbit of $f_i^{(d+1)}$ and are contained in $\mathbb{R}[V_d]_m$, there exists some permutation $\sigma \in G_d$ such that $f_j^{(d+1)} = \sigma f_{j'}^{(d+1)}$. Setting $f_j^{(d)} := \gamma_d^*(f_j^{(d+1)})$, it follows that

$$\gamma_d^*\left(\frac{1}{c_i}\sum_{g\in G_{d+1}} g\cdot f_i^{(d+1)}\right) = \frac{1}{c_j}\sum_{g\in G_d} g\cdot f_j^{(d)}$$

holds for some j and thus $\frac{1}{c_i}\sum_{g\in G_d} g\cdot f_i^{(d+1)}$ is a free element as desired. □

In particular, taking V_d and $G_d = \mathfrak{S}_d$ to be the vector spaces and groups appearing in Examples 2.1.1, 2.1.2, 2.1.3, and 2.1.4, we can apply Lemma 2.7.1 to conclude that the $\mathfrak{S}_d$-averaged monomials form a free basis of invariants for $\mathbb{R}[V_d]_m^{\mathfrak{S}_d}$. However, as the dimension d increases, calculating the $\mathfrak{S}_d$-orbit of monomials becomes prohibitively costly. Thus, it is desirable to design free bases of invariants that remain tractable as d grows. While we are not aware of such constructions for graphs (Example 2.1.3) and point clouds (Example 2.1.4), we provide efficiently-computable free bases for permutation invariant (Example 2.1.1) and set permutation invariant kernels (Example 2.1.2) in the following two sections, respectively.

2.7.1 Permutation Invariants

The foundation for the free bases of both permutation and set permutation invariant polynomials are *elementary symmetric polynomials*.

Definition 2.7.2 (Elementary Symmetric Polynomials)**.** The *elementary symmetric polynomial* of degree k in d variables $x = (x_1, \ldots, x_d)$ is $e_k(x) = \sum_{1\leq i_1<\ldots<i_k\leq d}\prod_{j=1}^k x_{i_j}$. For a given partition $\lambda \vdash m$, *the elementary symmetric polynomial indexed by* λ is $e_\lambda(x) = e_{\lambda_1}(x)\ldots e_{\lambda_d}(x)$ where $e_{\lambda_i}(x)$ is the λ_i^{th} elementary symmetric polynomial.

Elementary symmetric polynomials are well studied combinatorial objects. In particular, they form an algebra basis of the ring of symmetric polynomials, and the elementary

symmetric polynomials indexed by partitions form a vector space basis of the space of symmetric polynomials [32, 96, 99, 105]. The key to the efficient computability of elementary symmetric polynomials is the following recursion:

$$e_j(x_1, \dots, x_d) = x_d e_{j-1}(x_1, \dots, x_{d-1}) + e_j(x_1, \dots, x_{d-1}) \tag{2.22}$$

We leverage this recursion to show that the elementary symmetric polynomials indexed by partitions form a free basis of invariants.

Proposition 2.7.3 (Free basis of permutation invariant polynomials)**.** *Let $V_d = \mathbb{R}^d$ and consider the consistent sequence of $\{\mathfrak{S}_d\}_{d\in\mathbb{N}}$-representations $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}_{d\in\mathbb{N}}$ where γ_d is the zero-padding map. The set $\{e_\lambda(x_1, \dots, x_d) : \lambda \vdash j \leq m\}$ of all elementary symmetric polynomials indexed by partitions forms a free basis of invariants for $\mathcal{F}_m$.*

Proof. Elementary symmetric polynomials indexed by partitions of $j \leq m$ form a vector space basis of $\mathbb{R}[V_d]_m^{\mathfrak{S}_d}$ and so we need only prove that they are free elements. Since γ_d is the zero-padding map, for each $f(x_1, \dots, x_d) \in \mathbb{R}[V_d]_m$ we have that $\gamma_d^* f(x_1, \dots, x_d) = f(x_1, \dots, x_{d-1}, 0)$. So, by (2.22), for each elementary symmetric polynomial $e_j(x_1, \dots, x_d)$,

$$\begin{aligned}\gamma_d^*(e_j(x_1, \dots, x_d)) &= e_j(x_1, \dots, x_{d-1}, 0) \\ &= e_j(x_1, \dots, x_{d-1}) + 0 \cdot e_{j-1}(x_1, \dots, x_{d-1}) = e_j(x_1, \dots, x_{d-1}).\end{aligned}$$

It follows that the elementary symmetric polynomials are free elements of $\mathcal{F}_m$. Moreover, for a given partition λ, we can see that

$$\begin{aligned}\gamma_d^*(e_\lambda(x_1, \dots, x_d)) &= \gamma_d^*(e_{\lambda_1}(x_1, \dots, x_d) \cdots e_{\lambda_d}(x_1, \dots x_d)) \\ &= e_{\lambda_1}(x_1, \dots, x_{d-1}, 0) \cdots e_{\lambda_d}(x_1, \dots x_{d-1}, 0) \\ &= \prod_{i=1}^{d} (e_{\lambda_i}(x_1, \dots, x_{d-1}) + 0 \cdot e_{\lambda_i - 1}(x_1, \dots, x_{d-1})) \\ &= e_\lambda(x_1, \dots, x_{d-1})\end{aligned}$$

and so the elementary symmetric polynomials indexed by partitions of $j \leq m$ are also free elements of $\mathcal{F}_m$. □

For any given dimension d, and $r = \dim(\mathbb{R}[V_d]_m^{\mathfrak{S}_d})$, let $E_m^{(d)} \colon \mathbb{R}^d \to \mathbb{R}^r$ be a mapping producing a vector of all evaluations of elementary symmetric polynomials indexed by partitions of degree at most m. As before we will drop the index d and simply write E_m. Suppose

that one has access to the matrix C representing a kernel K with respect to the basis of elementary symmetric polynomials indexed by partitions. Then, for data points $x, y \in \mathbb{R}^d$, Procedure 1 describes the evaluation of $K(x, y)$.

Procedure 1 Permutation Invariant Kernel Computation

Input: Data points $x, y \in \mathbb{R}^d$, matrix C representing K.

1: Set $E_m(x) = E_m(y) = 0 \in \mathbb{R}^r$
2: **for all** j such that $0 \leq j \leq m$ **do**
3: Store $e_j(x)$
4: Store $e_j(y)$
5: **end for**
6: **for** $j = 0$ to m **do**
7: **for all** $\lambda \vdash j$ **do**
8: Set $E_m(x)_\lambda = \prod_{i=1}^d e_{\lambda_i}(x)$
9: Set $E_m(y)_\lambda = \prod_{i=1}^d e_{\lambda_i}(y)$
10: **end for**
11: **end for**

Output: $E_m(x)^\top C E_m(y)$

Proposition 2.7.4 (Complexity of evaluating a permutation invariant kernel)**.** *Suppose that* $\{K_d \colon V_d \times V_d \to \mathbb{R}\}$ *is a free kernel with rank equal to* $r = \dim(\mathbb{R}[V_d]_m^{\mathfrak{S}_d})$*. Suppose one has access to the matrix* $C \in \mathbb{R}^{r \times r}$ *representing* K_d *in the basis of elementary symmetric polynomials indexed by partitions, i.e.,* $K_d(x, y) = E_m(x)^\top C E_m(y)$*. Then, the number of floating point operations required to compute* $K_d(x, y)$ *following Procedure 1 is*

$$4md + \sum_{j=0}^{m} \sum_{\lambda \vdash j} (|\lambda| - 1) + \left(\sum_{j=0}^{m} p(j) + 1 \right) \left(2 \sum_{j=0}^{m} p(j) - 1 \right).$$

Proof. The number of flops necessary to compute $K_d(x, y)$ can be upper bounded as follows.

1. The number of flops required to complete the loop in Step 2 is given by the number of flops necessary to compute the elementary symmetric polynomials in the entries of x and y. Thanks to (2.22), computing the elementary symmetric polynomials requires filling in the entries of an $m \times d$ matrix, with each entry requiring one addition and one multiplication. So, in total, evaluating the elementary symmetric polynomials in x and y takes $4md$ flops.

2. For each partition λ, combining the elementary symmetric polynomials to form the elementary symmetric polynomial indexed by λ takes $|\lambda|-1$ multiplications, where $|\lambda|$ is the number of parts of λ. Thus, the number of flops required to complete the loop in Step 6 by generating all elementary symmetric polynomials indexed by partitions of all $j \leq m$ is $\sum_{j=0}^{m} \sum_{\lambda \vdash j} (|\lambda| - 1)$ operations.

3. Finally, evaluating the output $E_m(x)^\top C E_m(y)$ via naive matrix multiplication takes $(\sum_{j=0}^{m} p(j) + 1)(2\sum_{j=0}^{m} p(j) - 1)$ operations.

The sum of these operations amounts to the stated bound. □

One can further upper bound the number of flops by by $4md + (m^2+m)p(m) + 2m^2 p(m)^2$, and using the Hardy-Ramanujan asymptotic formula for $p(m)$, the number of flops grows at a rate of $4md + (m+1)\exp\left(\pi\sqrt{2m/3}\right) + \exp\left(2\pi\sqrt{2m/3}\right)$. Note that this completes the proof of the first statement in Theorem 2.1.7. While there is an unavoidable exponential dependence on the degree m, the dependence on the dimension of the data is linear and thus this method scales well as the degree is held constant and dimension increases.

Example 7 (Permutation Invariant Taylor Features)**.** We now turn our attention to a concrete invariant kernel that arises in practice and provide an efficient method for its computation. Consider the non-invariant kernel $K(x,y) = \sum_{j=0}^{m} \frac{\langle x,y\rangle^j}{j!}$, which is a rescaled Taylor features kernel. The permutation invariant Taylor features kernel is

$$K^*(x,y) = \sum_{\sigma \in \mathfrak{S}_d} \sum_{j=0}^{m} \frac{\langle x, \sigma(y)\rangle^j}{j!}.$$

We provide an efficient way to evaluate $K^*(x,y)$ by explicitly determining the matrix C representing the symmetric bilinear form corresponding to K^* with respect to the basis of elementary symmetric polynomials indexed by partitions. First, we write the kernel as a linear combination of *monomial symmetric polynomials*, a well-known vector space basis of the space of symmetric polynomials. Given a partition $\lambda \vdash j$, the monomial symmetric polynomial indexed by λ is $m_\lambda(x) = \frac{1}{\eta(\lambda)} \sum_{\sigma \in \mathfrak{S}_d} x^{\sigma(\lambda)}$ where $\eta(\lambda)$ is the product of the factorials of the multiplicities of the parts of λ.

For a given degree m we successively compute

$$\begin{aligned}\sum_{\sigma\in\mathfrak{S}_d}\langle x,\sigma(y)\rangle^m &= \sum_{\sigma\in\mathfrak{S}_d}(x_1y_{\sigma(1)}+\ldots+x_dy_{\sigma(d)})^m\\ &= \sum_{\sigma\in\mathfrak{S}_d}\sum_{\lambda\in C_d(m)}\binom{m}{\lambda}\prod_{i=1}^d x_i^{\lambda_i}y_{\sigma(i)}^{\lambda_i}\\ &= \sum_{\lambda\in C_d(m)}\binom{m}{\lambda}\left(\prod_{i=1}^d x_i^{\lambda_i}\right)\left(\sum_{\sigma\in\mathfrak{S}_d}\prod_{i=1}^d y_{\sigma(i)}^{\lambda_i}\right)\\ &= \sum_{\lambda\in C_d(m)}\binom{m}{\lambda}\left(\prod_{i=1}^d x_i^{\lambda_i}\right)\eta(\lambda)m_\lambda(y)\end{aligned}$$

where the constant $\eta(\lambda)$ appears since it counts how many times the monomial $\prod_{i=1}^d y_i^{\lambda_i}$ appears in the sum $\sum_{\sigma\in\mathfrak{S}_d}\prod_{i=1}^d y_{\sigma(i)}^{\lambda_i}$.

Since each composition λ is a permutation of some partition of m, and each monomial $\prod_{i=1}^d x_i^{\lambda_i}$ appears exactly once in the sum, the equality

$$\sum_{\lambda\in C_d(m)}\binom{m}{\lambda}\left(\prod_{i=1}^d x_i^{\lambda_i}\right)\eta(\lambda)m_\lambda(y)=\sum_{\lambda\vdash m}\binom{m}{\lambda}\eta(\lambda)m_\lambda(x)m_\lambda(y)\quad\text{holds.}$$

Note that the coefficients $\binom{m}{\lambda}$ and $\eta(\lambda)$ are independent of the dimension d.

Using the above computations, $K^*(x,y)$ can be written as

$$K^*(x,y)=M_m(x)^\top CM_m(y)$$

where $M_m(x)$ is the vector of all monomial symmetric polynomials in x and the matrix C is diagonal with entries $\frac{1}{j!}\cdot\binom{j}{\lambda}\eta(\lambda)$ for each partition $\lambda\vdash j\leq m$.

While the monomial symmetric polynomials do form a convenient vector space basis for the ring of symmetric polynomials, they are not efficiently computable in high dimensions. In contrast to monomial symmetric polynomials, we have seen that the elementary symmetric polynomials can be efficiently computed using the recursive relationship (2.22). Thus, in order to efficiently compute the kernel $K^*(x,y)$ we now need only compute the change of basis matrix from monomial symmetric polynomials to elementary symmetric polynomials indexed by partitions.

It is known that each elementary symmetric polynomial $e_\lambda(x)$ of degree i can be written as a linear combination of monomial symmetric polynomials: $e_\lambda(x)=\sum_{\mu\vdash i}M_{\lambda\mu}m_\mu(x)$

where $M_{\lambda\mu}$ is the number of binary matrices with row sum equal to λ and column sum equal to μ, see [99]. The quantities $M_{\lambda\mu}$ can be calculated recursively due to [72]. Let B be the change of basis matrix from monomial symmetric polynomials to elementary symmetric polynomials indexed by partitions and $E_m(x)$ the vector of all elementary symmetric polynomials indexed by partitions of integers up to degree m. Then,

$$K^*(x,y) = E_m(x)^\top B^{-\top} C B^{-1} E_m(y).$$

For a given kernel $K(x,y)$, the matrix $B^{-\top}CB^{-1}$ is independent of both the dimension d and the individual data points and thus can be precomputed. Then, to evaluate $K^*(x,y)$ for a given pair of points x, y we need only evaluate the elementary symmetric polynomials, which can be done efficiently using the above recursion, and form the elementary symmetric polynomials indexed by partitions by multiplying elementary symmetric polynomials together. When the degree of the kernel m is small, the latter is fast.

2.7.2 Set-Permutation Invariants

We now turn our attention to set-permutation invariant kernels and build upon the theory in Section 2.7.1 to provide an efficiently computable free basis of set-permutation invariant polynomials. First, we overview the necessary background on *multisymmetric polynomials*, following the discussion in [32]. Then, we show that *$\alpha-$polarized elementary symmetric polynomials indexed by partitions* form a free basis of set-permutation invariant polynomials. Finally, we determine the complexity of evaluating a set-permutation invariant kernel using this basis and show that it is linear in the number of points d.

Definition 2.7.5 (Multisymmetric Polynomial)**.** Let $x_1, \dots, x_d$ be $k-$dimensional vectors of indeterminates. A polynomial $f(x_1, \dots, x_d)$ in the entries of $x_1, \dots, x_d$ is *multisymmetric* if it is invariant under the action of $\mathfrak{S}_d$ given by $\sigma \cdot f(x_1, \dots, x_d) = f\left(x_{\sigma(1)}, \dots, x_{\sigma(d)}\right)$.

Multisymmetric polynomials are the natural generalization of symmetric polynomials needed to study set-permutation invariant kernels and are classical objects in combinatorics and invariant theory, see [32, 96, 105] for an overview. We define *multi-degree*, the natural generalization of degree to polynomials in the entries of indeterminate vectors.

Definition 2.7.6 (Multi-Degree)**.** Let $f(x_1, \dots, x_d) = \prod_{i=1}^{d} \prod_{j=1}^{k} x_{ij}^{\alpha_{ij}}$ be a monomial in the $k-$dimensional indeterminate vectors $x_1, \dots, x_d$. The *multi-degree* of f is the $k-$dimensional vector $\alpha = \sum_{i=1}^{d} (\alpha_{i1}, \dots, \alpha_{ik})$. The *total degree* of f is $|\alpha| = \sum_{i=1}^{d} \sum_{j=1}^{k} \alpha_{ij}$. We denote by f_α the component of f with multi-degree α.

We now introduce the *polarization operator* Pol, which maps symmetric polynomials in d variables to multisymmetric polynomials in the $k-$dimensional vectors $x_1, \dots, x_d$.

Definition 2.7.7 (Polarization)**.** Let $f(t_1, \dots, t_d)$ be a polynomial in d variables. The *polarization* of f is $\mathrm{Pol}(f)(x_1, \dots, x_d) = f(x_{11} + \dots + x_{1k}, \dots, x_{d1} + \dots + x_{dk})$. For multi-degree α, the $\alpha-$polarization of f is $\mathrm{Pol}_\alpha(f)(x_1, \dots, x_d) = f_\alpha(x_{11} + \dots + x_{1k}, \dots, x_{d1} + \dots + x_{dk})$, the component of $\mathrm{Pol}(f)$ with multi-degree α.

Example 8. Let $d = 3$ and $k = 2$. First, consider the second elementary symmetric polynomial in 3 variables, $e_2(x_1, x_2, x_3) = x_1x_2 + x_1x_3 + x_2x_3$. The polarization of e_2 is

$$\mathrm{Pol}(e_2) = (x_{11} + x_{12})(x_{21} + x_{22}) + (x_{11} + x_{12})(x_{31} + x_{32}) + (x_{21} + x_{22})(x_{31} + x_{32}).$$

There three different multi-degrees with total degree 2 are $(2,0), (1,1), (0,2)$ and each of these multi-degrees corresponds to a different multi-homogeneous component of $\mathrm{Pol}(e_2)$:

$$\begin{aligned}
\mathrm{Pol}_{(2,0)}(e_2)(x_1, x_2, x_3) &= x_{11}x_{21} + x_{11}x_{31} + x_{21}x_{31} \\
\mathrm{Pol}_{(1,1)}(e_2)(x_1, x_2, x_3) &= x_{11}x_{22} + x_{11}x_{32} + x_{12}x_{21} + x_{12}x_{31} + x_{21}x_{32} + x_{22}x_{31} \\
\mathrm{Pol}_{(0,2)}(e_2)(x_1, x_2, x_3) &= x_{12}x_{22} + x_{12}x_{32} + x_{22}x_{32}.
\end{aligned}$$

Polarization commutes with both the addition and multiplication of polynomials:

$$\begin{aligned}
\mathrm{Pol}(f + g) &= \mathrm{Pol}(f) + \mathrm{Pol}(g), \\
\mathrm{Pol}(fg) &= \mathrm{Pol}(f)\mathrm{Pol}(g).
\end{aligned}$$

Thus, given the relationship (2.22), there is a recursive relationship between the polarizations of elementary symmetric polynomials:

$$\mathrm{Pol}(e_m)(x_1, \dots, x_d) = \mathrm{Pol}(e_m)(x_1, \dots, x_{d-1}) + (x_{d1} + \dots, x_{dk})\mathrm{Pol}(e_{m-1})(x_1, \dots, x_{d-1}).$$

We will be interested in the $\alpha-$polarizations of elementary symmetric polynomials. A simple computation yields the equality

$$\mathrm{Pol}_\alpha(e_m)(x_1, \dots, x_d) = \mathrm{Pol}_\alpha(e_m)(x_1, \dots, x_{d-1}) + \left(\sum_{j=1}^{k} x_{dj}\right) \mathrm{Pol}_{\alpha-\delta_j}(e_{m-1})(x_1, \dots, x_{d-1}),$$

where $\delta_j = (0, \dots, 0, 1, 0, \dots, 0)$ is the j^{th} indicator vector, see [32, Section 2.3].

By [96, Theorem 3.4.1], the set of all α-polarized elementary symmetric polynomials, with α varying over all possible multi-degrees, form an algebra basis of the space $\mathbb{R}[x_1, \dots, x_d]^{\mathfrak{S}_d}$ of all invariant polynomials in the entries of $x_1, \dots, x_d$. It follows that taking products of $\alpha-$polarized elementary symmetric polynomials of degree at most m generates a vector space basis of $\mathbb{R}[x_1, \dots, x_d]_m^{\mathfrak{S}_d}$. In order to avoid double-counting, we introduce *vector partitions* to index the different products of $\alpha-$polarized elementary symmetric polynomials.

Definition 2.7.8 (Vector Partition)**.** Fix an integer $m \geq 0$. A $k-$vector partition of m into d parts is a collection of $k-$dimensional non-negative integer vectors $\Lambda = (\lambda_1, \dots, \lambda_d)$ such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ in the graded lexicographic order and $\sum_{i=1}^{d} \sum_{j=1}^{k} \lambda_{ij} = m$. [4] Each λ_i is a part of Λ and we use the shorthand $\Lambda \vdash_k m$ to denote a k-vector partition of m.

For a vector partition $\Lambda \vdash_k j$, the polarized elementary symmetric polynomial indexed by Λ is given by

$$E_\Lambda(x_1, \dots, x_d) = \prod_{i=1}^{d} \mathrm{Pol}_{\lambda_i}\left(e_{|\lambda_i|}\right)(x_1, \dots, x_d),$$

where $|\lambda_i| = \sum_{j=1}^{k} \lambda_{ij}$. We now show that the set of all polarized elementary symmetric polynomials indexed by vector partitions of total degree at most m form a free basis of invariants for the sequence $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}_{d \in \mathbb{N}}$.

Proposition 2.7.9 (Free Basis of Set-Permutation Invariant Polynomials)**.** *Let V_d and G_d be the vector spaces and groups from Example 2.1.2. Consider the consistent sequence $\mathcal{F}_m = \{(\mathbb{R}[V_d]_m, \gamma_d)\}_{d \in \mathbb{N}}$ of $\{G_d\}_{d \in \mathbb{N}}-$representations where γ_d is the zero-padding map. The set $\{E_\Lambda(x_1, \dots, x_d) : \Lambda \vdash_k j \leq m\}$ of all polarized elementary symmetric polynomials indexed by k-vector partitions of all $j \leq m$ forms a free basis of invariants for $\mathcal{F}_m$.*

Proof. We first show that the set of all polarized elementary symmetric polynomials indexed by vector partitions forms a basis by showing that the number of $k-$vector partitions of integers at most m is exactly equal to the number of distinct products of $\alpha-$polarized elementary symmetric polynomials with degree at most m and is equal to $\dim \mathbb{R}[V_d]_m^{G_d}$ as given in Example 2.1.2.

First, recall that our proof of Example 2.1.2 amounted to choosing unique G_d-orbit representatives of non-negative integer $d \times k$ matrices whose entries totaled at most m.

[4] $\lambda_i \geq \lambda_{i'}$ in the graded lexicographic order if either $\sum_{j=1}^{k} \lambda_{ij} > \sum_{j=1}^{k} \lambda_{i'j}$ or if $\sum_{j=1}^{k} \lambda_{ij} = \sum_{j=1}^{k} \lambda_{i'j}$ and $\lambda_i \geq \lambda_{i'}$ in the lexicographic order.

We previously argued that each orbit contains at least one matrix whose row sums form a partition of some integer $j \leq m$. Note that there is exactly one matrix in each orbit such that the rows are ordered in the graded lexicographic order. This matrix is in direct correspondence with a $k-$vector partition and thus it follows that the number of $k-$vector partitions of integers at most m is equal to $\dim \mathbb{R}[V_d]_m^{G_d}$.

Since there is one elementary symmetric polynomial for each degree $j \leq m$, the number of $\alpha-$polarized elementary symmetric polynomials of degree j is equal to the number of distinct multi-degree vectors α with total degree $|\alpha| = j$. Note that each multi-degree vector is in direct correspondence with a weak composition of j and so there are exactly $\binom{j+k-1}{k-1}$ distinct $\alpha-$polarized elementary symmetric polynomials of degree j. To generate a vector space basis of $\mathbb{R}[x_1, \dots, x_d]_m^{\mathfrak{S}_d}$, we take products of $\alpha-$polarized elementary symmetric polynomials such that the total degree is at most m. We count the number of distinct products as follows: for each partition $\alpha = \{1^{\mu_1}, \dots, j^{\mu_j}\}$ of $j \leq m$, choose $\mu_i(\alpha)$ polarized elementary symmetric polynomial of degree i for all $i = 1, \dots, j$. The product of the chosen polarized elementary symmetric polynomials is an $\mathfrak{S}_d-$invariant polynomial of degree j. For each i, there are $\binom{c_k(i)+\mu_i(\alpha)-1}{\mu_i(\alpha)}$ possible choices for the polarized elementary symmetric polynomials of degree i. Taking the sum over all possible degrees, there are $\sum_{j=0}^{m} \sum_{\alpha \vdash j} \prod_{i=1}^{j} \binom{c_k(i)+\mu_i(\alpha)-1}{\mu_i(\alpha)}$ distinct products of polarized elementary symmetric polynomials of degree at most m. This is exactly the dimension of $\mathbb{R}[x_1, \dots, x_d]_m^{\mathfrak{S}_d}$ as given in Theorem 2.1.2 and so the set of all products of α-polarized elementary symmetric polynomials with total degree at most m forms a basis of the space $\mathbb{R}[x_1, \dots, x_d]_m^{\mathfrak{S}_d}$.

Since γ_d is the zero-padding map by assumption, for each polynomial $f(x_1, \dots, x_d) \in \mathbb{R}[V_d]_m$ we see that its projection onto $\mathbb{R}[V_{d-1}]_m$ is $\gamma_d^* f(x_1, \dots, x_d) = f(x_1, \dots, x_{d-1}, 0)$. Then, for each polarized elementary symmetric polynomial $\mathrm{Pol}_\alpha(e_j)(x_1, \dots, x_d)$ we compute the following sequence of equalities:

$$\begin{aligned}
\gamma_d^* \mathrm{Pol}_\alpha(e_j)(x_1, \dots, x_d) &= \mathrm{Pol}_\alpha(e_j)(x_1, \dots, x_{d-1}, 0) \\
&= \mathrm{Pol}_\alpha(e_j)(x_1, \dots, x_{d-1}) + \sum_{i=1}^{k} 0 \cdot \mathrm{Pol}_{\alpha - \delta_i}(e_{j-1})(x_1, \dots, x_{d-1}) \\
&= \mathrm{Pol}_\alpha(e_j)(x_1, \dots, x_{d-1})
\end{aligned}$$

It follows that the polarized elementary symmetric polynomials are free elements. Taking the relevant products, we have that polarized elementary symmetric polynomial indexed by vector partitions form a free basis of invariants. $\square$

For any given dimensions d, k, and $r = \dim(\mathbb{R}[V_d]_m^{\mathfrak{S}_d})$, let $P_m^{(d)} \colon \mathbb{R}^{d\times k} \to \mathbb{R}^r$ be a mapping producing a vector of all evaluations of $\alpha-$polarized elementary symmetric polynomials indexed by vector partitions of degree at most m. As before we will drop the index d and simply write P_m. Suppose that one has access to the matrix C representing a kernel K with respect to the basis of polarized elementary symmetric polynomials indexed by vector partitions. Then, for data points $x, y \in \mathbb{R}^d$, Procedure 2 describes the evaluation of $K(x, y)$.

Procedure 2 Set-Permutation Invariant Kernel Evaluation

Input: Data points $x, y \in \mathbb{R}^{d\times k}$, matrix C representing K.

Set $P_m(x) = P_m(y) = 0 \in \mathbb{R}^r$
for $j = 0$ to m **do**
  Compute $\mathrm{Pol}(e_j)(x)$
  Compute $\mathrm{Pol}(e_j)(y)$
  for all $\alpha \in C_k(j)$ **do**
    Store $\mathrm{Pol}_\alpha(e_j)(x)$
    Store $\mathrm{Pol}_\alpha(e_j)(y)$
  end for
end for
for $j = 0$ to m **do**
  for all $\Lambda \vdash_k j$ **do**
    Set $P_m(x)_\Lambda = \prod_{i=1}^d \mathrm{Pol}_{\lambda_i}(e_{|\lambda_i|})(x)$
    Set $P_m(y)_\Lambda = \prod_{i=1}^d \mathrm{Pol}_{\lambda_i}(e_{|\lambda_i|})(y)$
  end for
end for

Output: $P_m(x)^\top C P_m(y)$.

We use Procedure 2 to evaluate set-permutation kernels in the numerical experiments; see Section 2.8.1 for details.

Proposition 2.7.10 (Complexity of evaluating a set-permutation invariant kernel)**.** *Suppose that $\{K_d : \mathbb{R}^{d\times k} \times \mathbb{R}^{d\times k} \to \mathbb{R}\}$ is a free set-permutation invariant kernel. Suppose one has access to the matrix C representing K_d with respect to the basis of polarized elementary symmetric polynomials. Then, the number of floating point operations required to compute $K_d(x, y)$ via Procedure 2 is upper bounded by*

$$(m+1)\cdot \exp(\sqrt{2m/3})\cdot (k-1)^{mk} + \exp(2\sqrt{2m/3})\cdot (k-1)^{2mk} + \frac{d(2k-1)(k-1)^m}{m!}.$$

In particular, this quantity scales linearly in the dimension d.

Proof. We upper bound the number of flops necessary to compute $K_d(x, y)$ as follows.

1. Completing the loop in Step 2 requires evaluating all $\alpha-$polarized elementary symmetric polynomials. Given the recursive relationship between $\alpha-$polarized elementary symmetric polynomials, evaluating all of them entails filling the entries of a $d \times \binom{m+k-1}{k-1}$ matrix since there are $\binom{m+k-1}{k-1}$ multi-degrees with total degree m. Determining each entry requires k additions and $k-1$ products and so in total evaluating all $\alpha-$polarized elementary symmetric polynomials takes $d \cdot \binom{m+k-1}{k-1} \cdot (2k-1)$ flops. Note that $\binom{m+k-1}{k-1} = \binom{m+k-1}{m} \leq \frac{(k-1)^m}{m!}$.

2. Completing the loop in Step 10 requires computing all distinct products of $\alpha-$polarized elementary symmetric polynomials of degree at most m. Recall that the quantity $\dim \mathbb{R}[V_d]_m^{G_d}$ grows at most $\exp(\sqrt{2m/3}) \cdot (k-1)^{mk}$ and so there are at most $\exp(\sqrt{2m/3}) \cdot (k-1)^{mk}$ polarized elementary symmetric polynomials indexed by partitions to evaluate. Evaluating each polarized elementary symmetric polynomial indexed by vector partition requires taking the product of m $\alpha-$polarized elementary symmetric polynomials. It follows that the number of flops required to complete the loop in Step 10 is $m \exp(\sqrt{2m/3}) \cdot (k-1)^{mk}$.

3. Evaluating the output $P_m(x)^\top C P_m(y)$ via naive matrix multiplication takes $(r+1)(2r-1)$ operations and so the number of operations required is at most $(\exp(\sqrt{2m/3}) \cdot (k-1)^{mk} + 1)(2(\exp(\sqrt{2m/3}) \cdot (k-1)^{mk}) - 1)$.

Taking the sum of the number of operations at each step yields the stated upper bound. Note that while there is an unavoidable exponential dependence on both k and m, but a linear dependence on the dimension d. □

The proof of Proposition 2.7.10 also completes the proof of Theorem 2.1.7. We now build on Illustration 7 and introduce a specific example of a set-permutation invariant kernel that can be evaluated efficiently in high dimensions.

Example 9 (Set-Permutation Invariant Taylor Features)**.** While there is an unavoidable exponential dependence on k in the complexity of evaluating the basis of polarized elementary symmetric polynomial indexed by vector partitions, there are set-permutation invariant kernels for which computation is much quicker. For instance, consider the kernel defined by

$$K(x, y) = \frac{1}{|\mathfrak{S}_d|} \sum_{\sigma \in \mathfrak{S}_d} \sum_{j=0}^{m} \frac{\langle x_1, \sigma(y_1) \rangle^j + ... + \langle x_k, \sigma(y_k) \rangle^j}{j!}$$

where x_i is the i^{th} *column* of the matrix $X \in \mathbb{R}^{d\times k}$. Note that K is the sum of k invariant Taylor features kernels as described in Example 7. Since the complexity of evaluating a permutation invariant kernel is linear in dimension d, and K is evaluated by evaluating k permutation invariant kernels, the complexity of evaluating K is linear in both d and k.

However, it is clear that the above kernel is not full rank; the rank of K is $k \cdot \sum_{j=1}^{m} p(j)$ which is in general less than the maximum rank of a set-permutation invariant kernel given in Example 2.1.2. With this in mind, consider instead the invariant Taylor features kernel for sets given by

$$\begin{aligned} K(x,y) &= \frac{1}{|\mathfrak{S}_d|}\sum_{j=0}^{m}\sum_{\sigma\in\mathfrak{S}_d}(\langle x_1,\sigma(y_1)\rangle + ... + \langle x_k,\sigma(y_k)\rangle)^j \\ &= \frac{1}{|\mathfrak{S}_d|}\sum_{j=0}^{m}\sum_{\lambda\in C_k(j)}\sum_{\sigma\in\mathfrak{S}_d}\prod_{i=1}^{k}\langle x_i,\sigma(y_i)\rangle^{\lambda_i}. \end{aligned}$$

We would like to be able to write the above kernel in terms of the invariant inner products $\sum_{\sigma\in\mathfrak{S}_d}\langle x_i,\sigma(y_i)\rangle^{\lambda_i}$, since we have seen in Illustration 7 that these inner products are easy to compute. Although the sum over $\sigma \in \mathfrak{S}_d$ does not commute with the product, we can approximate the invariant Taylor features kernel for sets with the kernel defined by

$$K(x,y) = \frac{1}{|\mathfrak{S}_d|}\sum_{j=0}^{m}\sum_{\lambda\in C_k(j)}\prod_{i=1}^{k}\sum_{\sigma\in\mathfrak{S}_d}\langle x_i,\sigma(y_i)\rangle^{\lambda_i}. \tag{2.23}$$

In Section 2.8.1, we use the kernel in (2.23) to classify a mixture of Gaussians and compare its performance to the non-invariant Taylor features kernel.

2.8 Numerical Experiments

In this section, we perform three numerical experiments that support the theory in this chapter. Our first experiment showcases the use of the invariant kernels defined in (2.23) to classify sets of points. For our second experiment, we use the invariant regression estimator (2.19) to learn sequences of symmetric polynomials that generalize well across dimensions. Our final experiment uses the formula in Theorem 2.1.5 to estimate the dimensions of different spaces of invariant polynomials where the true value is known. We wrote these experiments in Python with some key functions imported from SageMath. The code and data to reproduce these results can be found at

https://github.com/j4ck-k/invariant-kernels.

2.8.1 Classifying Sets of Points

The kernel K defined in (2.23) is an example of an invariant kernel under set permutations. We use this kernel to classify sets of points; the data points in each set are drawn from a fixed Gaussian distribution. We sample from a collection of seven mean-zero Gaussian distributions. Each covariance matrix is diagonal, with diagonal entries in $\{1, 2, 0.5\}$.

We run multiple experiments, for each of them approximately half of the sets have been sampled from the standard normal distribution, and the task is to correctly classify which data points are in this class. Since the distribution does not depend on the order in which the points were sampled, the classifier should be invariant under reshuffling of the data; therefore, an invariant kernel K is appropriate for this task.

Let $\{(x_i, y_i)\}_{i=1}^n$ be the training data set—x_i is the stacked set of Gaussian samples and $y_i \in \{-1, 1\}$ encodes whether or not the set was drawn from a standard Gaussian—and $M_K = [K(x_i, x_j)]_{i,j=1}^n$ the Gram matrix corresponding to this data. We construct a classifier by first computing α^* by solving

$$\alpha^\star \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \frac{1}{2}\|y_i - (M_K\alpha)_i\|^2 + \frac{\lambda}{2}\alpha^\top M_K \alpha, \quad \text{or, equivalently,} \quad \alpha^\star = (M_K + \lambda I_n)^{-1} y$$

where the regularizing parameter λ is chosen via 4−fold cross validation.[5] The classification function is then given by

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^n \alpha_i K(x, x_i)\right).$$

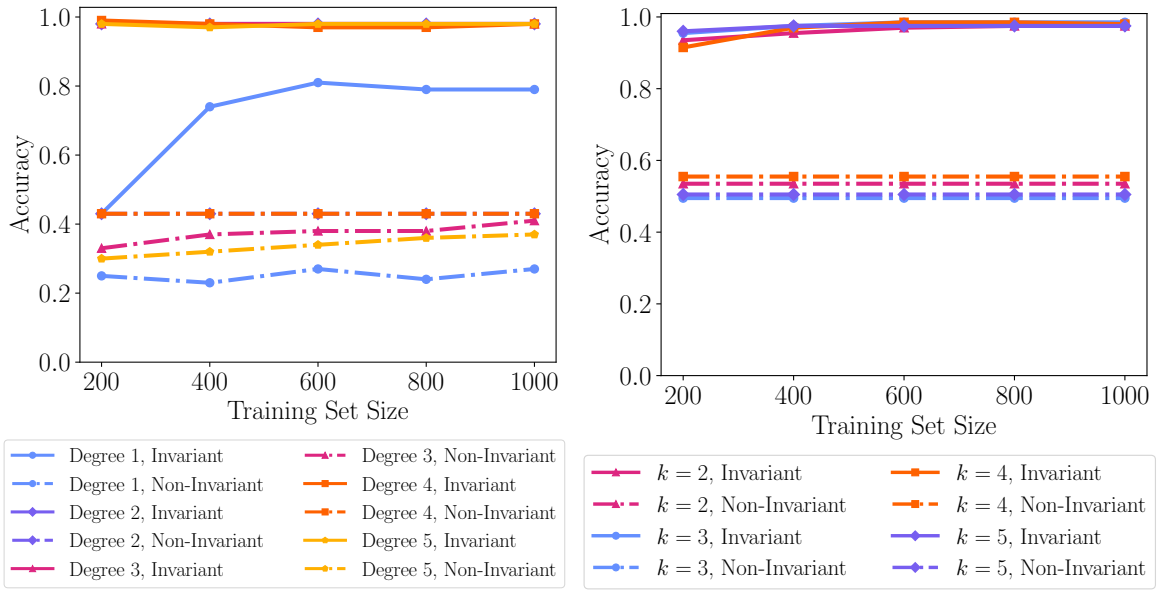
We run experiment using training sets of size $n \in \{200, 400, 600, 800, 1000\}$ Each classifier is then tested using an unseen set of 200 data points, and its success is measured based on the accuracy of the classifications on the test set. The accuracy of the classifier on the test set $\{(x_i, y_i)\}_{i=1}^n$ is calculated using the formula

$$\text{Accuracy}(f) = \frac{\text{Card}\{i : f(x_i) = y_i\}}{n}.$$

To construct the training and test sets, we sample the sets of points from a mixture of Gaussians with probability 0.5 of being sampled from the standard normal distribution and equal probability of being sampled from the remaining distributions.

[5] Note that we are using the ℓ_2 loss function for ease of computation.

We compare the performance of the invariant kernel K and non-invariant Taylor features approximation to the Gaussian kernel. For this comparison, we run two batches of experiments summarized in Figures 2.1a and 2.1b.



(a) Accuracy with Varying Degree m

(b) Accuracy with Varying Dimension k

Figure 2.1: Comparison of performance between (i) the set-permutation invariant kernel (2.23) and (ii) the non-invariant Taylor features kernel for classifying points drawn from a mixture of Gaussians (Illustration 7). We present two experimental scenarios: Figure 2.1a (left) shows results with fixed dimension $k = 2$ across varying degrees $m \in \{1, \dots, 5\}$, while Figure 2.1b (right) shows results with fixed degree $m = 2$ across varying dimensions $k \in \{1, \dots, m\}$. In each setting, the invariant kernel outperforms the non-invariant kernel.

Setting 1: Classifying points in $\mathbb{R}^2$ with Kernels of Varying Degree. We first fix the dimension $k = 2$ and test invariant and non-invariant kernels of degree $m \in \{1, 2, 3, 4, 5\}$. As shown in Figure 2.1a, invariant kernels of degrees higher than one far outperformed their non-invariant counterparts. Test accuracies of nonlinear invariant kernels were approximately 98%, whereas all non-invariant kernels had accuracy below 50%, which is worse than if the function simply randomly assigned classifications to each point cloud. Moreover, the linear invariant kernel also had much greater success than all of the non-invariant kernels. This demonstrates that for this task, even the simplest invariant kernel is more powerful than any non-invariant kernel for this task. We highlight that increasing the degree of both

the invariant and non-invariant kernels had little effect on the accuracy of the predictions, and thus, for this task, it suffices to consider only the relatively simple quadratic kernels.

Setting 2: Classifying point in $\mathbb{R}^k$ with Quadratic Kernels. We now fix the degree $m = 2$ of both the invariant and non-invariant kernel and test their performance on sets of points in dimension $k \in \{2, 3, 4, 5\}$. As shown in Figure 2.1b, the invariant kernel once again had almost perfect classification accuracy, whereas the non-invariant kernel had an accuracy of approximately 50%, which is equivalent to the accuracy of a random guess.

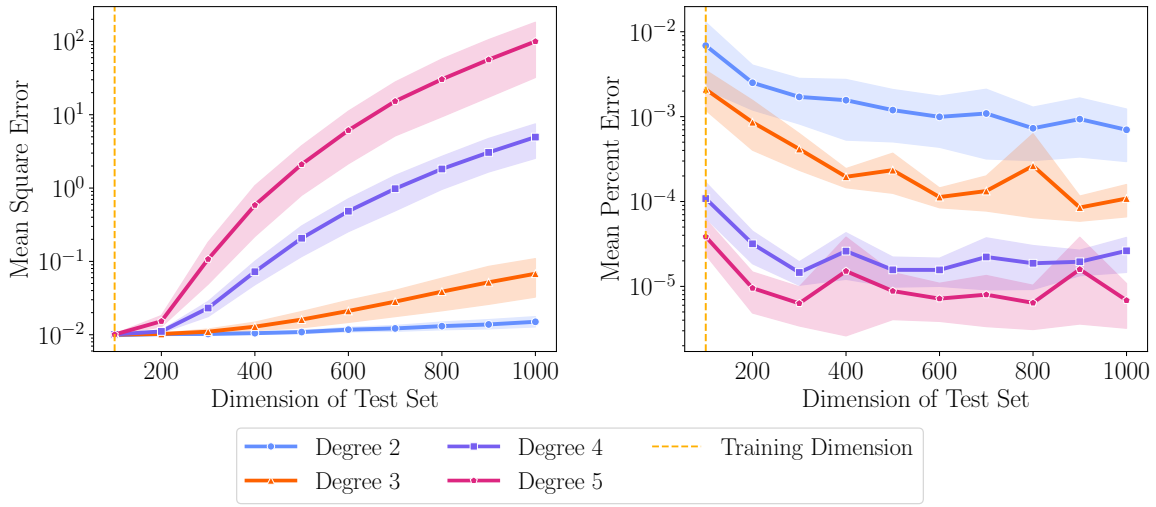


Figure 2.2: Mean Square Error (left) and Mean Percentage Error (right) versus test dimension using the estimator defined (2.6). For each degree, data was collected from ten randomly generated symmetric polynomials. Each estimator was trained on 10,000 noisy data points in dimension $d = 100$ and tested on 200 noisy data points in dimensions $d = 100, 200, \ldots, 1000$. The highlighted regions indicate a 95% confidence interval for the values.

2.8.2 Generalization Across Dimensions

We now leverage the theory in Section 2.6 to learn sequences of symmetric polynomials on data across different dimensions. We consider symmetric polynomials of degrees $m \in \{2, 3, 4, 5\}$. For each degree m, we learn ten randomly generated symmetric polynomials, and so in total we learn 40 different symmetric polynomials. Each polynomial is a random linear combination of elementary symmetric polynomials indexed by partitions with coefficients sampled from the standard Gaussian distribution. For each polynomial f, our training set consists of $10,000$ points $\{x_i\} \subset \mathbb{R}^{100}$ along with the noisy outputs $\{f(x_i) + \epsilon_i\}$ where x_i is sampled from the standard Gaussian distribution and ϵ_i is randomly sampled from a

Gaussian distribution with mean zero and variance $\sigma^2 = 10^{-2}$. We construct a predictor as in (2.19) and then test its performance as the dimension of the test set increased from $d = 100$ to $d = 1000$.

Figure 2.2 shows the mean squared error (MSE) and mean percentage error (MPE) of each learned predictor function as the dimension of the test sets increased. The MSE and MPE on the set $\{(x_i, y_i)\}_{i=1}^n$ for each predictor were calculated according to the formulas

$$\text{MSE}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \quad \text{and} \quad \text{MPE}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - f(x_i)|}{y_i}.$$

The mean MSE stays close to zero for each predictor of degree $m \in \{2, 3, 4\}$, but rapidly increased for degree $m = 5$ and dimensions greater than $d = 500$. However, the MPE of the predictors consistently decreased as both degree m and test set dimension d increased, with all predictors having less than a 1% error in dimension $d = 1000$.

2.8.3 Estimating $\dim(\mathbb{R}[V]_m^G)$ *Using Theorem 2.1.5*

Figure 2.3 displays Monte Carlo estimates of the dimension of $\mathbb{R}[V]_m^G$ for three different choices of vector space V and group G. The estimates were formed by calculating an empirical average of the formula in Theorem 2.1.5 with an increasing number of samples. Each experiment was repeated 10 times and Figure 2.3 plots the average estimate and the 95% confidence interval.

The first plot shows the estimate for the dimension of the space $\mathbb{R}[V]_4^G$ where $V = \mathbb{R}^d$ and $G = O(d)$, the group of orthogonal transformations. To construct the estimates of the dimension, we randomly sampled orthogonal matrices using the algorithm described in [69]. As any polynomial that is invariant under orthogonal transformations must be a polynomial in the square norm, the dimension d is arbitrary. Moreover, we can directly calculate that the true dimension is $\dim(\mathbb{R}[V]_4^G) = 3$. As shown in Figure 2.3, the estimated dimension quickly stabilizes to be approximately 3 as the number of samples increases.

The second plot shows the estimate for the dimension of $\mathbb{R}[V]_4^G$ where $V = \mathbb{R}^d$ for $d \geq 4$ and $G = \mathfrak{S}_d$ acts via coordinate permutation. Random elements of the group $\mathfrak{S}_d$ were sampled using the `random_element` function in Sage. Our estimate quickly approaches the true value computed in Example 2.1.1, i.e., $\dim(\mathbb{R}[V]_4^G) = \sum_{j=0}^{4} p(j) = 12$.

The final plot shows the estimate for the dimension of $\mathbb{R}[V]_2^G$ where $V \subset \mathbb{R}^{d\times d}$ is the space of all adjacency matrices of graphs on d vertices and $G = \mathfrak{S}_d$ acts via simultaneous permutation of rows and columns. As shown in Example 2.1.3, the rank of kernels on graphs

does not stabilize until the number of nodes d is at least twice the degree m of the kernel. Thus, for ease of computation, we use a smaller degree for this example. As in the previous example, random elements of the group $\mathfrak{S}_d$ were sampled using Sage's `random_element` function. The exact value of this dimension $\dim(\mathbb{R}[V]_2^G) = 10$ was calculated by summing the first three values in the sequence [31].

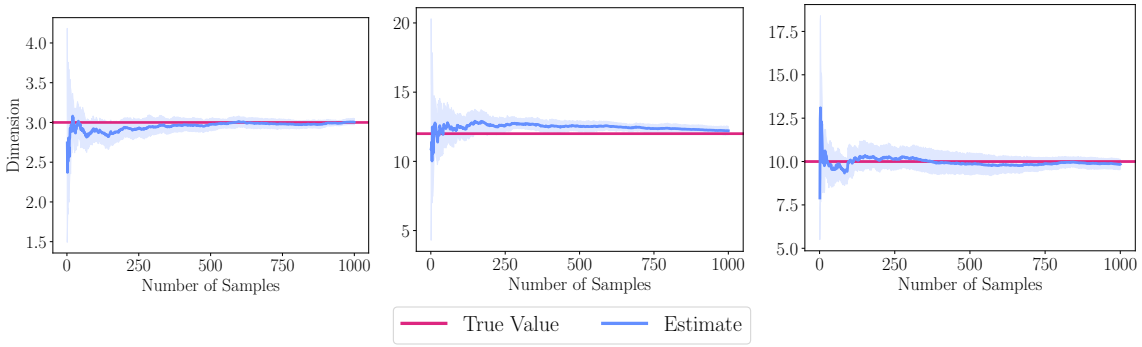


Figure 2.3: Monte Carlo estimates of $\dim(\mathbb{R}[V]_4^G)$ for orthogonal transformations, coordinate permutations and graph permutations. The estimates are calculated using the empirical average approximation to the formula in Theorem 2.1.5 with an increasing number of samples. In each case, we see that the estimate stabilises around the correct dimension relatively quickly. The highlighted regions are the 95% confidence intervals for the estimate.

2.9 Conclusion

In this chapter, we studied $G-$invariant polynomial kernels of fixed degree and proved that their ranks are independent of the input data dimension for important examples such as unordered sets, graphs, and point clouds. We provided elementary combinatorial arguments for these facts and also gave a general formula for computing invariant kernel ranks by leveraging representation theory. Moreover, we highlighted a direct connection between rank stabilization of invariant kernels and the phenomenon of representation stability. The theory of representation stability in turn allows us to learn sequences of polynomials that generalize well to data of arbitrary dimension. In particular, we provided a minimax optimal regression procedure for cross-dimensional generalization. Finally, we provided efficiently computable bases for the spaces of permutation invariant and set-permutation invariant polynomials, enabling high dimensional computation. Our developed theory was validated by the results of numerical experiments.

Chapter 3

DENSITY ESTIMATION ON RECTIFIABLE SETS

This chapter is the content of the paper [45]. Kernel density estimation is a popular method for estimating unseen probability distributions. However, the convergence of these classical estimators to the true density slows down in high dimensions. Moreover, they do not define meaningful probability distributions when the intrinsic dimension of data is much smaller than its ambient dimension. We build on previous work on density estimation on manifolds to show that a modified kernel density estimator converges to the true density on $d-$rectifiable sets. As a special case, we consider algebraic varieties and semi-algebraic sets and prove a convergence rate in this setting. We conclude the chapter with a numerical experiment illustrating the convergence of this estimator on sparse data.

3.1 Introduction

Density estimation is a canonical problem in statistics: given n random vectors $X_1, \dots, X_n \in \mathbb{R}^D$ sampled independently from a probability distribution P, we would like to estimate the density of P at some unseen point $x \in \mathbb{R}^D$. One popular method for estimating unseen density functions is through the use of a *kernel density estimator*:

$$\mathrm{KDE}_n(x) = \frac{1}{h_n^D} \sum_{i=1}^{n} K\left(\frac{\|X_i - x\|}{h_n}\right) \tag{3.1}$$

where K is some kernel function and h_n is a user-defined bandwidth parameter. The kernel density estimator defines a new probability distribution on $\mathbb{R}^D$ that we hope converges to the true distribution P as the number of samples n increases. Indeed, foundational work by Parzen [79] (when $D = 1$) and Cacoullos [15] (when $D > 1$) showed that when the bandwidth parameter h_n is chosen appropriately so that $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^d = \infty$, then the estimator KDE_n converges to the true density p. In particular, when h_n is chosen in proportion to $n^{-1/(D+4)}$, the mean square error (MSE) of $\hat{p}$ is of the order

$$\mathrm{MSE}\left[\mathrm{KDE}_n(x)\right] = O\left(\frac{1}{n^{4/(D+4)}}\right) \tag{3.2}$$

as $n \to \infty$. So, while the estimator $\text{KDE}_n(x)$ does well-approximate the true density $p(x)$ as the number of samples increases, this convergence slows down as the ambient dimension of the data D increases.

In many modern applications, although data is sampled in some high dimensional space $\mathbb{R}^D$, we see that the *intrinsic* dimension of the data is much smaller. In these cases, the support of the probabiliy measure P is some domain $\Omega \subset \mathbb{R}^D$ with dimension d that is much smaller than the ambient dimension D. Now, the kernel density estimator defined in (3.1) no longer defines a meaningful probability distribution on Ω : in order to account for the low-dimensionality of Ω, we must normalize the estimator to be zero outisde of Ω and infinite on Ω. One proposed workaround is to define the estimator

$$\hat{p}_n(x) = \frac{1}{h_n^d} \sum_{i=1}^{n} K\left(\frac{\|X_i - x\|}{h_n}\right), \tag{3.3}$$

a modified version of the estimator in (3.1) that uses the intrinsic dimension d of the data, rather than its ambient dimension D. We expect this new estimator to approximate the density of each point $x \in \Omega$ as the number of samples n grows.

The setting where Ω is a $d-$dimensional Riemannian submanifold of $\mathbb{R}^D$ was studied in [80] (when Ω is known), [77] (when only $\dim \Omega = d$ is known) and [10] (where $\dim \Omega$ is possibly unknown). It was shown that when the bandwidth parameters h_n are chosen proportionally to $n^{-1/(d+4)}$, then the mean square error $\text{MSE}\left[\hat{p}_n(x)\right]$ is of the order $O\left(\frac{1}{n^{4/(d+4)}}\right)$ as $n \to \infty$, mirroring the convergence rate of KDE_n.

However, the assumption that data lies on a manifold is very strong and is unreasonable in many important examples. For example, a natural way to ensure vector data has a low intrinsic dimension is to assume some level of sparsity. In a similar vein, if data is matrix valued, we may assume that the data is low rank or positive semi-definite. The sets of sparse vectors, low rank matrices, and positive semi-definite matrices do not lie on manifolds yet do have intrinsic senses of dimension that should be exploited when estimating densities on these spaces. In this chapter, we study the estimator $\hat{p}_n(x)$ on $d-$*rectifiable sets*, which can be thought of as the natural measure-theoretic generalizations of $d-$dimensional submanifolds and include the aforementioned examples as special cases. An important property of rectifiable sets is that they have *approximate tangent spaces* at almost every point. When these approximate tangent spaces satisfy a certain condition, discussed in Section 3.3, we prove the following theorem, whcih is the main result of this chapter.

Theorem 3.1.1: (KDE on rectifiable sets)

Let $\Omega \subset \mathbb{R}^D$ be a $d-$rectifiable set and P a probability measure on Ω with density p with respect to the $d-$dimensional Hausdorff measure $\mathcal{H}^d$. Assume that the kernel K is continuous and vanishes outside of $[0,1]$ Then, when bandwidth parameters h_n are chosen appropriately, the equality

$$\mathrm{MSE}\left[\hat{p}_n(x)\right] = O\left(\frac{1}{n^{2m/(d+2m)}}\right). \tag{3.4}$$

holds with probability 1, where m is a measure of how well Ω is approximated by its approximate tangent spaces.

Note that this theorem mirrors the results of [10, 77, 80] in a more general setting. As a consequence of this theorem, we prove that the convergence rate achieved in these previous works holds when Ω is locally a smooth manifold.[1] We emphasise that although this result can also be achieved with minor modifications to the arguments in [10, 77, 80], our perspective that this is a special case of a more general setting is novel.

Theorem 3.1.2: (KDE on a.e. smooth spaces)

Suppse Ω is locally a smooth manifold almost everywhere. Assume that K is twice differentiable and vanishes outside $[0,1]$ and the density p is twice differentiable. Then, when the bandwidth parameters h_n are chosen in proportion to $n^{-1/(d+4)}$, the equality

$$\mathrm{MSE}\left[\hat{p}_n(x)\right] = O\left(\frac{1}{n^{4/(d+4)}}\right). \tag{3.5}$$

holds with probability 1.

The proof of Theorem 3.1.2 amounts to showing that a neighbourhood of $x \in \Omega$ is approximated by the tangent space $T_x\Omega$ with only $O(h_n^2)$ error. Note that since algebraic varieties and semi-algebraic sets are smooth almost everywhere, this includes the case that Ω is a $d-$dimensional algebraic variety or semi-algebraic set.

[1] While our results are stated for sets that are locally smooth manifolds since we are interested in examples such as semi-algebraic sets, the hypothesis can be weakened to spaces that are locally only C^3 without changing the proof. Minor modifications may be made to the proof so that the result holds on spaces that are locally only C^2.

Outline

The rest of this section contains a brief overview of related literature. In Section 3.2, we discuss all necessary background material on $d-$rectifiable sets. The proofs of Theorems 3.1.1 and 3.1.2 are contained in Section 3.3 and Section 3.4 respectively. We conclude with Section 3.5 which details a numerical experiment using $\hat{p}_n$ to estimate a known density function on sparse vectors.

Related Literature

Although little attention has been paid to density estimation on non-smooth spaces, there is a wealth of literature on nonparametric density estimation on manifolds, beginning with [41]. The work [41] uses the techniques of Fourier analysis to estimate probability densities on manifolds, yet is limited since it requires knowledge of the eigenfunctions of the Laplace-Beltrami operator on the space. The kernel density estimator (3.3) on Riemannian submanifolds was first studied in [80], but again is limited since knowledge of the space is required. This was generalized to unknown submanifolds by [77]. The analysis in [77] relies on the observation that compactly supported kernels essentially become delta functions as the bandwidth parameter approaches zero. This fact was first observed in [21] in the context of estimating the Laplace-Beltrami operator on manifolds. The setting where data lies on a manifold with boundary was studied in [11], where an additional estimator known as the boundary direction estimator is used to renormalize kernels so that density estimations are consistent as points approach the boundary of the space without needing prior knowledge of the boundary's location.

Our analysis in the proofs of Theorems 3.1.1 and 3.1.2 relies on the existence of (approximate) tangent spaces to the domain Ω. The work [48] also focuses on the existence of tangent spaces and introduces a kernel density estimator that is defined on the tangent space to the manifold Ω. Although it is shown that this estimator does converge to the true denstiy on the domain, this approach is severely limited since it requires knowledge and evaluation of the exponential map on Ω.

More recently, kernel density estimation on unknown submanifolds without boundary was studied in [10]. Under certain conditions relating to the reach of the submanifold, [10] provides convergence rates of the kernel density estimator that depend on the smoothness of the probability density function. This convergence rate mirrors the convergence rate in our Theorem 3.1.1, however our result is dependent on the smoothness of the data's domain and not the probability density function itself. The analysis in [10] uses local parametrizations

of the space by its tangent space, much like our arguments in the proof of Theorem 3.1.2. We emphasise that our main result Theorem 3.1.1 is not predicated on the existence of such parametrizations since they do not exist in general for $d-$rectifiable sets.

3.2 Preliminaries

In this section we provide a brief background on rectifiable sets. For a more thorough treatment of the material, we refer the reader to [95]. First, we recall the definition of $d-$rectifiability.

Definition 3.2.1 ($d-$Rectifiable)**.** A set $\Omega \subset \mathbb{R}^D$ is $d-$rectifiable if it can be decomposed as $\Omega = \Omega_0 \cup (\cup_{j\in\mathbb{N}} F_j(A_j))$ where $\mathcal{H}^d(\Omega_0) = 0$, $A_j \subset \mathbb{R}^d$ and $F_j : \mathbb{R}^d \to \mathbb{R}^D$ is Lipschitz.

Rectifiable subsets of $\mathbb{R}^D$ can be viewed as a natural measure-theoretic generalization of embedded submanifolds of $\mathbb{R}^D$. Indeed, for each point $x \in \Omega$ there is some neighbourhood that is covered by countably many embedded $d-$dimensional C^1-submanifolds of $\mathbb{R}^D$. An alternative characterization of $d-$rectifiable sets arises using *approximate tangent spaces*.

Definition 3.2.2 (Approximate tangent space)**.** Let $\Omega \subset \mathbb{R}^D$ be $\mathcal{H}^d-$measurable and θ a positive $\mathcal{H}^d-$measurable function on Ω with $\int_{\Omega\cap K} \theta d\mathcal{H}^d < \infty$ for all compact sets $K \subset \mathbb{R}^D$. For $x \in \Omega$, the $d-$dimensional subspace $T_x\Omega \subset \mathbb{R}^D$ is an approximate tangent space to Ω with respect to θ if the equality

$$\lim_{h\to 0} \int_{(\Omega-x)/h} f(y)\theta(x+hy)d\mathcal{H}^d(y) = \theta(x) \int_{T_x\Omega} f(y)d\mathcal{H}^d(y) \tag{3.6}$$

holds for all continuous compactly supported functions f. Here, the set $(\Omega - x)/h = \{y \in \mathbb{R}^D : x + hy \in \Omega\}$ is the *blow up* of Ω around x.

Note that any tangent space is also an approximate tangent space. In particular, if a set Ω is locally a differentiable manifold at x, then Ω has an approximate tangent space at x. The following theorem shows that a $d-$rectifiable set $\Omega \subset \mathbb{R}^D$ has approximate tangent spaces $\mathcal{H}^d-$almost everywhere.

Theorem 3.2.3 ([95, Theorem 3.1.9])**.** *Suppose $\Omega \subset \mathbb{R}^D$ is $\mathcal{H}^d-$measurable and θ a positive $\mathcal{H}^d-$measurable function on Ω with $\int_{\Omega\cap K} \theta d\mathcal{H}^d < \infty$ for all compact sets $K \subset \mathbb{R}^D$. Then, Ω is $d-$rectifiable if and only if Ω has an approximate tangent space $T_x\Omega$ with respect to θ for $\mathcal{H}^d-$almost every $x \in \Omega$.*

Theorem 3.2.3 will be the main tool necessary in the proof of Theorem 3.1.1. We will see that the rate at which the mean square error $\mathrm{MSE}\,[\hat{p}_n(x)]$ converges to 0 is controlled by the rate of convergence in (3.6). Since any tangent space to Ω is also an approximate tangent space, if Ω is locally a $d-$dimensional differentiable manifold almost everywhere then Ω is $d-$rectifiable as a direct consequence of Theorem 3.2.3. In particular, objects such as $d-$dimensional algebraic varieties and semi-algebraic sets are also $d-$rectifiable sets.

Lemma 3.2.4. *Suppose Ω is a $d-$dimensional algebraic variety or semi-algebraic set. Then, Ω is $d-$rectifiable.*

Proof. If Ω is a $d-$dimensional algebraic variety or semi-algebraic set, there exists a *Whitney stratification* of Ω - a decomposition of Ω into finitely many non-interecting smooth manifolds with dimension at most d that satisfy certain compatibility conditions, see [106]. In particular, for $\mathcal{H}^d-$almost every $x \in \Omega$, there exists a $d-$dimensional tangent space $T_x\Omega$ to Ω at x. Since $T_x\Omega$ is also an *approximate* tangent space, Ω is $d-$rectifiable. □

3.3 Proof of Theorem 3.1.1

In this section, we prove Theorem 3.1.1. First, we fix some assumptions on the set Ω and the kernel K.

Assumptions. We fix the following setting:

A.1 The kernel K is continuous and vanishes outside of the interval $[0, 1]$, and the equality $\int_{\mathbb{R}^d} K(\|u\|)du = 1$ holds.

A.2 The domain $\Omega \subset \mathbb{R}^D$ is $d-$rectifiable and the approximate tangent spaces to Ω with respect to p satisfy

$$\int_{(\Omega-x)/h} f(y)p(x+hy)d\mathcal{H}^d(y) = p(x)\int_{T_x\Omega} f(y)d\mathcal{H}^d(y) + O(h^m)$$

for some fixed $m > 0$ and each compactly supported continuous function f.

The assumption on K is standard and is satisfied by most reasonable examples. The assumption on the approximate tangent spaces of Ω is necessary to quantify the rate at which the mean squared error $\mathrm{MSE}\,[\hat{p}_n(x)]$ converges to zero as the number of samples n grows. Note that without this assumption, we may still conclude that the equality $\lim_{n\to\infty} \mathrm{MSE}\,[\hat{p}_n(x)] = 0$ holds, but cannot quantify the rate of this convergence.

Recall that the mean square error of the predictor $\hat{p}_n(x)$ can be decomposed into its squared bias and variance:

$$\begin{aligned}\mathrm{MSE}\left[\hat{p}_n(x)\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[\hat{p}_n(x)\right] - p(x)\right)^2\right] \\ &= \left(\mathbb{E}\left[\hat{p}_n(x)\right] - p(x)\right)^2 + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{p}_n(x)\right] - p(x)\right)^2\right] \\ &= \mathrm{Bias}\left[\hat{p}_n(x)\right]^2 + \mathrm{Var}\left[\hat{p}_n(x)\right]\end{aligned}$$

We prove that both the bias and variance of $\hat{p}_n$ converge to zero as the number of samples n grows. This argument is similar in spirit to those of [79], [80], and [77]. We begin by considering the bias $\mathrm{Bias}\left[\hat{p}_n(x)\right]$.

Lemma 3.3.1. *For $\mathcal{H}^d-$almost every $x \in \Omega$ the equality*

$$\frac{1}{h^d}\int_\Omega K\left(\frac{\|x-y\|}{h}\right) dP(y) = p(x) + O(h^m).$$

holds. In particular, if h_n is chosen such that $\lim_{n\to\infty} h_n = 0$, then $\mathrm{Bias}\left[\hat{p}_n(x)\right] = O(h_n^m)$ with probability 1.

Proof. Since p is the density of the probability measure P with respect to the Hausdorff measure $\mathcal{H}^d$, the equalities

$$\begin{aligned}\frac{1}{h^d}\int_\Omega K\left(\frac{\|x-y\|}{h}\right) dP(y) &= \frac{1}{h^d}\int_\Omega K\left(\frac{\|x-y\|}{h}\right) p(y) dP(y) \\ &= \int_{(\Omega-x)/h} K(\|u\|) p(x+uh) d\mathcal{H}^d(u)\end{aligned}$$

hold, with the second equality following from the change of variables $u = \frac{1}{h}(x-y)$. As Ω is $d-$rectifiable, there exists an approximate tangent space $T_x\Omega$ with respect to the density p for almost every x in Ω and the equality

$$\int_{(\Omega-x)/h} K(\|u\|) p(x+uh) d\mathcal{H}^d(u) = p(x)\int_{T_x\Omega} K(\|u\|) d\mathcal{H}^d(u) + O(h^m)$$

holds by assumption A.2 since K is a compactly supported continuous function. Now, since Ω is assumed to be a subset of $\mathbb{R}^D$, the approximate tangent space $T_x\Omega$ is a $d-$dimensional

subspace of $\mathbb{R}^D$. So, there exists some isometry mapping $T_x\Omega$ to $\mathbb{R}^d$. Thus, the equality holds:

$$\int_{T_x\Omega} K(\|u\|)d\mathcal{H}^d(u) = \int_{\mathbb{R}^d} K(\|u\|)d\mathcal{H}^d(u) = 1$$

Then, we have that $\frac{1}{h^d}\int_\Omega K\left(\frac{\|x-y\|}{h}\right)dP(y) - p(x) = O(h^m)$ holds for almost every $x \in \Omega$.

We now show that with probability 1, the equality $\text{Bias}\left[\widehat{p}_n(x)\right] = O(h_n^m)$ holds. Recall that by the definition of bias,

$$\begin{aligned}\text{Bias}\left[\widehat{p}_n(x)\right] &= \frac{1}{h_n^d}\mathbb{E}\left[K\left(\frac{\|x-y\|}{h_n}\right)\right] - p(x) \\ &= \frac{1}{h_n^d}\int_\Omega K\left(\frac{\|x-y\|}{h_n}\right)dP(y) - p(x)\end{aligned}$$

By our previous argument, we have that $\text{Bias}\left[\widehat{p}_n(x)\right] = O(h_n^m)$. This concludes the proof. □

We now utilize similar arguments to show that the variance $\text{Var}\left[\widehat{p}_n(x)\right]$ converges to zero as the number of samples n increases.

Lemma 3.3.2. *Suppose that h_n is chosen such that $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^d = \infty$. Then, with probability 1, the equality holds for all $h < h_0$:*

$$\text{Var}\left[\widehat{p}_n(x)\right] = O\left(\frac{1}{nh_n^d}\right).$$

Proof. Recall that by the definition of variance,

$$\text{Var}\left[\widehat{p}_n(x)\right] = \frac{1}{n}\mathbb{E}\left[\frac{1}{h_n^{2d}}K^2\left(\frac{\|x-y\|}{h_n}\right)\right] - \frac{1}{n}\left(\mathbb{E}\left[\frac{1}{h_n^d}K\left(\frac{\|x-y\|}{h_n}\right)\right]\right)^2. \tag{3.7}$$

We will consider each term in the sum independently. Consider the first term in (3.7)

$$\frac{1}{n}\mathbb{E}\left[\frac{1}{h_n^{2d}}K^2\left(\frac{\|x-y\|}{h_n}\right)\right] = \frac{1}{nh_n^d}\int_\Omega \frac{1}{h_n^d}K^2\left(\frac{\|x-y\|}{h_n}\right)p(y)dP(y).$$

Using a similar argument to that in the proof of Lemma 3.3.1, we compute:

$$\begin{aligned}\frac{1}{nh_n^d}\int_\Omega \frac{1}{h_n^d}K^2\left(\frac{\|x-y\|}{h_n}\right)p(y)dP(y) &= \frac{1}{nh_n^d}\left(\int_{\mathbb{R}^d} K^2(\|u\|)du + O(h_n^m)\right) \\ &\leq \frac{Cp(x)}{nh_n^d} + O\left(\frac{h_n^m}{nh_n^d}\right),\end{aligned}$$

where $C \geq \int_{\mathbb{R}^d} K^2(\|u\|)du$. Note that such a constant C exists since K is continuous and compactly supported. By assumption, $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^d = \infty$ and so it follows that the first term in (3.7) converges to zero at the rate $O(1/(nh_n^d))$.

Now, observe that by the arguments in the proof of Lemma 3.3.1, the equality holds:

$$\left(\mathbb{E}\left[\frac{1}{h_n^d}K\left(\frac{\|x-y\|}{h_n}\right)\right]\right)^2 = p^2(x) + O(h_n^m).$$

Dividing by the number of samples n, the second term in (3.7) is $O(1/n)$ as n increases. As we may assume that $h_n < 1$, note that $O(1/n) \subset O(1/(nh_n^d))$. So, we have that the equality

$$\operatorname{Var}\left[\hat{p}_n(x)\right] = O\left(\frac{1}{nh_n^d}\right)$$

holds, as desired. This completes the proof. □

The proof of Theorem 3.1.1 is now a simple combination of Lemmas 3.3.1 and 3.3.2.

Proof of Theorem 3.1.1. Since the mean square error of $\hat{p}_n(x)$ can be decomposed into squared bias and variance, we compute:

$$\begin{aligned}\operatorname{MSE}\left[\hat{p}_n(x)\right] &= \operatorname{Bias}\left[\hat{p}_n(x)\right]^2 + \operatorname{Var}\left[\hat{p}_n(x)\right] \\ &= O\left(h_n^{2m} + \frac{1}{nh_n^d}\right)\end{aligned}$$

Choosing the bandwidth parameter h_n in proportion to $n^{-1/(d+2m)}$, we achieve our desired rate of convergence:

$$\lim_{n\to\infty} \operatorname{MSE}\left[\hat{p}_n(x)\right] = O\left(\frac{1}{n^{2m/(d+2m)}}\right)$$

This concludes our proof of Theorem 3.1.1. □

3.4 Proof of Theorem 3.1.2

In this section, we prove that the convergence rate (3.4) holds when Ω is locally a smooth manifold almost everywhere. To this end, we introduce the following lemma, which quantifies the approximation error when integrating over the tangent space $T_x\Omega$ as opposed to the blow up $(\Omega - x)/h$. The proof uses standard techniques but is included for completeness.

Lemma 3.4.1. *Suppose there exists some $h_0 > 0$ such that $B(x, h_0) \cap \Omega$ is a smooth $d-$dimensional submanifold of $\mathbb{R}^D$. Let f be even, twice differentiable and compactly supported and θ a twice differentiable function on Ω. Then, the equality holds:*

$$\int_{(\Omega - x)/h} f(y)\theta(x + hy)d\mathcal{H}^d(y) = \int_{T_x\Omega} f(y)\theta(x)d\mathcal{H}^d(y) + O(h^2)$$

Proof. As the Hausdorff measure is preserved under affine isometries, we may assume without loss of generality that $x = 0$ and $T_x\Omega = \mathbb{R}^d \subset \mathbb{R}^D$, i.e. $T_x\Omega = \{(z, 0) \in \mathbb{R}^D : z \in \mathbb{R}^d\}$. So, we prove the following equality:

$$\int_{\Omega/h} f(y)\theta(hy)d\mathcal{H}^d(y) = \int_{\mathbb{R}^d} f(z, 0)\theta(0)dz + O(h^2) \tag{3.8}$$

Note that since f is assumed to be compactly supported, we may intersect the domain of the integral with some compact set. Thus, we may assume that all functions and derivatives are bounded on our domain. First, we expand the integrand using the Taylor expansion of the function θ. Since θ is assumed to be twice differentiable, the equality holds:

$$\int_{\Omega/h} f(y)\theta(hy)d\mathcal{H}^d(y) = \theta(0)\int_{\Omega/h} f(y)d\mathcal{H}^d(y) + h\int_{\Omega/h} f(y)\langle D\theta(0), y\rangle d\mathcal{H}^d(y) + O(h^2) \tag{3.9}$$

We will independently consider each integral in the sum. Since Ω is a smooth manifold about 0, for small enough h we may assume that the domain Ω/h can be locally parametrized as $\{(y/h), \phi(y)/h : y \in U\}$ where ϕ is a smooth map with $\phi(0) = 0$ and $D\phi(0) = 0$ and U is some neighbourhood of the origin in $\mathbb{R}^d$. Define the map $\psi_h : \mathbb{R}^d \to \mathbb{R}^D$ as $\psi_h(z) = (z, \phi(hz)/h)$. We now perform a change of variables using this parametrization of Ω/h :

$$\int_{\Omega/h} f(y)d\mathcal{H}^d(y) = \int_{\mathbb{R}^d} f(z, \phi(hz)/h)J_h(z)d\mathcal{H}^d(y)$$

where $J_h(y) = \sqrt{\det(D\psi_h(z)^\top D\psi_h(z))}$ is the Jacobian determinant of the map ψ_h. By definition, the differential $D\psi_h(z)$ is

$$D\psi_h(z) = \begin{bmatrix} I_d \\ D\phi(hz) \end{bmatrix}$$

where I_d is the $d \times d$ identity matrix. We approximate $D\phi(hz)$ with a second order Taylor

approximation around 0. First, note that the Taylor expansion of $\phi(hz)$ is given by

$$\phi(hz) = \phi(0) + hD\phi(0)(z) + \frac{h^2}{2}D^2\phi(0)(z,z) + O(h^3) \tag{3.10}$$

and so taking the derivative with respect to z,

$$D\phi(hz) = D\phi(0) + h^2D^2\phi(0)(z) + O(h^3) = h^2D^2\phi(0)(z) + O(h^3)$$

since $D\phi(0) = 0$ by assumption. So, $D\psi_h(z)^\top D\psi_h(z) = I_d + h^4(D\phi(0)(z))^\top(D\phi(0)(z)) + O(h^5)$. Using the Taylor expansion of the determinant near the identity, we compute

$$\begin{aligned}\det(I_d + h^4(D\phi(0)(z))^\top(D\phi(0)(z)) + O(h^5)) &= 1 + h^4\mathrm{Trace}((D\phi(0)(z))^\top(D\phi(0)(z))) + O(h^4)\\ &= 1 + O(h^4).\end{aligned}$$

It follows that the equality

$$\sqrt{\det(D\psi_h(z)^\top D\psi_h(z))} = 1 + O(h^4)$$

holds. In particular, we have shown that the equality holds:

$$\int_{\Omega/h} f(y)d\mathcal{H}^d(y) = \int_{\mathbb{R}^d} f(z,\phi(hz)/h)dz + O(h^2)$$

We now compute the Taylor expansion of $f(z,\phi(hz)/h)$ in the normal direction to $\mathbb{R}^d \subset \mathbb{R}^D$. Given the Taylor expansion of $\phi(hz)$ in (3.10), it follows that the equality

$$f(z,\phi(hz)/h) = f(z,0) + \frac{h}{2}\langle D_\perp f(z,0), D^2\phi(0)(z,z)\rangle + O(h^2) \tag{3.11}$$

holds, where $D_\perp f(z,0)$ is the differential of f at $(z,0)$ in the normal direction. Note that since f is assumed to be even, $D_\perp f(z,0)$ is an odd function of z. We compute

$$\begin{aligned}\int_{\mathbb{R}^d} f(z,\phi(hz)/h)dz &= \int_{\mathbb{R}^d} f(z,0)dz + \frac{h}{2}\int_{\mathbb{R}^d}\langle D_\perp f(z,0), D^2\phi(0)(z,z)\rangle dz + O(h^2)\\ &= \int_{\mathbb{R}^d} f(z,0)dz + O(h^2),\end{aligned}$$

where the final equality holds since the inner product $\langle D_\perp f(z,0), D^2\phi(0)(z,z)\rangle$ is an odd

function of z. Thus, we have shown that the equality

$$\theta(0)\int_{\Omega/h} f(y)d\mathcal{H}^d(y) = \theta(0)\int_{\mathbb{R}^d} f(z,0)dz + O(h^2)$$

holds. We now need only show that the second term in the sum (3.9) is $O(h^2)$. Again, we use the change of variables $y = (z, \phi(hz)/h)$:

$$h\int_{\Omega/h} f(y)\langle D\theta(0), y\rangle d\mathcal{H}^d(y) = h\int_{\mathbb{R}^d} f(z,\phi(hz)/h)\langle D\theta(0),(z,\phi(hz)/h)\rangle dz + O(h^2)$$

Using the Taylor expansion of $\phi(hz)/h$ in (3.10) note that the equalities hold:

$$\begin{aligned}\langle D\theta(0),(z,\phi(hz)/h)\rangle &= \langle D\theta(0),(z,0)+(0,\frac{h}{2}D^2\phi(0)(z,z)+O(h^2))\rangle \\ &= \langle D\theta(0),(z,0)\rangle + \frac{h}{2}\langle D\theta(0),(0,D^2\phi(0)(z,z))\rangle + O(h^2)\end{aligned}$$

Taking the product with the expansion of $f(z,\phi(hz)/h)$ in (3.11), we compute

$$h\int_{\mathbb{R}^d} f(z,\phi(hz)/h)\langle D\theta(0),(z,\phi(hz)/h)\rangle dz = h\int_{\mathbb{R}^d} f(z,0)\langle D\theta(0),(z,0)\rangle dz + O(h^2)$$

Since f is an even function by assumption, the product $f(z,0)\langle D\theta(0),(z,0)\rangle$ is an odd function of z. It follows that the equality

$$\int_{\mathbb{R}^d} f(z,0)\langle D\theta(0),(z,0)\rangle dz = 0$$

holds and thus only the $O(h^2)$ term remains. This completes the proof. □

Proof of Theorem 3.1.2. By Lemma 3.4.1, the convergence rate in Theorem 3.1.1 holds with $m = 2$. It follows that the equality

$$\text{MSE}\left[\hat{p}_n(x)\right] = O\left(\frac{1}{n^{4/(d+4)}}\right)$$

holds for almost every $x \in \Omega$ when h_n is chosen in proportion to $n^{-1/(d+4)}$ as desired. □

We have shown that the kernel density estimator $\hat{p}_n(x)$ converges to to the true density $p(x)$ at the rate $O(n^{-4/(d+4)})$ on sets that are locally smooth $d-$dimensional manifolds, mirroring the classical results of [79, 15]. By Lemma 3.2.4, this includes all $d-$dimensional

algebraic varieties and semi-algebraic sets. This is a much more rich class of spaces than just $d-$dimensional manifolds. For example, the spaces of sparse vectors and low-rank matrices are algebraic varieties and the space of positive semi-definite matrices is a semi-algebraic set. Our result ensures that kernel density estimation can be used to estimate unknown density functions on these spaces and converges to the true denstiy at the same rate as classical kernel density estimation. The next section illustrates this convergence rate by estimating a known density function on sparse vectors.

3.5 Example: Density Estimation on Sparse Data

One common structure used in dimensionality reduction is assuming that data is $d-$sparse: although the data lives in the high-dimensional space $\mathbb{R}^D$, each sample only has d non-zero coordinates where d is typically much smaller than D. We now illustrate the convergence of the estimator $\hat{p}_n$ by estimating a probability density on $d-$sparse vectors.

We fix the following distribution on the set of $d-$sparse vectors in $\mathbb{R}^D$: data is sampled from a mixture of Gaussians by uniformly at randomly selecting a $d-$dimensional coordiante hyperplane and randomly sampling a vector according to the standard $d-$dimensional normal distribution in the hyperplane. For a $d-$sparse vector $x \in \mathbb{R}^D$, the true density of this distribution at the point x is given by

$$p(x) = \frac{1}{\binom{D}{d}} \cdot \frac{\exp\left(-\|x\|^2/2\right)}{\sqrt{(2\pi)^d}}. \tag{3.12}$$

since there are $\binom{D}{d}$ $d-$dimensional coordinate planes and each hyperplane is chosen uniformly at random.

To estimate the density p we use the Kernel density estimator $\hat{p}_n(x)$ with K chosen to be the truncated Gaussian kernel

$$K\left(\frac{\|x-y\|}{h}\right) = \begin{cases} \frac{1}{h\sqrt{2\pi}(F(1)-F(-1))} \cdot \exp\left(-\frac{\|x-y\|^2}{2h^2}\right) & \|x-y\| \leq h \\ 0 & \|x-y\| > h \end{cases}$$

where F is the cumulative distribution function of the standard normal distribution. We use the truncated Gaussian kernel rather than the standard Gaussian kernel since it is compactly supported on the interval $[0, 1]$.

Note that the set of $d-$sparse vectors is a $d-$dimensional variety in $\mathbb{R}^D$ since it is the vanishing locus of the set of all degree $d+1$ monomials. Thus, from the result of Theorem

3.1.2 we expect the MSE of the estimator $\hat{p}_n(x)$ to converge to 0 at the rate $O(n^{-4/(d+4)})$ regardless of the ambient dimension D of the data. We fixed the level of sparsity of the data at $d = 3$ and tested the estimator $\hat{p}_n$ in ambient dimensions $D = 5, 10, 50$. In each setting, we constructed an estimator using training sets of various sizes: we began with using only 100 training points and increased the number of data points in increments of 100 until the training set consisted of 10,000 points. For a training set of size n, the bandwidth parameter h_n was chosen to be $n^{-1/7}$ in accordance with the choice of parameter in Theorem 3.1.2. We calculated the empirical MSE of these predictors using a test set of 500 points. For each regime, we repeated the experiment 10 times. Our results are shown in Figure 3.1.

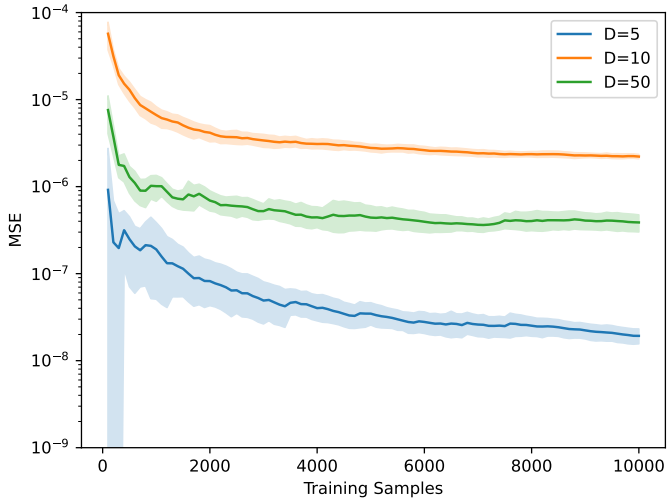


Figure 3.1: MSE of the predictor $\hat{p}_n$ in various ambient dimensions and using varying sizes of training sets. The shaded regions indicate a 95% confidence interval. Note that the curve corresponding to each ambient dimension D has the same shape, illustrating that the convergence rate is independent of the ambient dimension.

3.6 Conclusion

In this chapter, we studied the problem of probability density estimation on $d-$rectifiable sets and proved that a suitably modified kernel density estimator converges to the true density at a rate consistent with previous results. This work extends the literature on density estimation on manifolds to a much richer class of spaces. The theory developed is supported by a numerical experiment estimating a known density on sparse vectors.

Chapter 4

UNIVERSAL GRÖBNER BASES OF (UNIVERSAL) MULTIVIEW VARIETIES

This chapter is the content of the paper [28] and is joint work with Duff and Thomas. Multiview ideals arise from the geometry of image formation in pinhole cameras, and universal multiview ideals are their analogs for unknown cameras. We prove that a natural collection of polynomials form a universal Gröbner basis for both types of ideals using a criterion introduced by Huang and Larson, and include a proof of their criterion in our setting. Symmetry reduction and induction enable the method to be deployed on an infinite family of ideals. We also give an explicit description of the matroids on which the methodology depends, in the context of multiview ideals.

4.1 Introduction

In this chapter, we apply a recent universal Gröbner basis criterion due to Huang and Larson [43] to show that certain sets of polynomials form universal Gröbner bases for two families of ideals from computer vision.

First, we recover a result for *multiview ideals*, originally [2, Theorem 2.1].

Theorem 4.1.1: (Universal Gröbner Basis for Multiview Ideals)

or an arrangement of n generic pinhole cameras, the associated 2-,3-, and 4-focals form a universal Gröbner basis of the multiview ideal I_n.

We then extend this result to prove a new result about a universal Gröbner basis of *universal multiview ideals*, introduced in [1].

Theorem 4.1.2: (Universal Gröbner Basis for Universal Multiview Ideals)

or an arrangement of n unknown cameras, the associated 2-,3-, and 4-focals form a universal Gröbner basis of the universal multiview ideal $\tilde{I}_n$.

We define multiview and universal multiview ideals in Sections 4.4 and 4.5. Both form families of ideals indexed by $2 \leq n \in \mathbb{N}$. The Huang-Larson criterion for a set of polynomials

to form a universal Gröbner basis is for a fixed ideal, and it relies on an abstract simplicial complex that is generated by the supports of the candidate polynomials. In order to apply this criterion to a family of ideals, we first need a handle on the corresponding family of simplicial complexes. We rely on combinatorics, symmetry reduction, and finally induction, to reduce the checks to a small finite number. We also give explicit descriptions of the complexes as transversal matroids.

This chapter is organized as follows. In Section 4.2 we give the necessary background on simplicial complexes and matroids, followed by the Huang-Larson criterion for a set of square-free polynomials to form a universal Gröbner basis. For our results, we only need to apply this criterion in the special setting where the underlying varieties are irreducible. In Section 4.3, we include a self-contained proof of the Huang-Larson criterion in this special setting. In Section 4.4 we apply the criterion to multiview ideals and in Section 4.5 to universal multiview ideals. In the former case we recover a known result (Theorem 4.1.1) using this new technique. Using a similar approach we prove Theorem 4.1.2, a new result which generalizes [1, Theorem 3.2], and resolves the first open question in [1, §8.1].

4.2 Background

4.2.1 Simplicial Complexes and Matroids

In this subsection, we collect some standard combinatorial terminology and facts that play a role in our results.

An *abstract simplicial complex* $\Delta = (S, \mathcal{I})$ consists of a finite ground set S and a downward-closed family of subsets $\mathcal{I}$: if $A \subseteq S$ satisfies $A \in \mathcal{I}$, then we have $B \in \mathcal{I}$ whenever $B \subseteq A$. The elements of $\mathcal{I}$ are known as the faces of Δ, and inclusion-maximal faces are called *facets*. The *dimension* of a face is one less than its cardinality, and the dimension of a complex is the largest dimension over all of its facets. A complex is *pure* if all of its facets' dimensions are equal. Two complexes are *isomorphic* if there is a face-preserving bijection between their ground sets.

Next we recall some basic aspects of *Stanley-Reisner theory* —see eg. [71, 98, Ch. 1] for more details. The *Stanley-Reisner ideal* of a simplicial complex Δ on S is the monomial ideal J_Δ in $\mathbb{C}[x_i : i \in S]$ generated by the square-free monomials $x^A := \Pi_{i \in A} x_i$ such that A is a *nonface* of Δ, i.e., $A \notin \Delta$. If Δ is pure then J_Δ is equidimensional, i.e., all its associated prime ideals have the same dimension. On the other hand, every ideal J in $\mathbb{C}[x_i : i \in S]$ generated by square-free monomials is the Stanley-Reisner ideal of a simplicial complex Δ_J

on S; $x^A \in J$ if and only if A is a nonface of Δ_J. The inclusion-minimal nonfaces of Δ_J are in bijection with the minimal generators of J. The simplicial complex Δ_J is called the *Stanley-Reisner complex* of J.

A *matroid* $\mathcal{M} = (S, \mathcal{I})$ on a finite ground set S is an abstract simplicial complex, known also as an *independence complex*, which satisfies the *augmentation axiom*: if $I_1, I_2 \in \mathcal{I}$ with $\#I_1 < \#I_2$, then there exists $s \in I_2 \setminus I_1$ with $I_1 \cup \{s\} \in \mathcal{I}$. Basic matroid theory shows that all independence complexes are pure; their faces and facets are known, respectively, as the *independent sets* and the *bases* of the matroid. The *rank function* $r_{\mathcal{M}}$ of a matroid $\mathcal{M} = (S, \mathcal{I})$ is the set function satisfying

$$r_{\mathcal{M}}(X) = \max\{\#I \mid I \in \mathcal{I},\ I \subseteq X\} \qquad \forall X \subseteq S.$$

The rank of $\mathcal{M} = (S, \mathcal{I})$ is defined as $r_{\mathcal{M}}(S)$ and equals the size of any basis of $\mathcal{M}$.

Example 10. The *uniform* matroid $\mathcal{U}_{k,n}$ has ground set $[n] := \{1, \ldots, n\}$, and its independent sets are all subsets of $[n]$ of size at most k.

Example 11. Let G be a bipartite graph on vertex set $S \sqcup T$. We define the *G-transversal matroid on S*, denoted $\mathcal{M}[G, S] = (S, \mathcal{I})$, to consist of all subsets $I \subset S$ that can be covered by a matching in G.

In general, we say a matroid is uniform (resp. transversal) if its is isomorphic to some uniform (resp. transversal) matroid. Thus, all uniform matroids are transversal: for the complete bipartite graph $K_{k,n}$, we have $\mathcal{U}_{k,n} \cong \mathcal{M}[K_{k,n}, [n]]$

New matroids can be built from old ones using the *matroid union theorem.* Given n matroids $\mathcal{M}_1 = (S_1, \mathcal{I}_1), \ldots, \mathcal{M}_n = (S_n, \mathcal{I}_n)$, their *union*

$$\mathcal{M} = \mathcal{M}_1 \vee \cdots \vee \mathcal{M}_n = (S, \mathcal{I})$$

is defined so that $S = S_1 \cup \cdots \cup S_n$ and

$$\mathcal{I} = \{I_1 \cup \cdots \cup I_n \mid I_1 \in \mathcal{I}_1, \ldots, I_n \in \mathcal{I}_n\}.$$

Note that the ground sets S_i need not be disjoint in the definition of matroid union. When the ground sets are *pairwise disjoint*, meaning $S_i \cap S_j = \emptyset$ whenever $i \neq j$, the matroid union is instead called the *direct sum* and denoted $\mathcal{M}_1 \oplus \cdots \oplus \mathcal{M}_n$.

Theorem 4.2.1. *[76, Theorem 12.3.1] The union of matroids $\mathcal{M} = \mathcal{M}_1 \vee \cdots \vee \mathcal{M}_n$ is also a matroid, whose rank function $r_{\mathcal{M}}$ is related to the rank functions $r_1, \ldots, r_n$ of $\mathcal{M}_1, \ldots, \mathcal{M}_n$ as follows:*

$$r_{\mathcal{M}}(X) = \min\left\{ \sum_{i=1}^{n} r_i(Y) + \#(X \setminus Y) \mid Y \subseteq X \right\}.$$

Remark 4.2.2. In general, unions of bases need not be bases. For example, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are the rank-1 uniform matroids on $\{1\}$ and $\{1, 2\}$, respectively, then

$$B_1 = B_2 = \{1\} \quad \Rightarrow \quad B_1 \cup B_2 = \{1\} \subsetneq \{1, 2\},$$

so $B_1 \cup B_2$ is not a basis in the union $\mathcal{M} = \mathcal{M}_1 \vee \mathcal{M}_2$. However, if $B = B_1 \cup \cdots \cup B_n$ with the B_i pairwise disjoint, then B is a basis for $\mathcal{M} = \mathcal{M}_1 \vee \cdots \vee \mathcal{M}_n$, as this implies $r_{\mathcal{M}}(B) = \#B$, with the minimum in Theorem 4.2.1 attained at any $Y \subseteq B$.

4.2.2 Universality Criterion for Gröbner Bases

The paper [43] gives a criterion for determining whether a set of nonzero square-free polynomials is a universal Gröbner basis of an ideal. In this section, we give a brief overview of this criterion and describe our general strategy for proving Theorems 4.1.1 and 4.1.2.

Fix a variety $V \subset \mathbb{C}^N$. For any polynomial $f \in \mathbb{C}[x_1, \ldots, x_N]$, define the *spread* of f to be the collection of $i \in [N]$ such that x_i is in the support of f. For a collection of nonzero polynomials $f_1, \ldots, f_r$ let $\Delta(f_1, \ldots, f_r)$ be the simplicial complex on $[N]$ whose nonfaces are generated by the spreads of $f_1, \ldots, f_r$.

There is a natural embedding of V in the projective space $\mathbb{P}^N$ obtained by homogenizing each coordinate x_i, i.e., sending $x_i \mapsto [x_i : 1]$. Let $\overline{V}$ denote the closure of V inside $\mathbb{P}^N$. Given a subset $U \subset [N]$, consider the projection

$$\begin{aligned} \overline{\pi}_U : \mathbb{P}^N &\to \mathbb{P}^U \\ ([x_1 : x_{N+1}], \ldots, [x_N : x_{2N}]) &\mapsto ([x_i : x_{i+N}] \mid i \in U). \end{aligned} \tag{4.1}$$

The following theorem gives a sufficient condition for when $\{f_1, \ldots, f_r\}$ forms a universal Gröbner basis for the vanishing ideal $\mathcal{I}(V)$.

Theorem 4.2.3 (Theorem 2.7 in [43])**.** *Let $V \subset \mathbb{C}^N$ be a closed subvariety, and $f_1, \ldots, f_r \in \mathcal{I}(V) \setminus \{0\}$ be square-free polynomials. If $\overline{\pi}_U(\overline{V}) = \mathbb{P}^U$ for every facet U of $\Delta(f_1, \ldots, f_r)$, then $\{f_1, \ldots, f_r\}$ is a universal Gröbner basis of $\mathcal{I}(V)$.*

To apply Theorem 4.2.3, we must show that the projection $\overline{\pi}_U : \overline{V} \to \mathbb{P}^U$ is for each facet of the simplicial complex $\Delta(f_1, \dots, f_r)$. Since our varieties are affine, it will be convenient to instead consider the projections $\pi_U : V \to \mathbb{C}^U$. The following simple lemma expresses the the surjectivity, and hence dominance, of $\overline{\pi}_U$ in terms of the original affine variety V, when V is irreducible.

Lemma 4.2.4. *Let $V \subset \mathbb{C}^N$ be irreducible. For any subset $U \subset \{1, \dots, N\}$, the projection $\pi_U : V \to \mathbb{C}^U$ is dominant if and only if $\overline{\pi}_U : \overline{V} \to \mathbb{P}^U$ is surjective.*

Proof. Consider the following diagram:

$$\begin{array}{ccc} V & \xhookrightarrow{i_V} & \overline{V} \\ \downarrow{\scriptstyle \pi_U} & & \downarrow{\scriptstyle \overline{\pi}_U} \\ \mathbb{C}^U & \xhookrightarrow{i_U} & \mathbb{P}^U \end{array}$$

Since the horizontal inclusion maps are birational isomorphisms, the image of π_U is dense if and only if the image of $\overline{\pi}_U$ is dense. Furthermore, dominance and surjectivity of $\overline{\pi}_U$ are equivalent by the projective elimination theorem. □

While Theorem 4.2.3 gives a sufficient condition for a collection of polynomials to form a universal Gröbner basis of a single ideal, in this chapter we wish to prove that a given sequence of sets of polynomials form universal Gröbner bases for a nested sequence of ideals I_n, indexed by $n \in \mathbb{N}$. Our general strategy is as follows:

1. For each n, describe the simplicial complex Δ_n needed in Theorem 4.2.3 for I_n, and determine a "growth rule" as n increases.
2. Use symmetry to reduce the number of facets that need to be checked in Theorem 4.2.3.
3. Prove a base case via direct computation in Macaulay2 [38].
4. Finally, apply induction to prove the general case.

4.3 A proof of Theorem 4.2.3 in the irreducible case

In this section, we provide a short proof of Theorem 4.2.3 under the assumption that the variety V is irreducible. This gives us a self-contained route towards our main results, where this assumption always holds.

Our task is to show that under the hypothesis of Theorem 4.2.3, $\{f_1, \dots, f_r\}$ is a Gröbner basis for $\mathcal{I}(V)$ with respect to any monomial order $<$ on $\mathbb{C}[x_1, \dots, x_N]$.

Let $\mathbb{C}[x_1, \dots, x_N, x_{N+1}, \dots, x_{2N}]$ be the coordinate ring of $\mathbb{P}^N$ where variable x_i is paired with x_{N+i} for $i = 1, \dots, N$. The *multi-homogenization* of a polynomial $f \in \mathbb{C}[x_1, \dots, x_N]$ is $f^h \in \mathbb{C}[x_1, \dots, x_{2N}]$ defined as

$$f^h(x_1, \dots, x_N) = x_{N+1}^{\deg_{x_1}(f)} \cdots x_{2N}^{\deg_{x_N}(f)} \cdot f(x_1/x_{N+1}, \dots, x_N/x_{2N}). \tag{4.2}$$

The vanishing ideal of $\overline{V}$ in $\mathbb{C}[x_1, \dots, x_{2N}]$ is

$$\mathcal{I}(\overline{V}) = \langle f^h \mid f \in \mathcal{I}(V) \rangle. \tag{4.3}$$

In order to show that $\{f_1, \dots, f_r\}$ is a Gröbner basis of $\mathcal{I}(V)$ with respect to any monomial order on $\mathbb{C}[x_1, \dots, x_N]$, it suffices to argue that $\{f_1^h, \dots, f_r^h\}$ is a Gröbner basis for $\mathcal{I}(\overline{V})$ with respect to any product order such that $x_{j+N} < x_i$ for all i, j. These orders can be viewed as natural extensions of monomial orders on $\mathbb{C}[x_1, \dots, x_N]$ to monomial orders on $\mathbb{C}[x_1, \dots, x_{2N}]$. If $f_1^h, \dots, f_r^h$ form a Gröbner basis of $\mathcal{I}(\overline{V})$ with respect to $<$, then $f_1, \dots, f_r$ form a Gröbner basis of $\mathcal{I}(V)$ with respect to the corresponding monomial order on $\mathbb{C}[x_1, \dots, x_N]$, see for example, [29, Exercise 15.39].

Fix such an order $<$ on $\mathbb{C}[x_1, \dots, x_{2N}]$ and define the monomial ideal

$$J := \langle \mathrm{in}_<(f_1^h), \dots, \mathrm{in}_<(f_r^h) \rangle \subset \mathbb{C}[x_1, \dots, x_{2N}]. \tag{4.4}$$

To show that $f_1^h, \dots, f_r^h$ form a Gröbner basis of $\mathcal{I}(\overline{V})$ with respect to $<$, we will apply the following result of Conner, Han, and Michalek [24] to $I = \mathcal{I}(\overline{V})$.

Lemma 4.3.1 (Lemma 3.1 in [24])**.** *Suppose I is a homogeneous prime ideal and $g_1, \dots, g_r$ are elements of I. Assume the following hold for the monomial order $<$:*

1. *The initial terms $\mathrm{in}_<(g_1), \dots, \mathrm{in}_<(g_r)$ are square-free,*
2. *The initial ideal $J := \langle \mathrm{in}_<(g_1), \dots, \mathrm{in}_<(g_r) \rangle$ is equidimensional, and*
3. *The equalities $\dim I = \dim J$, and $\deg I = \deg J$ hold.*

Then, $g_1, \dots, g_r$ is a Gröbner basis of I with respect to the monomial order $<$.

Since the variety V is irreducible by assumption , so is the variety $\overline{V}$. Thus, the vanishing ideal $\mathcal{I}(\overline{V})$ is prime and homogenous. By assumption (in the setting of Theorem 4.2.3),

the initial monomials $\mathrm{in}_<(f_1^h), \dots, \mathrm{in}_<(f_r^h)$ are square-free and so we need only prove that conditions (2) and (3) hold. Let $\mathrm{spr} f$ denote the spread of a polynomial f.

The following lemma, summarizing two results in [43], will be useful for us.

Lemma 4.3.2. *Suppose $V \subset \mathbb{C}^N$ is a closed subvariety. For a subset $U \subset [N]$, we have that $\overline{\pi}_U(\overline{V}) \neq \mathbb{P}^U$ if and only if there is a nonzero $f \in \mathcal{I}(V)$ such that $\mathrm{spr} f \subset U$. Alternatively, $\overline{\pi}_U(\overline{V}) = \mathbb{P}^U$ if and only if $\mathcal{I}(V) \cap \mathbb{C}[x_i : i \in U] = 0$.*

Proof. This is an easy consequence of Lemmas 2.1 and 2.2 in [43]. □

Since the variety V is irreducible, the sets $U \subset [N]$ with $\overline{\pi}_U(\overline{V}) = \mathbb{P}^U$ are precisely the independent sets in an *algebraic matroid* [76, §6.7], whose rank is $\dim V$. The bases in this matroid correspond to transcendence bases of rational functions in the set $\{x_1, \dots, x_N\} \subset \mathbb{C}(V)$. As in [84], we call this the *algebraic matroid of V*.

Lemma 4.3.3. *Under the hypotheses of Theorem 4.2.3 (and since V is irreducible), the simplicial complex $\Delta(f_1, \dots, f_r)$ is the algebraic matroid of V. In particular, $\Delta(f_1, \dots, f_r)$ is a pure complex of dimension $\dim V - 1$.*

Proof. Under the hypotheses of Theorem 4.2.3, every face of $\Delta(f_1, \dots, f_r)$ is independent in the algebraic matroid of V. On the other hand, if U is a nonface of $\Delta(f_1, \dots, f_r)$, then U contains the spread of some f_i, and Lemma 4.3.2 implies U is dependent. Thus, the algebraic matroid of V coincides with the simplicial complex $\Delta(f_1, \dots, f_r)$. Furthermore, since the bases of a matroid all have the same size, $\Delta(f_1, \dots, f_r)$ is a pure simplicial complex. □

Returning to the proof of Theorem 4.2.3, we now prove that condition (2) holds, i.e., that $J = \langle \mathrm{in}_<(f_1^h), \dots, \mathrm{in}_<(f_r^h) \rangle$ is equidimensional, using Stanley-Reisner theory. Observe that the Stanley-Reisner complex Δ_J is closely related to the simplicial complex $\Delta(f_1, \dots, f_r)$; for any $f \in \mathbb{C}[x_1, \dots, x_N]$ there is a bijective correspondence between $\mathrm{spr} f$ in $[N]$ and the support of $\mathrm{in}_<(f^h)$ in $[2N]$. Hence, the nonfaces of Δ_J and $\Delta(f_1, \dots, f_r)$ are in bijection. Moreover, by [43, Lemma 2.10], there is a bijection sending facets $U \in \Delta(f_1, \dots, f_r)$ of size k to facets $\widetilde{U} \in \Delta_J$ of size $k + N$.

Example 12. For example, if $f = x_1 + x_2 x_3 \in \mathbb{C}[x_1, x_2, x_3]$, then $\mathrm{spr} f = \{1, 2, 3\}$ and $\Delta_{\langle f \rangle}$ has the single generating nonface $\{1, 2, 3\}$. Double the variables by letting $x_1 \leftrightarrow x_4, x_2 \leftrightarrow x_5, x_3 \leftrightarrow x_6$. Then $f^h = x_1 x_5 x_6 + x_4 x_2 x_3$, and for any term order with $\mathrm{in}_<(f^h) = x_1 x_5 x_6$, the complex $\Delta_{\langle \mathrm{in}_<(f^h) \rangle}$ is generated by the nonface $\{1, 5, 6\}$. The indices $2, 3$ in $\{1, 2, 3\}$

have been replaced by $5, 6$ in $\{1, 5, 6\}$. We also have the following bijective correspondence between facets:

$$\begin{aligned}
F = \{1, 2\} \in \Delta_{\langle f \rangle} &\leftrightarrow \tilde{F} = \{1, 2, 3, 4, 5\} \in \Delta_{\langle \mathrm{in}_<(f) \rangle}, \\
F = \{1, 3\} \in \Delta_{\langle f \rangle} &\leftrightarrow \tilde{F} = \{1, 2, 3, 4, 6\} \in \Delta_{\langle \mathrm{in}_<(f) \rangle}, \\
F = \{2, 3\} \in \Delta_{\langle f \rangle} &\leftrightarrow \tilde{F} = \{2, 3, 4, 5, 6\} \in \Delta_{\langle \mathrm{in}_<(f) \rangle}.
\end{aligned}$$

Lemma 4.3.4. *The ideal J is equidimensional.*

Proof. As observed in the previous discussion, the ideal J is the Stanley-Reisner ideal of the simplicial complex $\Delta_J = \Delta_{\langle \mathrm{in}_<(f_1^h), \ldots, \mathrm{in}_<(f_r^h) \rangle}$. Recall that the Stanley-Reisner ideal of a pure simplicial complex is equidimensional and so it suffices to prove that Δ_J is pure.

There exists a bijection mapping each facet $U \in \Delta_J$ of size k to a facet $\widetilde{U} \in \Delta_J$ of size $k + N$. Therefore, since $\Delta(f_1, \ldots, f_r)$ is pure, so is Δ_J. □

We now establish condition (3). First we show the equality $\dim \mathcal{J}(\overline{V}) = \dim J$.

Lemma 4.3.5. *The equality holds:*

$$\dim \mathcal{J}(\overline{V}) = \dim J$$

Proof. Note that $\dim \mathcal{J}(\overline{V})$ is given in terms of the affine variety V by the equalities

$$\dim \mathcal{J}(\overline{V}) = N + \dim \mathcal{J}(V) = N + \dim(V). \tag{4.5}$$

Since J is the Stanley-Reisner ideal of $\Delta_{\langle \mathrm{in}_<(f_1^h), \ldots, \mathrm{in}_<(f_r^h) \rangle}$, we have that

$$\begin{aligned}
\dim J &= \dim \Delta_{\langle \mathrm{in}_<(f_1^h), \ldots, \mathrm{in}_<(f_r^h) \rangle} + 1 \\
&= N + \dim \Delta(f_1, \ldots, f_r) + 1 && ([43, \text{Lemma } 2.10]) \\
&= N + \operatorname{rank} \Delta(f_1, \ldots, f_r) \\
&= N + \dim(V) && (\text{by Lemma } 4.3.3) \\
&= \dim \mathcal{J}(\overline{V}) && (\text{by } (4.5))
\end{aligned}$$

This completes the proof. □

Finally, we show that the degrees of the ideals $\mathcal{J}(\overline{V})$ and J are equal.

Lemma 4.3.6. *The equality holds:*

$$\deg \mathcal{J}(\overline{V}) = \deg J$$

Proof. The variety $\mathcal{V}(J) \subset \mathbb{P}^N$ is a union of linear subspaces indexed by facets,

$$\mathcal{V}(J) = \bigcup_{\substack{F \in \Delta_{\langle \mathrm{in}_<(f_1^h), \ldots, \mathrm{in}_<(f_r^h)\rangle} \\ F \text{ a facet}}} L_F.$$

Since J is generated by square-free monomials, it is a radical ideal and defines a subvariety of $\mathbb{P}^{2N-1}$, of degree

$$\deg(J) = \#\text{facets of } \Delta_J = \#\mathcal{B},$$

where $\mathcal{B}$ denotes the set of bases in the algebraic matroid of V.

Let $I := \mathcal{J}(\overline{V})$ and $\mathrm{in}_<(I) = \langle \mathrm{in}_<(f^h) \mid f \in I \rangle$ be the initial ideal of $\mathcal{J}(\overline{V})$. Note that the $\deg \mathrm{in}_<(I) = \deg I$ and $\dim \mathrm{in}_<(I) = \dim I = \dim J$ by Lemma 4.3.5. Then, since $J \subseteq \mathrm{in}_<(I)$, we have that

$$\deg I = \deg \mathrm{in}_<(I) \leq \deg J = \#\mathcal{B} \leq \sum_{U \in \mathcal{B}} \#\overline{\pi}_U^{-1}(p_U),$$

where each $p_U \in \mathbb{P}^U$ is a generic point in the codomain of the projection π_U. Each term $\#\overline{\pi}_U^{-1}(p_U)$ in the sum above is a *multidegree* of the variety $\overline{V} = \mathcal{V}(\mathcal{J}(\overline{V}))$. A result of van der Waerden [102] expresses the standard degree $\deg \mathcal{J}(\overline{V})$ as the sum of multidegrees,

$$\sum_{U \in \mathcal{B}} \#\overline{\pi}_U^{-1}(p_U) = \deg(\mathcal{J}(\overline{V})).$$

(See also [50, Proposition 1.7.3], for a more general statement.) Thus, we have

$$\deg I \leq \deg J \leq \sum_{U \in \mathcal{B}} \#\overline{\pi}_U^{-1}(p_U) = \deg I$$

and so $\deg \mathcal{J}(\overline{V}) = \deg J$ as desired □

Thus, condition (3) is also satisfied, and $f_1^h, \ldots, f_r^h$ form a Gröbner basis of $\mathcal{J}(\overline{V})$ with respect to the monomial order $<$.

4.4 Multiview Ideals

In this section, we define the multiview ideal and the family of polynomials known as $k-$focals, which will form a universal Gröbner basis of the multiview ideal. We begin with a brief overview of the necessary background from the literature on computer vision. For further details, we refer the reader to [1, 2].

The standard model of a pinhole camera is a surjective linear map $\mathbb{P}^3 \dashrightarrow \mathbb{P}^2$ which is represented up to scale by a 3×4 matrix A of full rank. The camera maps a *world point* $q \in \mathbb{P}^3$ to its *image point* $Aq \in \mathbb{P}^2$.

A configuration of n cameras, represented by the matrices $A_1, \dots, A_n$, is *generic* if all 4×4 minors of the matrix $\begin{bmatrix} A_1^\top & \dots & A_n^\top \end{bmatrix}$ are nonzero. We fix a generic configuration of cameras $(A_1, \dots, A_n)$. The *multiview variety* of $(A_1, \dots, A_n)$ is the closure in $(\mathbb{P}^2)^n$ of the image of the rational map,

$$\begin{aligned} \mathbb{P}^3 &\dashrightarrow (\mathbb{P}^2)^n \\ q &\mapsto (A_1 q, \dots, A_n q). \end{aligned}$$

This is an irreducible variety containing the n-tuples of images of world points $q \in \mathbb{P}^3$. Its vanishing ideal is the multiview ideal associated to cameras $(A_1, \dots, A_n)$. Equivalently, this is the vanishing ideal of the affine cone over the multiview variety. Although the multiview ideals for n generic cameras will vary with the cameras, the combinatorics governing such ideals will depend only on n—thus, abusing notation, we let I_n denote the multiview ideal associated to n generic cameras. To mirror the setting of [43], we work primarily with the affine cone over the multiview variety, which we denote by M_n.

For a subset $\sigma \subset [n]$ of size k, and $x_i = (x_{i1}, x_{i2}, x_{i3})^\top$, a $k-$focal is a maximal minor of the $3k \times (4+k)$ matrix

$$\begin{bmatrix} A_{\sigma_1} & x_1 & 0 & \cdots & 0 \\ A_{\sigma_2} & 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{\sigma_k} & 0 & 0 & \cdots & x_k \end{bmatrix}. \tag{4.6}$$

The work [2] proves that the set of all $2-, 3-$, and $4-$focals forms a universal Gröbner basis of I_n. We recover this result using Theorem 4.2.3.

Note that, since $3k \geq 4 + k$ for $k \geq 2$, a maximal minor is determined by a choice of $4 + k$ rows of the above matrix. Thus, each $k-$focal is uniquely determined by a choice of k cameras and a choice of $4 + k$ coordinates from the 3-tuples $x_1, \dots, x_k$. It will be convenient

to consider a coarse classification of focals which we call the *profile*. For any polynomial $f \in \mathbb{C}[x_1, \dots, x_n]$, the profile of f is an $n-$dimensional vector $\text{prof}(f)$ with each coordinate corresponding to the index of a camera and each entry being the number of coordinates appearing as a variable in f, from the corresponding x_i. For example, consider the 2-focal

$$f = \det \begin{bmatrix} A_1 & x_1 & 0 \\ A_2 & 0 & x_2 \end{bmatrix}.$$

We see that $\text{prof}(f) = (3, 3, 0 \dots, 0)$ all three coordinates of x_1 and x_2 appear as variables in the determinant. Note that if f is a $k-$focal and any entry of $\text{prof}(f)$ is equal to 1, then f is a monomial multiple of some focal of lower degree. For example, consider the 3-focal g given by the minor

$$g = \det \begin{bmatrix} A_1 & x_1 & 0 & 0 \\ A_2 & 0 & x_2 & 0 \\ A_{31} & 0 & 0 & x_{31} \end{bmatrix}.$$

Then $\text{prof}(g) = (3, 3, 1, 0 \dots, 0)$ and by Laplace expansion we see that g is $x_{31} \cdot f$. Since any monomial multiple of a focal is a redundant generator of the multiview ideal, we only consider focals f where $\text{prof}(f)$ has no entry equal to 1.

4.4.1 The simplicial complex

For $n \geq 4$, define Δ_n to be the simplicial complex whose nonfaces are generated by the spreads of the $2-, 3-, 4-$focals of a generic camera configuration $(A_1, \dots, A_n)$. We determine the dimension and number of facets of Δ_n and show that each facet has one of two profiles up to permutation of cameras.

Note that each face $U \in \Delta_n$ corresponds to the support of some square-free monomial $x^U = \prod_{ij \in U} x_{ij} \in \mathbb{C}[x_1, \dots, x_n]$. Thus, for $U \in \Delta_n$ we define its profile as $\text{prof}(U) = \text{prof}(x^U)$. The following results show that if $U \in \Delta_n$ is a facet, it has one of two profiles up to permutation of cameras.

First, consider the case where $n = 4$. Up to permutation of cameras, any $2-, 3-$, or $4-$focal has profile $(3, 3, 0, 0), (3, 2, 2, 0)$, or $(2, 2, 2, 2)$. To each $k-$focal f we associate a *spread monomial*: the product of all variable x_{ij} in the spread of f. The facets of the simplicial complex Δ_4 correspond to all square-free monomials in the 12 variables

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{31}, x_{32}, x_{33}, x_{41}, x_{42}, x_{43}$$

with maximal support that are not divisible by any spread monomial. For example, one such monomial is

$$x_{11}x_{12}x_{13}x_{21}x_{22}x_{31}x_{41}$$

which has profile $(3,2,1,1)$. Another possibility is the monomial

$$x_{11}x_{12}x_{21}x_{22}x_{31}x_{32}x_{41}$$

which has profile $(2,2,2,1)$. We show that up to permutation, any facet of Δ_4 has profile $(3,2,1,1)$ or $(2,2,2,1)$.

Lemma 4.4.1. *The complex Δ_4 is pure, 6-dimensional and has 648 facets whose profiles are obtained from all permutations of $(3,2,1,1)$ and $(2,2,2,1)$, and choices of image variables from each camera.*

Proof. Recall that the Stanley-Reisner ideal of Δ_4 is generated by all square-free monomials $\prod_{ij\in W} x_{ij}$ where W is a *nonface* of Δ_4. The nonfaces of Δ_4 are generated by the spreads of all $2-, 3-$, and $4-$focals and each of these focals has profile $(3,3,0,0), (3,2,2,0)$ or $(2,2,2,2)$ up to permutation of cameras. Therefore, the Stanley-Reisner ideal of Δ_4 is generated by monomials with profile $(3,3,0,0), (3,2,2,0)$ or $(2,2,2,2)$ up to permutation. Moreover, any focal is uniquely determined by a choice of k cameras and $4+k$ coordinates x_{ij} and so for any permutation and choice of $\mathbf{x}$ coordinates that could lead to a square-free monomial with profile $(3,3,0,0), (3,2,2,0)$ or $(2,2,2,2)$, it is simple to construct a corresponding $k-$focal. Thus, the Stanley-Reisner ideal of Δ_4 is generated by all square-free monomials with profile $(3,3,0,0), (3,2,2,0)$ or $(2,2,2,2)$ up to permutation.

The profile of each facet U of Δ_4 is a 4-dimensional vector $\text{prof}(U)$ with entries in $\{0,1,2,3\}$. Up to permutation of cameras, $\text{prof}(U)$ may be written in lexicographic order. Since the monomial corresponding to U must not be divisible by any monomial in the Stanley-Reisner ideal, $\text{prof}(U)$ can have at most one entry equal to 3. Once the first entry is chosen to be 3, $\text{prof}(U)$ must be the largest vector below $(3,2,2,0)$ in lexicographic order. It follows that $\text{prof}(U) = (3,2,1,1)$. If the first entry of $\text{prof}(U)$ is 2, then $\text{prof}(U)$ is the largest vector below $(2,2,2,2)$ in the lexicographic order and so $\text{prof}(U) = (2,2,2,1)$. It follows that any facet of Δ_4 has profile $(3,2,1,1)$ or $(2,2,2,1)$. On the other hand, any square-free monomial with profile $(3,2,1,1)$ or $(2,2,2,1)$ is not divisible by any monomial in the Stanley-Reisner ideal of Δ_4 and thus corresponds to a facet of the complex. Thus, the facets of Δ_4 are in bijection with the set of all monomials with profile $(3,2,1,1)$ of $(2,2,2,1)$ up to permutation. In particular, Δ_4 is a pure 6-dimensional simplicial complex.

By the above discussion, all facets of Δ_4 have profile $(3,2,1,1)$ or $(2,2,2,1)$. For facets with profile $(3,2,1,1)$ there are 4 ways to choose the camera that contributes 3 variables and then 3 ways to choose the one that contributes 2 variables which is a total of 12. The cameras that contribute 2 or 1 variables have 3^3 ways to pick the variables. Therefore, there is a total of $12 \times 27 = 324$ such facets. By a similar argument, for the facet $(2,2,2,1)$, there are 4 ways to choose the camera that contributes 1 variable and 3^4 ways for the various cameras to choose the variables they want to contribute, making another 324 possibilities. Together, we have 648 facets, each of dimension 6. □

We now consider the general case $n \geq 4$. Again, we show that up to permutation of cameras the facets of Δ_n have profile $(3,2,1,1,\dots,1)$ or $(2,2,2,1,\dots,1)$.

Theorem 4.4.2. *The simplicial complex Δ_n for $n \geq 4$ is pure, $(n+2)-$dimensional and has only two profiles of facets up to permutation of cameras and choice of variables in each camera plane. There are $n(n-1)3^{n-1}$ facets with profile $(3,2,1,1,\dots,1)$ and $\binom{n}{3}3^n$ facets with profile $(2,2,2,1,\dots,1)$.*

Proof. Following the same argument as in Lemma 4.4.1, the Stanley-Reisner ideal of Δ_n is generated by all square-free monomials with profile one of

$$(3,3,0\dots,0),(3,2,2,0,\dots,0) \text{ or } (2,2,2,2,0,\dots,0)$$

up to permutation of cameras.

Choose a facet $U \in \Delta_n$. As in Lemma 4.4.1, U corresponds to a maxmimally supported square-free monomial that is not divisible by any monomial in the Stanley-Reisner ideal. Up to permutation of cameras, we may assume that $\text{prof}(U)$ is in lexicographic order. In particular, the first entry of $\text{prof}(U)$ is either 2 or 3.

The non-faces of Δ_n are generated by the spreads of all $2-, 3-, 4-$focals and so, as in Lemma 4.4.1, the Stanley-Reisner ideal of Δ_n is generated by all monomials corresponding to vectors of the form $(3,3,0,0,0,\dots),(3,2,2,0,0,\dots)$ or $(2,2,2,2,0\dots)$ up to permutation of cameras. If the first entry of $\text{prof}(U)$ is 3, then $\text{prof}(U) = (3,2,1,1,\dots,1)$ since this is the maximal element below $(3,2,2,0,\dots,0)$ in the lexicographic order. If the first entry of $\text{prof}(U)$ is 2, then $\text{prof}(U) = (2,2,2,1,1\dots,1)$ since this is the maximal element below $(2,2,2,2,0,\dots,0)$ in the lexicographic order. Thus, every facet of Δ_n has profile $(3,2,1,1,\dots,1)$ or $(2,2,2,1,\dots,1)$. Note that the sum of entries in each of these vectors is $n+3$ and so the simplicial complex Δ_n is pure and $(n+2)-$dimensional. Moreover, any

square-free monomial with profile $(3,2,1,1\dots,1)$ or $(2,2,2,1,\dots,1)$ up to permutation of cameras is not divisible by any element of the Stanley-Reisner ideal. It follows that the facets of Δ_n are in bijective correspondence with all square-free monomials with profile $(3,2,1,1\dots,1)$ or $(2,2,2,1,\dots,1)$ up to permutation.

We now count the number of facets. For facets with profile $(3,2,1,1,\dots,1)$, there are n choices of camera contributing 3 variables and $(n-1)$ choices of camera contributing 2 variables, and the remaining cameras each contribute 1 variable. There are 3 choices of variables for each camera contributing either 1 or 2 variables and so in total there are $n(n-1)3^{n-1}$ facets with profile $(3,2,1,1,\dots,1)$. For facets with profile $(2,2,2,1,\dots,1)$, there are $\binom{n}{3}$ choices of cameras contributing 2 variables and the remaining cameras each contribute one variable. For each of the n cameras there are 3 choices of variables and so in total there are $\binom{n}{3}3^n$ facets with profile $(2,2,2,1,\dots,1)$. □

The simplicial complex Δ_n is in fact the independence complex of a matroid, a result that will follow from Lemma 4.3.3 and the next subsection. Our next result answers the question, *what kind of matroid?*

Theorem 4.4.3. *The simplicial complex Δ_n is a transversal matroid of rank $n+3$, isomorphic to the union of the uniform matroid $\mathcal{U}_{3,3n}$ on all x-variables and the direct sum $\mathcal{U}_{1,n}^{\oplus n}$ over the n subsets $\{x_{i1},x_{i2},x_{i3}\}$.*

Proof. Let U be any facet of Δ_n. Whether U has profile $(3,2,1,1,\dots,1)$ or $(2,2,2,1,\dots,1)$, it must must contain at least one element from $\{x_{i1},x_{i2},x_{i3}\}$ for each $i=1,\dots,n$. In other words, U must contain a basis B of the direct sum $\mathcal{U}_{1,n}^{\oplus n}$. Now, since the remaining 3 elements of $U\setminus B$ form a basis of $\mathcal{U}_{3,3n}$, we have that U is a basis in $\mathcal{U}_{3,3n}\vee\mathcal{U}_{1,n}^{\oplus n}$. A similar proof shows that any face of Δ_n is a union of independent sets in $\mathcal{U}_{3,3n}$ and $\mathcal{U}_{1,n}^{\oplus n}$. The class of transversal matroids is closed under union ([76, Corollary 12.3.8]), completing the proof. □

The structure of Δ_n as a transversal matroid can be seen by explicitly constructing a bipartite graph $G=(S\sqcup T,E)$ with $\Delta_n\cong\mathcal{M}[G,S]$, according to the following rule: set vertices $S=\{x_{11},\dots,x_{n3}\}$, $T=\{a,b,c\}\sqcup\{v_1,\dots,v_n\}$, and edges

$$E=\bigcup_{\substack{1\le i\le n\\1\le j\le 3}}\{x_{ij}v_i,x_{ij}a,x_{ij}b,x_{ij}c\}.$$

In matroid theory, it is well-known that every transversal matroid is representable over a sufficiently large (finite) field (cf. [76, Ch. 6].) On the other hand, the matroids Δ_n are not

realizable over arbitrary fields. For example, starting from Δ_4, we may delete the variables x_{21}, x_{31}, x_{41} and contract the variables $x_{11}, x_{12}, x_{22}, x_{32}, x_{42}$ to obtain the uniform matroid $\mathcal{U}_{2,4}$ on the remaining variables $x_{13}, \dots, x_{43}$ as a minor. This implies that Δ_n for $n \geq 4$ is not representable over the field with two elements. In particular, the matroid Δ_n is non-graphic [76, Chapter 5].

4.4.2 Applying the Huang-Larson theorem

Let $M_n \subset \mathbb{C}^{3n}$ denote the affine cone over the multiview variety from n generic cameras and let I_n denote its vanishing ideal. By Lemma 4.2.4, it is sufficient to prove that for each facet $U \in \Delta_n$ the projection $\pi_U : M_n \to \mathbb{C}^U$ is dominant.

First, we prove that it suffices to check that the projection is dominant for a small number of facets. The structure of the simplicial complex Δ_n is highly symmetric. There is a natural permutation action of the symmetric group $\mathfrak{S}_n$ on both the simplicial complex and the polynomial ring $\mathbb{C}[x_1, \dots, x_n]$ that corresponds to the permutation of cameras. For each permutation $\tau \in \mathfrak{S}_n$ and face $U = \{x_{ij}\}$ of Δ_n, we write τU to denote the face $\{x_{\tau(i)j}\}$ of Δ_n. Note that by Theorem 4.4.2, if U is a facet of Δ_n then so is τU for each permutation $\tau \in \mathfrak{S}_n$ and any facet $W \in \Delta_n$ can be written as τU for some $\tau \in \mathfrak{S}_n$ and some facet U with profile $(3, 2, 1, 1, \dots, 1)$ or $(2, 2, 2, 1, \dots, 1)$.

The symmetric group $\mathfrak{S}_n$ acts on the polynomial ring $\mathbb{C}[x_1, \dots, x_n]$ in the same way and we see that the $k-$focals of the camera configuration $(A_1, \dots, A_n)$ exhibit similar symmetries. For each polynomial $f \in \mathbb{C}[x_1, \dots, x_n]$, we write τf to denote the polynomial achieved by replacing x_{ij} with $x_{\tau(i)j}$. Note that if f is a $k-$focal arising as a maximal minor from the subset $\sigma \subset [n]$, then τf is a $k-$focal arising as a maximal minor from the subset $\tau(\sigma) \subset [n]$.

We now use these symmetries to reduce the number of computations needed to prove that $2-, 3-, 4-$focals form a universal Gröbner basis of the multiview ideal.

Lemma 4.4.4. *It suffices to prove that the map* (4.1) *is dominant for facets with profile* $(3, 2, 1, 1, \dots, 1)$ *and* $(2, 2, 2, 1, \dots, 1)$.

Proof. By Theorem 4.4.2, each facet of Δ_n corresponds to some permutation of either $(3, 2, 1, 1 \dots, 1)$ or $(2, 2, 2, 1, \dots, 1)$. Thus, if W is a facet of Δ_n there exists a facet U with profile $(3, 2, 1, 1, \dots, 1)$ or $(2, 2, 2, 1 \dots, 1)$ and a permutation $\tau \in \mathfrak{S}_n$ such that $W = \tau U$.

By Lemma 4.2.4 it suffices to work with π_U. Suppose that the projection π_U is dominant for all facets U with profile $(3, 2, 1, 1, \dots, 1)$ and $(2, 2, 2, 1, \dots, 1)$. Let $W = \tau U$. Choose g in the intersection $I_n \cap \mathbb{C}[x_{ij} : ij \in W]$ where I_n is the multiview ideal. It follows that

we can write $g = \sum_\alpha h_\alpha f_\alpha$ where each f_α is a $2-, 3-$,or $4-$focal. Note that $\tau^{-1} \circ g = \tau^{-1} \circ \sum_\alpha h_\alpha f_\alpha \in I_n \cap \mathbb{C}[x_{ij} : ij \in U]$ since $\tau^{-1} f_\alpha$ is still a $2-, 3-, 4-$focal for each focal f_α and each permutation τ. Since we assume that π_U is dominant, it follows from Lemma 4.3.2 that $g = 0$. Thus, it suffices to show that π_U is dominant when U is a facet with profile $(3, 2, 1, 1, \dots, 1)$ or $(2, 2, 2, 1, \dots, 1)$, as desired. □

We can now prove that the set of all $2-, 3-$ and $4-$focals forms a universal Gröbner basis of the multiview ideal I_n.

Proof of Theorem 4.1.1. By Lemmas 4.2.4 and 4.4.4, it is sufficient to check that (4.1) is a dominant map for facets with profile $(3, 2, 1, 1, 1, \dots, 1)$ and $(2, 2, 2, 1, 1, \dots, 1)$.

Assume the statement holds for up to $n-1$ cameras. Now suppose we have n cameras. Without loss of generality, we assume that the nth camera contributes one index to a facet $U \in \Delta_n$. Suppose this variable is x_{nj}. Choose a generic point $p = (p_1, \dots, p_{n-1}, p_n)$ in the affine space $\mathbb{C}^U$, where p_i corresponds to an entry in U for camera i. Since p is generic, then so is $q = (p_1, \dots, p_{n-1})$, so that by induction there is a generic point $a = (a_1, \dots, a_{n-1}) \in M_{n-1}$ that projects to q under π_W where W is obtained from U by dropping the last entry.

Since $a \in M_{n-1}$ there is some point in $\mathbb{P}^3$ that is imaged to a_i by camera A_i for $1 \leq i \leq n-1$. Let $x(a) \in \mathbb{C}^4$ be a nonzero point for which $A_i x(a)$ and a_i are scalar multiples for all $i = 1, \dots, n-1$. Since $(A_1, \dots, A_n)$ is generic, note that each row of A_n must be nonzero. Furthermore, since a is generic, we may assume the j^{th} coordinate of $A_n x(a)$ is some nonzero scalar c. By construction,

$$(p_1, \dots, p_{n-1}, c) = \left(\pi_W(a), e_j^T A_n x(a)\right) \in \pi_U(M_n)$$

Since M_n is a cone in each factor, it follows that

$$(p_1, \dots, p_{n-1}, p_n) = (p_1, \dots, p_{n-1}, (p_n/c) \cdot c) \in \pi_U(M_n),$$

completing the proof that π_U is dominant.

We verified that the statement holds in the base case of $n = 4$ by a direct computation in Macaulay2. Thus, by induction on the number of cameras, the set of $2-, 3-, 4-$focals forms a universal Gröbner basis of the multiview ideal I_n for any $n \geq 4$. □

4.5 The universal multiview ideal

For n unknown cameras, the universal multiview variety consists of all points

$$(A_1, \dots, A_n, p_1, \dots, p_n) \in (\mathbb{P}^{3\times 4})^n \times (\mathbb{P}^2)^n$$

such that for each $i = 1, \dots, n$ there is some world point $q \in \mathbb{P}^3$ such that $A_i q = p_i$. Much like in the previous section, the universal multiview variety is irreducible, as it is the closed image of a rational map $(\mathbb{P}^{11})^n \times \mathbb{P}^3 \dashrightarrow (\mathbb{P}^{11} \times \mathbb{P}^2)^n$. Let $\tilde{M}_n$ be the affine cone over the universal multiview variety and $\tilde{I}_n$ the vanishing ideal of $\tilde{M}_n$. Note that the difference between M_n and $\tilde{M}_n$ is that the former lies in $(\mathbb{P}^2)^n$ and its vanishing ideal I_n lies in $\mathbb{C}[x_1, \dots, x_n]$ where x_i consists of the three variables corresponding to the ith camera plane, while $\tilde{M}_n$ lies in $(\mathbb{P}^{3\times 4})^n \times (\mathbb{P}^2)^n$ and its vanishing ideal $\tilde{I}_n$ lies in a polynomial ring $\mathbb{C}[A_1, \dots, A_n, x_1, \dots, x_n]$ with $12n$ variables coming from the n cameras $A_1, \dots, A_n$ and $3n$ variables of the image plane coordinates $x_1, \dots, x_n$.

In this section, we use the Huang-Larson theorem to show that the set of all $2-, 3-, 4-$focals of n unknown camera matrices and image points in the polynomial ring $\mathbb{C}[\mathbf{A}, \mathbf{x}]$ form a universal Gröbner basis for the universal multiview ideal $\tilde{I}_n$. It was shown in [1, Theorem 3.2] that the $2-, 3-, 4-$focals form a Gröbner basis for $\tilde{I}_n$ under certain product orders. Thus, in particular, these focals generate $\tilde{I}_n$. Our result that the $2-, 3-, 4-$focals form a universal Gröbner basis of $\tilde{I}_n$ strengthens [1, Theorem 3.2], and is a new result.

4.5.1 The simplicial complex

For $n \geq 4$, let $\tilde{D}_n$ be the simplicial complex whose nonfaces are generated by the spreads of $2-, 3-, 4-$focals in the polynomial ring $\mathbb{C}[\mathbf{A}, \mathbf{x}]$. Recall that a $k-$focal is a maximal minor of the matrix (4.6). Both Δ_n and $\tilde{D}_n$ have nonfaces generated by the spreads of $2-, 3-, 4-$focals, except that in $\tilde{D}_n$ (and $\tilde{I}_n$), the entries of $\mathbf{A}$ are variables. Regardless, we expect a relationship between the facets and nonfaces of Δ_n and $\tilde{D}_n$. We describe a simple process for generating all facets of $\tilde{D}_n$ from the facets of Δ_n.

Identify the ground set of $\tilde{D}_n$ with the union of the variables in $\mathbf{A}$ and $\mathbf{x}$ coming from all n cameras – a total of $15n = 12n + 3n$ elements.The spread of a polynomial $f \in \mathbb{C}[\mathbf{A}, \mathbf{x}]$ is identified with the set of $\mathbf{A}$ and $\mathbf{x}$ variables present in f while profiles, as before, will record the number of $\mathbf{x}$ variables from the n cameras. The following observations will help us understand $\tilde{D}_n$.

Lemma 4.5.1. *If $U \in \tilde{D}_n$ is a facet, then the following three conditions hold:*

1. *If $x_{ij} \notin U$ then $a_{ijk} \in U$ for all values of $k = 1, 2, 3, 4$.*
2. *There are exactly $n+3$ variables $x_{ij} \in U$ such that $a_{ijk} \in U$ for all values of $k = 1, 2, 3, 4$. Moreover, these $\mathbf{x}-$variables form a facet of Δ_n.*
3. *If U contains more than $n + 3$ $\mathbf{x}-$variables then (2) holds and for each remaining $x_{ij} \in U$ there is exactly one value of k such that $a_{ijk} \notin U$.*

It follows that $\tilde{D}_n$ is a pure $(13n + 2)-$dimensional simplicial complex.

Proof. The nonfaces of $\tilde{D}_n$ are generated by the spreads of $2-, 3-$, and $4-$focals with profiles $(3, 3, 0, 0, \dots, 0)$, $(3, 2, 2, 0 \dots, 0)$ and $(2, 2, 2, 2, \dots, 0)$ up to permutation of cameras. Any such focal, f, is the determinant of a submatrix of (4.6) with $4 + k$ chosen rows. It will be convenient to think of f as a polynomial in all the $\mathbf{x}$-variables from the chosen rows with (symbolic) coefficients the 4×4 minors of the camera rows appearing in these chosen rows. Note that each camera variable from the chosen rows appears in at least one 4×4 determinant and hence, in the spread of f. This means that

1. if $x_{ij} \in \text{spr}f$, then $a_{ijk} \in \text{spr}f$ for all $k = 1, 2, 3, 4$, and
2. if $x_{ij} \notin \text{spr}f$, then $a_{ijk} \notin \text{spr}f$ for all $k = 1, 2, 3, 4$.

From this, we conclude that if $U \in \tilde{D}_n$ then the collection of variables $x_{ij} \in U$ such that $a_{ijk} \in U$ for all $k = 1, 2, 3, 4$ must form a face of Δ_n. Otherwise, U contains the spread of some $k-$focal. In particular, U has at most $n + 3$ such $\mathbf{x}-$variables.

We now consider the faces of $\tilde{D}_n$. Choose $U \in \tilde{D}_n$ and let $U_{\mathbf{x}}$ be the collection of $\mathbf{x}-$variables in U. There are two cases to consider: either $U_{\mathbf{x}} \in \Delta_n$ or $U_{\mathbf{x}} \notin \Delta_n$.

Suppose that $U_{\mathbf{x}} \in \Delta_n$. Then, U does not contain the $\mathbf{x}-$spread of a $2-, 3-$, or $4-$focal, and the same is true for $U \cup \{\mathbf{A}\}$. It follows that $U \cup \{\mathbf{A}\} \in \tilde{D}_n$. Moreover, if U is a facet of $\tilde{D}_n$ then it must be that $U_{\mathbf{x}}$ is a facet of Δ_n. In particular, there are $n+3$ variables $x_{ij} \in U$, as well as all the $\mathbf{A}$ variables, and (1) and (2) hold.

Now, suppose that $U_{\mathbf{x}} \notin \Delta_n$. Assume the $U_{\mathbf{x}}$ contains $n + 3$ variables. Recall that any facet of Δ_n also contains $n + 3$ variables and the profile of each facet is nonzero in every entry. It follows that $\text{prof}(U_{\mathbf{x}})$ has at least one entry equal to 0 since any other valid way to arrange $n + 3$ $\mathbf{x}-$variables leads to the profile of a facet. So, there exists some index i such that $x_{ij} \notin U$ for all choices of j. Consider $W = U \cup \{x_{i1}\}$. Since U does not contain the spread

of any $k-$focal, if W contains the spread of the $k-$focal f then the i^{th} entry of prof(f) must be equal to 1. However, this implies that f is a monomial multiple of a generating focal. So, since U does not contain the spread of a focal, neither does W and therefore $W \in \tilde{D}_n$. It follows that U is not a facet of $\tilde{D}_n$. Using the same argument, if $U_{\mathbf{x}}$ contains fewer then $n+3$ variables then U is not a facet of $\tilde{D}_n$. We conclude that if $U \in \tilde{D}_n$ is a facet, then $U_{\mathbf{x}}$ contains at least $n+4$ variables.

Assume $U_{\mathbf{x}}$ contains at least $n+4$ variables. As discussed above, the variables $x_{ij} \in U_{\mathbf{x}}$ such that $a_{ijk} \in U$ for all $k = 1, 2, 3, 4$ form a face of Δ_n. For the remaining $\mathbf{x}-$variables, there is at least one value of k such that $a_{ijk} \notin U$. If U is a facet, and thus maximally supported, there are exactly $n+3$ variables in $U_{\mathbf{x}}$ such that $a_{ijk} \in U$ for all values $k = 1, 2, 3, 4$ and these $\mathbf{x}-$variables form a facet of Δ_n. Moreover, there is at exactly one value of k such that $a_{ijk} \notin U$ for each remaining $\mathbf{x}-$variable. For each $x_{ij} \notin U$ we have that $a_{ijk} \in U$ for all values of $k = 1, 2, 3, 4$ again because U is maximally supported.

Given each of the properties (1), (2), and (3), a counting argument shows that any facet of $\tilde{D}_n$ contains $13n+3$ variables. It follows that $\tilde{D}_n$ is a pure $(13n+2)-$dimensional simplicial complex. □

From the proof of Lemma 4.5.1 we see that if U is a facet of Δ_n then the union $U \cup \{\mathbf{A}\}$ of U with all variables corresponding to entries in all n camera matrices is a facet of $\tilde{D}_n$. Moreover, every facet of $\tilde{D}_n$ has the same size. We now show that all facets of $\tilde{D}_n$ can be found using a simple recursive process.

Theorem 4.5.2. *Every facet of $\tilde{D}_n$ can be found using the following process:*

1. *Choose a facet $U \in \Delta_n$ and set $U_0 := U \cup \{\mathbf{A}\}$.*
2. *For $\ell \geq 1$, choose $x_{ij} \notin U_{\ell-1}$ and set $U_\ell := (U_{\ell-1} \cup \{x_{ij}\}) \setminus \{a_{ijk}\}$ for some a_{ijk}.*

Proof. By the previous discussion, U_0 is a facet of $\tilde{D}_n$ for each choice of U. Moreover, each U_ℓ is a face of $\tilde{D}_n$. Since $\tilde{D}_n$ is pure and U_ℓ is the same size as U_0, we see that U_ℓ is also facet. Thus, we only need to show that the process is exhaustive.

Choose some facet $W \in \tilde{D}_n$. By Lemma 4.5.1, if $x_{ij} \notin W$, then $a_{ijk} \in W$ for all values of k and there are exactly $n+3$ variables $x_{i_1 j_1}, \ldots, x_{i_{n+3} j_{n+3}} \in W$ with $a_{ijk} \in W$ for all k. Set $U = \{x_{i_1 j_1}, \ldots, x_{i_{n+3} j_{n+3}}\}$ and $U_0 = U \cup \{\mathbf{A}\}$. If W contains exactly $n+3$ of the $\mathbf{x}-$variables, then $U_0 = W$. Suppose then that W contains $n+3+m$ of the $\mathbf{x}-$variables with $m > 0$. Then, for each $x_{i_{n+3+\ell} j_{n+3+\ell}} \in W$ there exists some value of $k_\ell = 1, 2, 3, 4$ such that $a_{ijk_\ell} \notin W$. So, setting $U_\ell = U_{\ell-1} \cup \{x_{i_{n+3+\ell} j_{n+3+\ell}}\} \setminus \{a_{ijk_\ell}\}$, we see that $W = U_m$. Thus, we have constructed W using the desired process. □

Analogously to Theorem 4.4.3, we now describe the matroid structure of $\widetilde{\Delta}_n$.

Theorem 4.5.3. *The simplicial complex $\widetilde{\Delta}_n$ is a transversal matroid of rank $13n+3$, isomorphic to the union of the uniform matroid $\mathcal{U}_{3,15n}$ on all variables and the direct sum $\mathcal{U}_{13,15}^{\oplus n}$ over the n subsets $\{x_{i1}, x_{i2}, x_{i3}, a_{11}, \dots, a_{34}\}$.*

Proof. Note first that every facet $U_0 = U \cup \{\mathbf{A}\}$ appearing in Theorem 4.5.2 is a basis of the described matroid union. Indeed, Theorem 4.4.3 implies $U = U' \cup U''$, where U' is a basis of $\mathcal{U}_{3,3n}$ and U'' is a basis of $\mathcal{U}_{1,3}^{\oplus n}$. Furthermore, since $U'' \cup \{\mathbf{A}\}$ is a basis of $\mathcal{U}_{13,15}^{\oplus n}$ disjoint from U', we deduce that U_0 is a basis of $\mathcal{U}_{3,15n} \vee \mathcal{U}_{13,15}^{\oplus n}$.

Next, observe that any of the facets U_ℓ constructed in Theorem 4.5.2 is a basis in the matroid union $\mathcal{U}_{3,15n} \vee \mathcal{U}_{13,15}^{\oplus n}$. For example, we have

$$U_1 = \big(U_0 \cup \{x_{ij}\}\big) \setminus \{a_{ijk}\} = U' \cup \big(U'' \cup \{\mathbf{A}\} \setminus \{a_{ijk}\} \cup \{x_{ij}\}\big)\,,$$

with U' and U'' as in the preceding paragraph, and similarly for $\ell \geq 2$. Finally, combining the observations of Lemma 4.5.1 and Theorem 4.5.2, we deduce that all facets of the union $\mathcal{U}_{3,15n} \vee \mathcal{U}_{13,15}^{\oplus n}$ have the form U_ℓ for some choice of U_0 and ℓ. □

4.5.2 Applying the Huang-Larson theorem

Our goal now is to prove that the $2-, 3-, 4-$focals form a universal Gröbner basis of $\tilde{I}_n$. By Lemma 4.2.4, it is sufficient to show that for each facet $U \in \tilde{D}_n$ the projection $\pi_U : \tilde{M}_n \to \mathbb{C}^U$ is dominant. Since the number of facets of $\tilde{D}_n$ is large, even in the case where $n = 4$, it is necessary to reduce the dominance checks to a limited number of facets. To this end, we show that we need only consider facets of $\tilde{D}_n$ corresponding to facets of Δ_n followed by a symmetry reduction analogous to Lemma 4.4.4.

Lemma 4.5.4. *It suffices to prove that π_U is dominant for all facets U of the form $U = W \cup \{\mathbf{A}\}$ where W is a facet of Δ_n.*

Proof. Assume that for some facet U' of $\tilde{D}_n$ the projection $\pi_{U'}$ is dominant. Let $U = (U' \cup \{x_{ij}\}) \setminus \{a_{ijk}\}$ for some values of i, j, k.

Choose a generic point $Q \in \mathbb{C}^U$. We will argue that there is a point $P \in \tilde{M}_n$ such that $\pi_U(P) = Q$. We consider a generic point $Q' \in \mathbb{C}^{U'}$ obtained from Q by deleting the entry in position x_{ij} and adding an entry in position a_{ijk}. By assumption, there is a point $P' \in \tilde{M}_n$ such that $\pi_{U'}(P') = Q'$. The projection $\pi_U(P')$ agrees with Q in all positions except the one indexed by x_{ij}. Note that the entry of P' in position a_{ijk} is irrelevant for the projection

π_U since U does not contain a_{ijk}. We will use this flexibility to construct a $P \in \tilde{M}_n$, from P', such that $\pi_U(P) = Q$.

Suppose $P' = (A'_1, \dots, A'_n, p'_1, \dots, p'_n)$ and q_{ij} is the entry of Q in position x_{ij}. We modify p'_i in the jth position to obtain a preimage for Q in $\tilde{M}_n$ under π_U. Since $P' \in \tilde{M}_n$, there is some y and scalars λ'_ℓ such that $A'_\ell y = \lambda'_\ell p'_\ell$ for $1 \leq \ell \leq n$. Modify P' to P by replacing A'_i with A_i as follows: replace the jk entry of A'_i with $A_{ijk} = \frac{a'_{ijk} y_k + \lambda'_i (q_{ij} - p'_{ij})}{y_k}$. Then, $(A_i y)_j = \lambda'_i q_{ij}$ and so $\pi_U(P) = Q$. It follows that $\pi_U : \tilde{M}_n \to \mathbb{C}^U$ is dominant.

By Theorem 4.5.2, we now need only check that the projections π_U are dominant for facets of the form $U = W \cup \{\mathbf{A}\}$ for some facet W of Δ_n, as desired. □

The result of Lemma 4.5.4 greatly reduces the number of facets that must be considered to check the hypotheses of Theorem 4.2.3. We can further reduce the number of facets to be checked using symmetry, as in Lemma 4.4.4.

Lemma 4.5.5. *In Lemma 4.5.4, we can further restrict to facets W of Δ_n with profile $(3, 2, 1, 1, \dots, 1)$ and $(2, 2, 2, 1, \dots, 1)$.*

Proof. This follows from the same argument as in Lemma 4.4.4 □

We can now prove that the set of all $2-, 3-$, and $4-$focals forms a universal Gröbner basis of the universal multiview ideal $\tilde{I}_n$.

Proof of Theorem 4.1.2 By Lemmas 4.5.4 and 4.5.5, it suffices to check that (4.1) is a dominant map for facets of $\tilde{D}_n$ of the form $U = W \cup \{\mathbf{A}\}$ where $W \in \Delta_n$ is a facet with profile $(3, 2, 1, 1, \dots, 1)$ or $(2, 2, 2, 1, \dots, 1)$.

Suppose the statement holds for $\tilde{I}_{n-1}$ and we now have n cameras. This n^{th} camera contributes $12n$ camera variables, and without loss of generality, one variable x_{nj} to the facet $U \in \tilde{D}_n$. Choose a generic point $Q = (B_1, \dots, B_n, q_1, \dots, q_n) \in \mathbb{C}^U$. Since Q is generic, so is the point $Q' = (B_1, \dots, B_{n-1}, q_1, \dots, q_{n-1}) \in \mathbb{C}^{U'}$ where U' is attained from U by removing each variable corresponding to the n^{th} camera. Note that U' is also of the form $U' = W' \cup \{\mathbf{A}_{n-1}\}$ for a facet $W' \in \Delta_{n-1}$. By induction, there exists $(B_1, \dots, B_{n-1}, p_1, \dots, p_{n-1}) \in \tilde{M}_{n-1}$ that projects to Q' under $\pi_{U'}$. We can assume that the cameras are exactly $B_1, \dots, B_{n-1}$ since U' contains all camera variables. In particular, there exists $y \in \mathbb{C}^4$ and scalars λ_i such that for $1 \leq i \leq n-1$, we have that $B_i y = \lambda_i p_i$. Since B_n, y are generic, we may assume that the j^{th} coordinate of $B_n y = c \neq 0$. By construction,

$$(B_1, \dots, B_{n-1}, B_n, q_1, \dots, q_{n-1}, c) \in \pi_U(\tilde{M}_n).$$

Then, since $\tilde{M}_n$ is a cone in each factor,

$$Q = (B_1, \dots, B_n, q_1, \dots, q_n) = (B_1, \dots, B_n, q_1, \dots, (q_n/c) \cdot c) \in \pi_U(\tilde{M}_n),$$

completing the proof that π_U is dominant.

We verified that the statement holds in the base case of $n = 4$ by a direct computation in Macaulay2. Thus, by induction on the number of cameras, the set of $2-, 3-, 4-$focals forms a universal Gröbner basis of the universal multiview ideal $\tilde{I}_n$ for any $n \geq 4$. □

4.6 Conclusion

In conclusion, this chapter demonstrates that Theorem 4.2.3 is a useful tool for proving universal Gröbner basis results in the emerging field of *algebraic vision* [47]. Specifically, we have obtained a new proof of the previously-known result Theorem 4.1.1, and an entirely new result in Theorem 4.1.2. We are optimistic that the techniques developed here can be adapted to other problems; in particular, to families of ideals appearing in [1], such as the resectioning varieties studied in [23].

Chapter 5

VISUALIZING DATA WITH TREE-SNE

This chapter is the content of the paper [46]. The clustering and visualisation of high-dimensional data is a ubiquitous task in modern data science. Popular techniques include nonlinear dimensionality reduction methods like t-SNE or UMAP. These methods face the 'scale-problem' of clustering: when dealing with the MNIST dataset, do we want to distinguish different digits or do we want to distinguish different ways of writing the digits? The answer is task dependent and depends on scale. We revisit an idea of Robinson & Pierce-Hoffman that exploits an underlying scaling symmetry in t-SNE to replace 2-dimensional with (2+1)-dimensional embeddings where the additional parameter accounts for scale. This gives rise to the t-SNE tree (short: tree-SNE). We prove that the optimal embedding depends continuously on the scaling parameter for all initial conditions outside a set of measure 0: the tree-SNE tree exists. This idea conceivably extends to other attraction-repulsion methods and is illustrated on several examples.

5.1 Introduction and Results

5.1.1 The problem of clustering.

From fields such as computer vision to computational linguistics, large sets of high dimensional data are increasingly commonplace. Thus, it is necessary to develop tools that can be used for their analysis and visualisation. While traditional linear techniques are often remarkably powerful, the visualisation problem poses a new type of challenge. Even if data in 1000 dimensions is intrinsically close to a 5-dimensional submanifold (the manifold hypothesis), embedding five dimensions into two or even three dimensions poses a considerable challenge. One popular dimensionality reduction technique is t-distributed stochastic neighbour embedding (t-SNE), first introduced by van der Maarten and Hinton in [62]. As a completely non-linear method, t-SNE does not try to produce an isometric or low-distortion embedding. Instead, it tries to group the data into clusters. This works remarkably well in practice and has been very widely used. Indeed, it has become apparent [13] that a broad range of methods (including Laplacian eigenmaps [9], ForceAtlas2 [44], UMAP [65, 51] and

t-SNE [62]) can be seen as special cases of a general attraction-repulsion method: we start with points in $\mathbb{R}^2$ that represent the high-dimensional data and then induce each point to move towards other points in $\mathbb{R}^2$ that represent 'similar' high-dimensional data. Simultaneously, points are repelled by all of the other data points. If the underlying data is highly clustered, then the same cluster structure in $\mathbb{R}^2$ is a steady state of that dynamical system, see [58, 6, 16] for details.

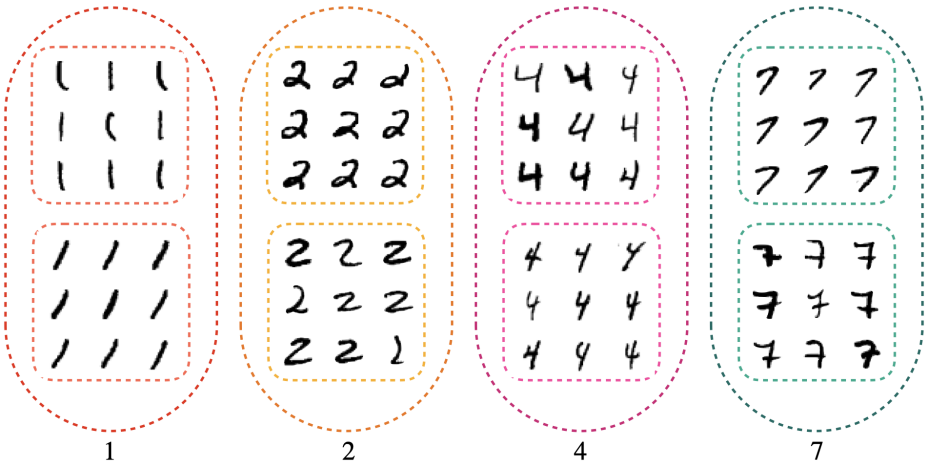


Figure 5.1: A schematic of subclusters in the MNIST dataset.

5.1.2 The problem of scale

Since t-SNE embeddings exhibit cluster structures that are representative of those seen in the high-dimensional data, a typical use case for t-SNE is producing a 2D visualisation that aids with identification of clusters within a dataset. However, the question of scale is often overlooked in this clustering process. Although often treated as a static problem, clustering may be viewed as a dynamic problem with a hierarchical structure. As a motivating example, consider the MNIST dataset which contains 70,000 images of handwritten numerical digits. Typically, these images are clustered based on what digits they represent. However, this clustering is not helpful if we are, for example, trying to distinguish different handwriting styles within the dataset. In this case, a more granular clustering is needed. Since many digits can be written in multiple ways, we expect there to be subclusters corresponding to different handwriting styles within the clusters corresponding to each digit. Figure 5.1 gives a schematic of example clusters and subclusters within the MNIST dataset.

One solution to the problem of scale are *hierarchical clustering* methods, see [74, 75] for an overview. These methods are often computationally expensive and so an alternative method is desirable. It has been observed that variants of t-SNE using kernels with heavier tails tend to produce clusters with finer granularity, see for example [52]. However, there is currently no good way to choose a kernel that corresponds to a particular granularity outside of trial and error. Thus, the problem of scale persists within t-SNE embeddings and clusterings. Note that this is not exclusive to t-SNE and is a problem with *all* clustering methods yet is often ignored in practice.

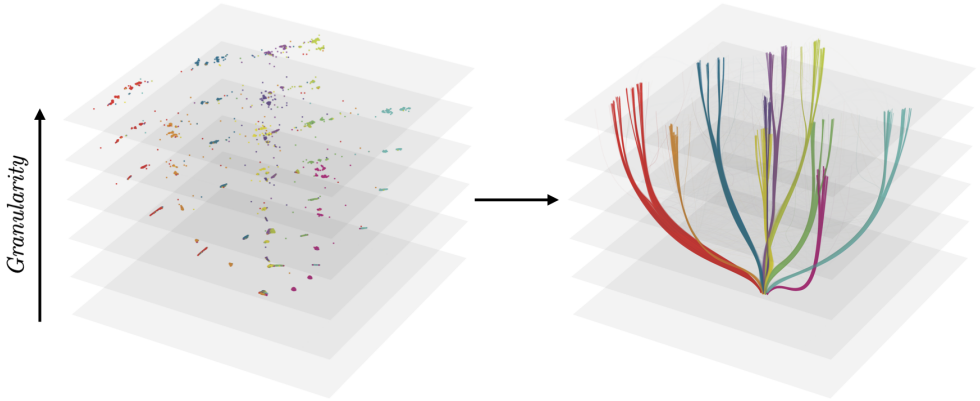


Figure 5.2: Slices of the MNIST tree-SNE embedding are interpolated to create a tree. Each color corresponds to a digit.

5.1.3 Tree-SNE

To address the problem of scale, we explore a hierarchical variant of t-SNE known as tree-SNE that was introduced in [83]. The idea is to exploit a one-parameter family [52] of kernels that produce t-SNE-type embeddings with finer and finer granularity. We begin with generating an ordinary t-SNE embedding of the data. Then, additional layers are generated by changing the parameter in the kernel and optimizing a certain loss function. Each embedding forms a layer of the tree-SNE embedding and the points in each layer can be interpolated to plot smooth trajectories of each of the data points. Figure 5.2 contains slices of a tree-SNE embedding of the MNIST dataset, along with the tree that comes from interpolating each layer.

We see in Figure 5.2 that each of the main branches of the tree splits into smaller branches as the layers increase. These correspond to different ways of writing the same

digit. In Figure 5.3, we show example branches and subbranches that correspond to the clusters and subclusters in Figure 5.1. Each branch is labelled with the average image of elements belonging to the cluster.

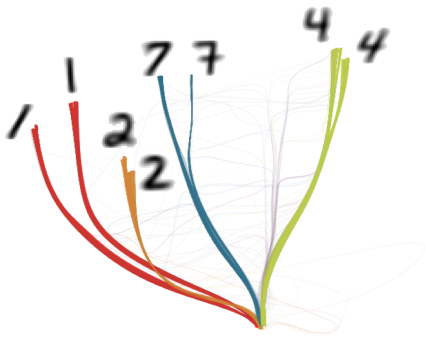


Figure 5.3: Subbranches of the MNIST tree correspond to different handwriting styles. Images are the average of elements in the corresponding cluster.

In particular, we see that as the layers of the embedding increase there are clusters that reveal structure with finer granularity and there is a continuous structure to the tree: the points follow a continuous trajectory and do not jump around from layer to layer. As has been previously observed in [52], t-SNE embeddings generated with heavier tailed kernels tend to produce finer clusters and so it is expected that clusters at the top of the tree-SNE embedding are more granular than those at the bottom. However, the continuous structure of the tree is non-trivial and will be the main focus of this chapter.

5.1.4 Main Result

Our main result concerns the continuous structure of tree-SNE embeddings, as observed in Figure 5.2. We prove that this structure is not coincidental and is a consequence of the tree-SNE algorithm's design. Recall that each layer of a tree-SNE embedding is generated using a kernel from a one parameter family. We show that as the parameter varies continuously, so does the embedding.

Theorem 5.1.1: (The Tree-SNE Tree exists)

Suppose $(y_1, \dots, y_n) \in \mathbb{R}^{nd}$ *is a* t-SNE *embedding in* $\mathbb{R}^d$ *with parameter* α*. Under mild conditions, for each* α' *in some neighbourhood of* α *there is a* t-SNE *embedding* $(y'_1, \dots, y'_n) \in \mathbb{R}^{nd}$ *in a neighbourhood of* $(y_1, \dots, y_n)$*.*

To be more specific, we view the point $(\alpha, y_1, \dots, y_n) \in \mathbb{R}^{1+nd}$ as belonging to the zero set of a function F. The function F is the gradient of the loss function minimized to find t-SNE embeddings when the parameter α is not treated as a constant. The 'mild conditions' referred to in our theorem ensure that F is smooth and generically has a constant rank. This will allow us to conclude that the zero set of F is generically a smooth manifold of a dimension $1 + d(d+1)/2$ around $(\alpha, y_1, \dots, y_n)$, with one of these dimensions accounting for changes in the parameter α. These 'mild conditions' are necessary but are not restrictive. Indeed, they are satisfied by generic datasets and initializations, see Section 5.3.

Organisation.

In Section 5.2, we recount the necessary background information on t-SNE and describe the tree-SNE algorithm. In Section 5.3, we prove our main theorem to show that the tree-SNE algorithm indeed produces tree-like embeddings for generic datasets and initialisations. Finally, in Section 5.4 we use tree-SNE to visualise two different datasets from the HathiTrust library and examine structures revealed in the embeddings.

5.2 Preliminaries

In this section, we give a brief overview of both t-SNE and tree-SNE. Throughout, we fix a dataset $\{x_1, \dots, x_n\} \subset \mathbb{R}^D$, where one may think of the dimension D as large. We begin by describing dimensionality reduction and data visualisation with t-SNE.

5.2.1 T-SNE

Given a high-dimensional dataset $\{x_1, \dots, x_n\} \subset \mathbb{R}^D$, the central goal of t-SNE is to find a low-dimensional set of points $\{y_1, \dots, y_n\} \subset \mathbb{R}^d$ that preserves the overall structure of the data. In particular, we aim to find an embedding such that if x_i and x_j are "close" in $\mathbb{R}^D$, then y_i and y_j are "close" in $\mathbb{R}^d$ with high probability. Typically, the dimension d is chosen such that $d = 2$ or $d = 3$ as this enables us to use the t-SNE embedding $\{y_1, \dots, y_n\}$ as a proxy for visualizing the high-dimensional dataset $\{x_1, \dots, x_n\}$.

The t-SNE method consists of two stages. First, *affinities* p_{ij} are determined between each pair of points x_i and x_j in $\mathbb{R}^D$. The affinity p_{ij} between the points x_i and x_j is a probability that represents the similarity of x_i and x_j within the dataset: a high affinity is associated to pairs that are close together, whereas low affinities are assigned to pairs of points that are far away from each other. The affinity p_{ij} between x_i and x_j is determined

using the pair of *directional affinities* $p_{j|i}$ and $p_{i|j}$. For each $j \neq i$, the directional affinity $p_{j|i}$ is defined as

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/\sigma_i^2)},$$

where the variance σ_i^2 is a user determined parameter. The directional affinity between x_i and itself is set to $p_{i|i} = 0$ for all i. Note that for each value of i, the directional affinities determine a probability distribution on $\{x_1, \dots, x_n\}$. The variance σ_i^2 is chosen such that the *perplexity* of this distribution, defined

$$\mathcal{P} = \exp\left(-\log 2 \cdot \sum_{j\neq i} p_{j|i}\log_2 p_{j|i}\right),$$

has some pre-specified value. Typical values of the perplexity $\mathcal{P}$ are between 5 and 50. For each pair i, j the affinity p_{ij} is the normalised average of $p_{j|i}$ and $p_{i|j}$,

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$

Once affinities between all pairs of points x_i and x_j have been determined, a low dimensional embedding of the data $\{y_1, \dots, y_n\} \subset \mathbb{R}^d$ is found. For each pair of points $y_i, y_j \in \mathbb{R}^d$ the affinity q_{ij} is defined

$$q_{ij} = \frac{K(\|y_i - y_j\|)}{\sum_{k\neq\ell} K(\|y_k - y_\ell\|)}$$

where K is a kernel. Traditionally, the kernel K corresponds to the t-distribution with one degree of freedom, $K(d) = (1 + d^2)^{-1}$. However, more general t-distributions can be considered. As in [52], we consider kernels given by

$$K_\alpha(d) = \frac{1}{(1 + d^{2/\alpha})^\alpha}$$

where $\alpha > 0$ is a parameter. This corresponds to a scaled t-distribution with $2\alpha - 1$ degrees of freedom. Note that choosing $\alpha = 1$ yields the standard t-distribution kernel and the taking the limit as $\alpha \to \infty$ gives the Gaussian kernel $K_\infty(d) = \exp(-d^2)$. This corresponds to the SNE method of Hinton and Roweis [42]. Moreover, $\alpha < 0.5$ corresponds to a t-distribution with negative degrees of freedom and so represents a distribution with heavier

tails than any (non-scaled) t-distribution. Note that when $\alpha < 0.5$ the resulting distribution is no longer a probability distribution. A common misconception about the t-SNE method is that the kernel must correspond to a probability distribution, however this is not correct. Since we only consider a finite number of points, distributions may always be renormalized and it is not, for instance, necessary for any particular integral to be finite. As in the high-dimensional space, the affinity between y_i and itself is set as $q_{ii} = 0$. The embedding $\{y_1, \dots, y_n\}$ is chosen so that the Kullbeck-Leibler (KL) divergence

$$\mathcal{L}(y_1, \dots, y_n) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

is minimized. The gradient of $\mathcal{L}$ is given by

$$\frac{\partial \mathcal{L}}{\partial y_i} = \sum_{j=1}^{n} (p_{ij} - q_{ij}) K_\alpha(\|y_i - y_j\|)^{1/\alpha} (y_i - y_j)$$

and so an embedding that minimizes the loss function may be found using a random initialisation and applying first order methods such as gradient descent.

5.2.2 Tree-SNE

We now describe how tree-SNE builds on t-SNE to generate collections of embeddings that reveal finer substructures within data. The central idea of tree-SNE is that iteratively decreasing the degrees of freedom in the t-distribution kernel yields embeddings that capture the structure of data in increasing granularity. This sequence of embeddings can then be stacked atop one another and interpolated to reveal a tree-like structure with different branches corresponding to different subclusters in the data. Tree-SNE takes in a sequence of parameters $\{a^{(i)}\}_{i=1}^m$ and produces a sequence of embeddings $\{(y_1^{(i)}, \dots, y_n^{(i)})\}_{i=1}^m$. As with t-SNE, each embedding is produced in two stages. First, affinities between each pair x_i, x_j of data points in $\mathbb{R}^D$ are calculated so that the directional affinities at each point have some perplexity $\mathcal{P}^{(i)}$ that is dependent on $\alpha^{(i)}$. Then, the embedding $(y_1^{(i)}, \dots, y_n^{(i)})$ is a minimizer of the loss function $\mathcal{L}$. Each embedding $(y_1^{(i)}, \dots, y_n^{(n)})$ forms a layer of the tree-SNE embedding. The parameters $\{\alpha^{(i)}\}_{i=1}^m$ are chosen such that $\alpha^{(1)} = 1$ and $\alpha^{(i+1)} < \alpha^{(i)}$ for all i. Similarly, the perplexities $\mathcal{P}^{(i)}$ decrease as the layer of the embedding increases. Thus, as the layers of the tree-SNE embedding increase, the t-distribution has heavier and heavier tails and this leads to finer structures of the data being revealed.

Typically, t-SNE embeddings are found using a random initialisation. However, since

the loss function $\mathcal{L}$ is highly non-convex, different random initialisations lead to different embeddings. To ensure that the tree-SNE embedding has a continuous structure from one layer to the next, we use the embedding $\{y_1^{(i)}, \dots, y_n^{(i)}\}$ as the initialisation for layer $i+1$. Assuming that the difference between $\alpha^{(i)}$ and $\alpha^{(i+1)}$ is small, we expect that if $\{y_1^{(i)}, \dots, y_n^{(i)}\}$ minimizes $\mathcal{L}$ with parameter $\alpha^{(i)}$ then there should be a minimizer for $\mathcal{L}$ with parameter $\alpha^{(i+1)}$ in some small neighbourhood of $\{y_1^{(i)}, \dots, y_n^{(i)}\}$. We prove that this is the case in Section 5.3. Note that the random initialisation of the t-SNE embedding in the first layer may serve as an intrinsic sanity check: since different initialisations produce different trees, any structures that persist across multiple initialisations is likely to be trustworthy. This idea is used in the work [39].

5.3 Proof of Continuous Structure

In this section, we prove that, under mild assumptions, the tree-SNE method described in Section 5.2.2 produces a sequence of embeddings with a continuous structure. Although tree-SNE is a discrete process with a finite set of parameters producing a finite set of embeddings, we consider a continuous version of the problem and treat the embedding $y_1(\alpha), \dots, y_n(\alpha)$ as a function of the parameter α which may take any value in $(0, 1]$. Similarly, the perplexity $\mathcal{P}(\alpha)$ is also a function of the parameter α. Our goal is to prove that each point $y_i(\alpha)$ is a continuous function.

We consider the map $F : \mathbb{R}^{1+nd} \to \mathbb{R}^{nd}$ defined by

$$F(\alpha, y_1, \dots, y_n) = \left(\frac{\partial \mathcal{L}}{\partial y_1}, \dots, \frac{\partial \mathcal{L}}{\partial y_n} \right).$$

Since the embedding $y_1(\alpha), \dots, y_n(\alpha)$ is a minimizer of $\mathcal{L}$, in particular the condition $\nabla \mathcal{L} = 0$ is satisfied and so every embedding lies on the zero-set of F. As the function $\mathcal{L}$ is not convex, it is not true that every point in the zero-set of F corresponds to a minimizer of $\mathcal{L}$. However, given an appropriate choice of parameters such as step size, first order methods like gradient descent return the point in the zero-set of F closest to their initialisation. We restate a refined version of our main result in the following theorem.

Theorem 5.3.1 (The Tree-SNE tree exists - refined)**.** *Suppose $(\alpha, y_1, \dots, y_n)$ is a generic point in the zero set of F. Then, the zero set of F is a smooth, properly embedded, $1 + d(d+1)/2$ dimensional submanifold of $\mathbb{R}^{1+nd}$ $(\alpha, y_1, \dots, y_n)$. Moreover, in this neighbourhood the parameter α is not constant.*

We argue that this is sufficient to prove the contuous structure of a tree-SNE embedding. Indeed, fix a parameter α and suppose that $(y_1, \dots, y_n)$ is the corresponding layer of the tree-SNE embedding. The result of Theorem 5.3.1 tells us that for each α' in some neighbourhood of α there exists a set of points $(y'_1, \dots, y'_n)$ in a neighbourhood of $(y_1, \dots, y_n)$ so that $(\alpha', y'_1, \dots, y'_n)$ is in the zero set of F. So, by initializing the α' layer of the tree-SNE embedding at $(y_1, \dots, y_n)$, gradient descent will return the embedding $(y'_1, \dots, y'_n)$, or some other set of points in the neighbourhood of $(y_1, \dots, y_n)$. It follows that, when the parameters $\{a^{(i)}\}_{i=1}^m$ are chosen appropriately and each layer is initialized using the preceding layer, the tree-SNE embedding will have a continuous structure. The main tool that we use is the *constant-rank level-set theorem*. This is a standard result in the theory of manifolds. For details, see for example [55].

Theorem 5.3.2 (Constant-Rank Level Set)**.** *Let M and N be smooth manifolds, and $\phi : M \to N$ a smooth map with constant rank r in some neighbourhood of p. Each level set of ϕ is locally a properly embedded submanifold of codimension r in M.*

To prove Theorem 5.3.1, we will show that F is smooth and generically has rank $nd - d(d+1)/2$. Then, using the constant rank level-set theorem, we show that at a generic point, the zero set of F is locally a properly embedded submanifold of $\mathbb{R}^{1+nd}$ of dimension $1 + d(d+1)/2$ where α is not constant. First, we fix the following set of assumptions.

Assumption 5.3.3. We assume the following conditions are satisfied.

(A) The dataset $\{x_1, \dots, x_n\} \subset \mathbb{R}^D$ is fixed.

(B) The perplexity $\mathcal{P}(\alpha)$ is a smooth function of α

(C) Given the affinities p_{ij} calculated using perplexity $\mathcal{P}(\alpha)$, there is at least one set of points $y_1, \dots, y_n \in \mathbb{R}^d$ such that $\nabla^2 \mathcal{L}$ has rank $nd - d(d+1)/2$.

Assumptions (A) and (B) will ensure that the map F is smooth, whereas Assumption (C) will ensure that F generically has rank $nd - d(d+1)/2$. This is the key to ensuring that the constant-rank level set theorem applies in our setting.

5.3.1 A Discussion of Assumption (C)

Note that since the loss function $\mathcal{L}$ is defined using pairwise distances, it is invariant under all rigid transformations of $\mathbb{R}^d$. It follows that the rank of $\nabla \mathcal{L}^2$ is at most $nd - d(d+1)/2$ since the space of all rigid transformations is $d(d+1)/2$ dimensional. Since this is the only

invariance exhibited by the function, we expect the Hessian $\nabla\mathcal{L}^2$ to achieve this rank at any generic embedding $y_1, \dots, y_n$. While Assumption (C) is generically satisfied, it is indeed possible for the Hessian $\nabla^2\mathcal{L}$ to have rank lower than $nd - d(d+1)/2$, as shown in the following proposition.

Proposition (Rank Deficient Hessian)**.** *Let $n = 3$ and $d = 2$. Suppose that the set $\{y_1, y_2, y_3\} \subset \mathbb{R}^2$ is the vertex set of an equilateral triangle with side length r. Then, the rank of $\nabla^2\mathcal{L}$ is at most $nd - d(d+1)/2 - 1$.*

Proof. Recall that the rank of $\nabla^2\mathcal{L}$ is at most $nd - d(d+1)/2$ since $\mathcal{L}$ is invariant under all rigid transformations of $\mathbb{R}^d$. Note that any additional invariance of $\mathcal{L}$ causes the rank of the Hessian $\nabla^2\mathcal{L}$ to drop. Since the set $\{y_1, y_2, y_3\} \subset \mathbb{R}^2$ is the vertex set of an equilateral triangle, in particular we have that for each choice of $i \neq j$, the distance between y_i and y_j is $\|y_i - j_y\| = r$. Note that each of the probabilities q_{ij} is given by

$$q_{ij} = \frac{K(r)}{\sum_{k\neq\ell} K(r)} = \frac{1}{6}$$

and so is independent of the side length r. It follows that $\mathcal{L}(y_1, y_2, y_3)$ is invariant under scalings of the side length r. This is not a rigid transformation of $\mathbb{R}^2$ and so the rank of $\nabla^2\mathcal{L}$ is at most $nd - d(d+1)/2 - 1$. □

Although Assumption (C) only requires that the desired rank be achieved for one set of points, if this condition is satisfied, the Hessian at any generic embedding $y_1, \dots, y_n$ will have rank $nd - d(d+1)/2$ due to the lower semicontinuity of rank. This the content of the following lemma.

Lemma 5.3.4. *Suppose Assumption (C) is satisfied. Then, for any generic set of points $y_1, \dots, y_n$ the Hessian $\nabla^2\mathcal{L}$ has rank $nd - d(d+1)/2$.*

Proof. For any set of points $y_1, \dots, y_n$, the rank of $\nabla^2\mathcal{L}$ is at most $r = nd - d(d+1)/2$ since $\mathcal{L}$ is invariant under all rigid transformations of $\mathbb{R}^d$. Suppose the rank is equal to r at $y_1, \dots, y_n$. Then, there exists an $r \times r$ minor $m(y_1, \dots, y_n)$ of $\nabla^2\mathcal{L}$ that is non-zero at $y_1, \dots, y_n$. Note that $\mathcal{L}$ is analytic and so this minor is also analytic. Thus, the zero set of $m(y_1, \dots, y_n)$ is measure zero. Moreover, the zero set of all $r \times r$ minors of $\nabla^2\mathcal{L}$ is measure zero and so a generic embedding does not land in this set. Since the rank function is lower semicontinuous, $\operatorname{rank}\nabla^2\mathcal{L}$ may only increase in a neighbourhood of $y_1, \dots, y_n$. If the rank of the Hessian $\nabla^2\mathcal{L}$ is equal to $nd - d(d+1)/2$ at $y_1, \dots, y_n$, it follows that $\operatorname{rank}\nabla^2\mathcal{L}$ is

constant in a neighbourhood of $y_1, \dots, y_n$. It follows that at a generic set of points, the rank of $\nabla^2 \mathcal{L}$ is exactly $nd - d(d+1)/2$. □

Note that if Assumption (C) is not satisfied, then it will be satisfied for a random small perturbation of the affinities. This corresponds to a small perturbation of the perplexity of each distribution. Since t-SNE is known to be robust to changes in perplexity [62], this does not have a significant effect on the embedding.

Lemma 5.3.5. *Suppose for a set of affinities p_{ij} that Assumption (C) is not satisfied. Then, replacing p_{ij} with $p_{ij} + \epsilon_{ij}$ with ϵ_{ij} sampled i.i.d from $N(0, \sigma^2)$, satisfies Assumption (C) with probability 1.*

Proof. Since p_{ij} is not a function of the embedding $y_1, \dots, y_n$, each entry of $\nabla^2 \mathcal{L}$ is linear in p_{ij}. Note that $\nabla^2 \mathcal{L}$ has rank strictly less than $nd - d(d+1)/2$ if and only if all of its $nd - d(d+1)/2$ minors are equal to 0. The set of all p_{ij} satisfying these polynomial constraints is measure zero. Thus, any random perturbation of the affinities will not land in this set with probability 1. □

5.3.2 The Tree-SNE Tree Exists.

We now show that, in the setting of Assumption 5.3.3, the map $F : \mathbb{R}^{1+nd} \to \mathbb{R}^{nd}$ is smooth and has constant rank $nd - d(d+1)/2$ in some neighbourhood of a generic point $(\alpha, y_1, \dots, y_n) \in \mathbb{R}^{1+nd}$.

Lemma 5.3.6. *The map $F : \mathbb{R}^{1+nd} \to \mathbb{R}^{nd}$ is smooth and has constant rank $nd - d(d+1)/2$ around a generic point $(\alpha, y_1, \dots, y_n)$.*

Proof. Recall that the partial derivative of $\mathcal{L}$ with respect to y_i is given by

$$\frac{\partial \mathcal{L}}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) K_\alpha(\|y_i - y_j\|)^{1/\alpha} (y_i - y_j).$$

Assuming that $y_i \neq y_j$ for all $i \neq j$, note that q_{ij} and $K_\alpha(\|y_i - y_j\|)$ are smooth functions of $y_1, \dots, y_n$. It follows that $\frac{\partial \mathcal{L}}{\partial y_i}$ is a smooth function of $y_1, \dots, y_n$. We now justify that $\frac{\partial \mathcal{L}}{\partial y_i}$ is a smooth function of α. Assuming that $\alpha \neq 0$, it is clear that q_{ij} and $K_\alpha(\|y_i - y_j\|)^{1/\alpha}$ are smooth functions of α. Thus, we need only justify that p_{ij} is a smooth function of α. Recall that the affinities p_{ij} are smooth functions of the bandwidths σ_i chosen to assure that the distribution around each data point has the perplexity $\mathcal{P}(\alpha)$. Note that it is sufficient

to prove that the bandwidths σ_i are smooth functions of α. By assumption, the perplexity $\mathcal{P}(\alpha)$ is a smooth function of α. Consider the function $G : \mathbb{R}^{1+n} \to \mathbb{R}^n$ defined by

$$G(\alpha, \sigma_1, \dots, \sigma_n) = \left[\mathcal{P}(\alpha) - \exp(-\log 2 \sum_{j \neq i} p_{j|i} \log_2 p_{j|i})\right]_{i=1}^{n}$$

where we view each directional affinity $p_{j|i}$ as a smooth function of the bandwidth σ_i. The set of parameters and bandwidths used to find the tree-SNE embeddings are in the zero set of G. It is clear the G is a smooth function. Suppose that for a given α^*, $\sigma_1^*, \dots, \sigma_n^*$, the equality $G(\alpha^*, \sigma_1^*, \dots, \sigma_n^*) = 0$ holds. Then, by the implicit function theorem, $\sigma_1, \dots, \sigma_n$ are a continuous function of α in a neighbourhood of $(\alpha^*, \sigma_1^*, \dots, \sigma_n^*)$ if the matrix

$$J = \left[\frac{\partial}{\partial \sigma_k}(\mathcal{P}(\alpha) - \exp(-\log 2 \sum_{j \neq i} p_{j|i} \log_2 p_{j|i}))\right]_{i,k=1}^{n}$$

is invertible. Note that J is diagonal and so is not invertible whenever the equality

$$\frac{\partial}{\partial \sigma_i}(\mathcal{P}(\alpha) - \exp(-\log 2 \sum_{j \neq i} p_{j|i} \log_2 p_{j|i})) = 0$$

for some value of i. Generically, the equality does not hold and so the implicit function theorem applies. It follows that F is a smooth function as desired. We now show that F has constant rank $nd - d(d+1)/2$ in some neighbourhood of a generic point $(\alpha, y_1, \dots, y_n)$. Recall that the rank of F is the rank of its Jacobian ∇F. The Jacobian ∇F is given by

$$\nabla F = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial y_1 \partial \alpha} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial y_n \partial \alpha} \\ & \nabla^2 \mathcal{L} & \end{bmatrix}$$

and so the rank of F is at least the rank of the Hessian $\nabla \mathcal{L}^2$. Note that ∇F is an $(nd+1) \times nd$ matrix and so its rank cannot exceed nd. However, like $\mathcal{L}$, the function F is invariant under all rigid transformations of $\mathbb{R}^d$. The space of all rigid transformations of $\mathbb{R}^d$ is $d(d+1)/2-$dimensional and so it follows that $\operatorname{rank} \nabla F \leq nd - d(d+1)/2$. By assumption, $\nabla^2 \mathcal{L}$ generically has rank $nd - d(d+1)/2$ and so we must have that $\operatorname{rank} \nabla F = nd - d(d+1)/2$. By the lower semicontinuity of rank, F has constant rank of $nd - d(d+1)/2$ in a neighbourhood of a generic point, as desired. □

We now apply the constant-rank level-set theorem to prove our main theorem and conclude that tree-SNE embeddings have continuous structure.

Proof of Theorem 5.3.1. Choose a generic point $(\alpha, y_1, \dots, y_n)$ in the zero set of F.

Then, as shown in Lemma 5.3.6, the hypothesis of the constant-rank level-set theorem is satisfied. In particular, we have that the zero-set of F around $(\alpha, y_1, \dots, y_n)$ is a properly embedded, $1 + d(d+1)/2$ dimensional submanifold of $\mathbb{R}^{1+nd}$. It remains to argue that the parameter α is not constant in this neighbourhood. Indeed, since F is invariant under all rigid transformations of $\mathbb{R}^d$, any level set of F is at least $d(d+1)/2$ dimensional. Suppose that α is constant in this neighbourhood. Then, there exists some nontrivial vector $(v_1, \dots, v_n)^\top \in \mathbb{R}^{nd}$ such that $(\alpha, y_1 + v_1, \dots, y_n + v_n) = 0$. However, we would then have that $(v_1, \dots, v_n)^\top$ is in the nullspace of $\nabla^2 \mathcal{L}$. Since $\nabla^2 \mathcal{L}$ has rank $nd - d(d+1)/2$ by assumption, this is a contradiction. It follows that α is not constant, as desired. □

5.4 Examples

We used tree-SNE to visualize two datasets constructed from the HathiTrust Library. The full HathiTrust dataset consists of 13.6 million works described by millions of features that correspond to word counts on each page of the work. We use the preprocessed dataset from [90] that reduces each work to a 100-dimensional vector. We generated embeddings using the default parameters described in [83]. Each layer of the embeddings was found using FIt-SNE [59] with default parameters. Code to reproduce these examples, as well as the MNIST example seen in Section 5.1, can be found at `https://github.com/j4ck-k/tree-sne`.

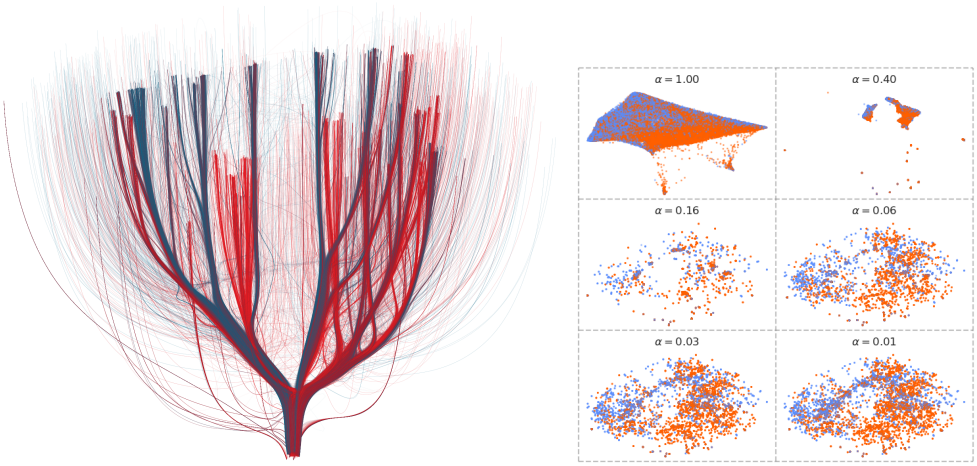


Figure 5.4: Tree-SNE embedding of 100,000 works of American and English literature. Red points correspond to English literature and blue points correspond to American literature.

5.4.1 English and American Literature

For our first example, we use tree-SNE to visualize a set of 100,000 works of English and American literature from the HathiTrust library dataset. This dataset was constructed from the full HathiTrust dataset by filetering works classified as either PR (the Library of Congress code for English literature) and PS (the Library of Congress code for American literature) and taking the first 100,000 works for computational simplicity and efficiency. Our dataset is approximately 55% English and 45% American literature.

We generates a 50 layer embedding with degrees of freedom between 1 and 0.01. Figure 5.4 shows the tree-SNE embedding of the dataset along with slices of several layers. Points are color coded depending on Library of Congress code. In each plot, red points are works classified as PR and blue points are works classified as PS. When $\alpha = 1$ we see that, although the two classifications are being pushed to opposite sides of the embedding, there is no separation between works of English and American literature. Clusters begin to appear in subsequent layers of the plot.

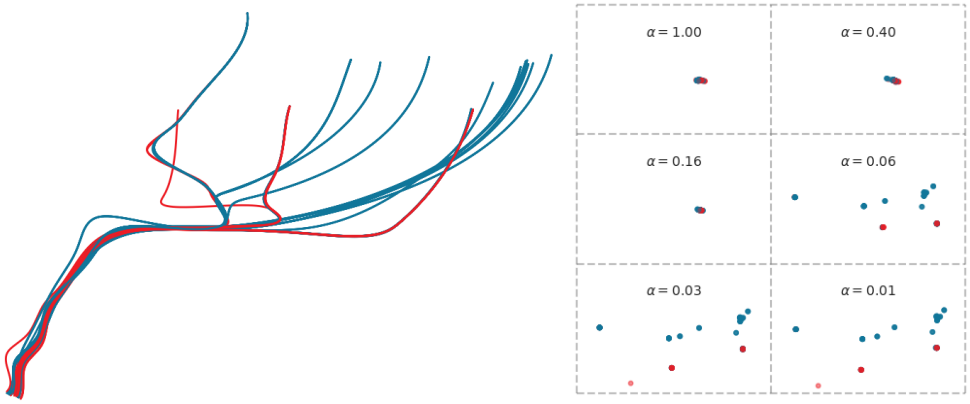


Figure 5.5: The trajectories of 109 works in the tree-SNE embedding in Figure 5.4. These works form a cluster in the tree approximately halfway through the embedding and subsequently split into more informative subclusters towards the top of the tree.

To illustrate how the tree-SNE embedding captures the evolution of clusters within the data, we track the paths of 109 works that form a single cluster roughly halfway into the embedding. This cluster was chosen because it contains a mix of American (80 works) and English (29 works) literature and its relatively small size allows us to easily examine the works that appear as well as subclusters that form within the data. The trajectories of these works are shown in Figure 5.5.

The 109 works in this cluster are mostly pieces of travel writing and journals by writers from the 19th century. Given that these works center a common theme, a macro-level cluster is expected. We see that as the cluster evolves within the tree, additional branches begin to appear. Notably, in the final layer of the embedding, all except one subcluster is formed entirely of either English or American literature. The one exception is a cluster containing 13 works of English literature and 18 works of American literature. However, the American literature in this subcluster consists entirely of travel writings from the novelist Nathaniel Hawthorne during his time in the United Kingdom and continental Europe.

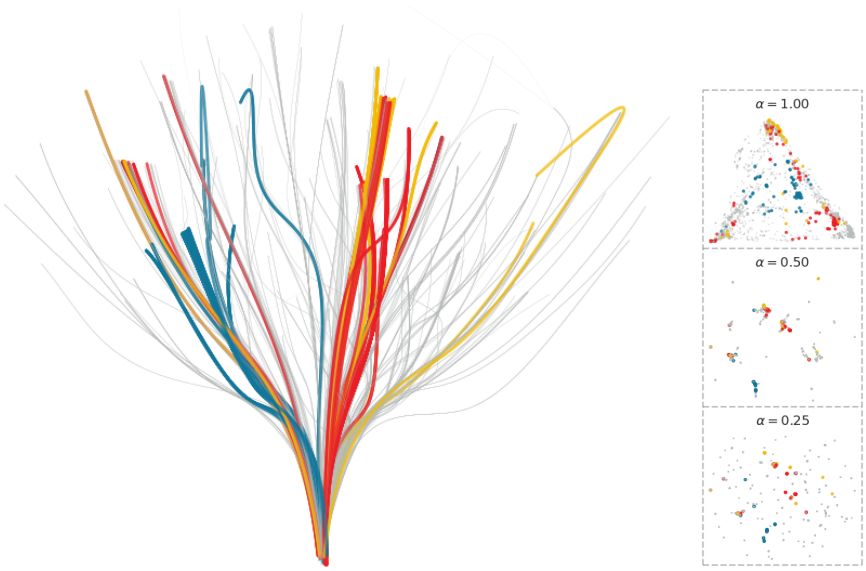


Figure 5.6: Tree-SNE embedding of Shakespeare and contemporaries. Gold points are works attributed to Shakespeare, red points are by Marlowe, and blue points are by Bacon.

5.4.2 The Anti-Stratfordian Theory

Although it is widely accepted that the plays and poems attributed to Shakespeare were indeed written by William Shakespeare of Stratford-upon-Avon, there is a fringe theory amongst literary historians that these works were produced by another author. This is known as the Anti-Stratfordian Theory. Motivated by the cluster formed by Shakespeare and his contemporaries in Figure 5.4, we apply tree-SNE to explore this authorship question.

Two prominent candidates for Shakespearean authorship are the philospher Sir Francis Bacon and the playwright Christopher Marlowe. We constructed a dataset of 2419 English

language works by Shakespeare (68 works), Bacon (87 works), Marlowe (114 works), and additional contemporaries from the HathiTrust database. Many of these contemporaries have also been proposed as candidate authors for Shakespeare's work yet are not as popular as Marlowe and Bacon. Since this dataset is relatively small, we generate a tree-SNE embedding with only 10 layers corresponding to degrees of freedom between 1 and 0.25. The embedding and slices of three layers are in Figure 5.6.

We applied DBSCAN to various layers in the embedding to determine clusters within the data. Initially, there is little separation between any of the included authors, although we see that works by Bacon are more central and works by Marlowe and Shakespeare are pushed towards the corners of the embedding. As the layers of the embedding increase, we see that clusters consisting of works attributed to a single author start to appear. In the final layer of the embedding, there are two larger clusters containing works of Shakespeare. More than half of the works attributed to Shakespeare appear in one large cluster of 304 works. This cluster mostly consists of plays and many authors including Marlowe, but not Bacon, are represented. Other than Shakespeare, ten authors (including Marlowe) appear at least ten times in this cluster. Bacon and Shakespeare only appear together in one cluster that contains 6 works of Bacon, 7 works of Shakespeare, and 256 works total. Thus, the vast majority of the works attributed to these authors do not appear in the same clusters within the data. Within the confines of this small experiment, it appears that Marlowe is a more likely candidate than Bacon for the true author of works attributed to Shakespeare, although there is limited evidence to support the Anti-Stratfordian theory.

5.5 Conclusion

We have successfully shown that the tree-SNE procedure introduced in [83] generates continuous structures that produce finer and finer clusterings within high-dimensional data. Moreover, by creating tree-SNE embeddings of three high-dimensional datasets, we observed that these clusterings are indeed meaningful and elucidate both macro- and micro-level structures within data. While we have shown that tree-SNE embeddings have continuous strucutre, it is yet unclear why and when different branches form within the tree. To fully understand this, a greater understanding of the inner workings of t-SNE is needed. An interesting question is to uncover these threshold degrees of freedom for highly structured data, such as points sampled from a mixture of mixtures of Gaussians.

Appendix A

APPENNDIX TO CHAPTER 2

A.1 Proofs from Section 2.2

A.1.1 Proof of Lemma 2.2.2

The fact that K^* is invariant follows directly from right-translation invariance of the Haar measure, $\mu(Sg) = \mu(S)$, for all $g \in G$ and all Borel sets S. Suppose now that K is PSD and fix arbitrary points $x_1, \dots, x_n \in V$. Using linearity of the integral, we compute

$$\sum_{i,j=1}^{n} c_i c_j K^*(x_i, x_j) = \int_G \int_G \underbrace{\sum_{i,j=1}^{n} c_i c_j K(gx_i, hx_j)}_{\geq 0} \, d\mu(g) d\mu(h),$$

and therefore K^* is PSD.

A.1.2 Proof of Lemma 2.2.5

Writing the polynomial kernel K in terms of the basis $f_1, \dots, f_N$ yields the equation

$$K(x, y) = \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij} f_j(x) f_i(y),$$

for some constants c_{ij}. Let us now argue uniqueness. To this end, suppose that there exists another matrix $\tilde{C}$ satisfying (2.8). Then we may write $K(x, y)$ in two ways:

$$\sum_{i=1}^{N} \left(\sum_{j=1}^{N} c_{ij} f_j(x) \right) f_i(y) = K(x, y) = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \tilde{c}_{ij} f_j(x) \right) f_i(y).$$

Linear independence of f_i implies the equality $\sum_{j=1}^{N} c_{ij} f_j(x) = \sum_{j=1}^{N} \tilde{c}_{ij} f_j(x)$ for any x and any index $i = 1, \dots, N$. Again using linear independence, we deduce $c_{ij} = \tilde{c}_{ij}$ for all i, j, as claimed. Finally, symmetry of $K(x, y)$ in x and y implies that (2.8) remains true with C replaced by $(C + C^\top)/2$. Uniqueness of the matrix C satisfying (2.8), which we already established, therefore implies that C is symmetric. Taking the expectation $\mathbb{E}_{g,h}$ of both sides

in (2.8) directly yields (2.9).

A.1.3 Proof of Theorem 2.2.7

Define $\mathcal{F}(x) = [f_1(x), \dots, f_N(x)]$, $\mathcal{F}_*(x) = [f_1^*(x), \dots, f_N^*(x)]$, and $\mathcal{V}(x) = [g_1(x), \dots, g_r(x)]$ where $\{g_j\}_{j=1}^r$ is a basis for $\operatorname{span}\{f_1^*, \dots, f_N^*\}$. Then expanding f_i^* in the $\{g_j\}$ basis we deduce that there exists a matrix $B \in \mathbb{R}^{r \times N}$ such that $\mathcal{F}_*(x) = \mathcal{V}(x)B$ for all $x \in \mathbb{R}^d$. Consequently using (2.8), we deduce

$$K(x, y) = \mathcal{V}(x) \cdot (BCB^\top) \cdot \mathcal{V}(x)^\top.$$

Forming a rank r factorization $BCB^\top = LR^\top$ for some $L, R \in \mathbb{R}^{r \times r}$ therefore yields the equality $K(x, y) = \langle \mathcal{V}(x)L, \mathcal{V}(x)R \rangle$, thereby verifying the inequality $\operatorname{rank} K \leq r$ in (2.10).

Next, we show the equality in (2.10). Clearly, since f_i^* lie in $\mathbb{R}[V]_m^G$, the inequality holds:

$$r \leq \dim(\mathbb{R}[V]_m^G). \tag{A.1}$$

It remains to argue that $\dim(\mathbb{R}[V]_m^G) \leq r$. Choose $h(x) \in \mathbb{R}[V]_m^G$. Since $\mathbb{R}[V]_m^G$ is a subspace of $\mathbb{R}[V]_m$, it follows that there exist coefficients $c_1, \dots, c_N$ such that $h(x) = \sum_{i=1}^N c_i f_i(x)$. Then, as h is invariant under the action of G and G acts linearly on $\mathbb{R}[V]_m$, the equalities

$$h(x) = h^*(x) = \sum_{i=1}^N c_i f_i^*(x)$$

hold. It follows that $f_1^*(x), \dots, f_N^*(x)$ form a spanning set for $\mathbb{R}[V]_m^G$ and thus the inequality $\dim(\mathbb{R}[V]_m^G) \leq r$ holds as claimed.

A.2 Proofs from Section 2.6

A.2.1 Proof of Theorem 2.6.1

The proof follows along similar lines as the standard argument for linear regression in a fixed dimension [8, Proposition 3.5]. The fact that the estimator is unbiased follows immediately from its definition and the fact that $y = \Phi(X)\alpha^\star + \varepsilon$ with $\mathbb{E}\varepsilon = 0$. Take any $\alpha \in \mathbb{R}^d$,

expanding and using the fact that the noise is i.i.d, we derive

$$\begin{aligned} R_T^{(\alpha^\star)}(\alpha) &= \frac{1}{|T|}\sum_{(x,y)\in T}\mathbb{E}\left(\langle\varphi(x),\alpha\rangle - \langle\varphi(x),\alpha^\star\rangle + \varepsilon\right)^2 \\ &= \sigma^2 + \frac{1}{|T|}\sum_{(x,y)\in T}\mathbb{E}\left(\langle\varphi(x),\alpha\rangle - \langle\varphi(x),\alpha^\star\rangle\right)^2 \\ &= R_T^{(\alpha^\star)}(\alpha^\star) + \frac{1}{|T|}\sum_{(x,y)\in T}\mathbb{E}\left(\langle\varphi(x),\alpha\rangle - \langle\varphi(x),\alpha^\star\rangle\right)^2. \end{aligned}$$

In particular, the risk R_T is minimized at $\alpha^\star$. Then, the excess risk of $\hat{\alpha}$ is equal to

$$\begin{aligned} R_T^{(\alpha^\star)}(\hat{\alpha}) - R_T^{(\alpha^\star)}(\alpha^\star) &= \mathbb{E}\,\|\hat{\alpha}-\alpha^\star\|_{\Sigma_T}^2 \\ &= \mathbb{E}\left\|\left(\Phi(X)^\top\Phi(X)\right)^{-1}\Phi(X)^\top y - \alpha^\star\right\|_{\Sigma_T}^2 \\ &= \mathbb{E}\left\|\left(\Phi(X)^\top\Phi(X)\right)^{-1}\Phi(X)^\top(y-\Phi(X)\alpha^\star)\right\|_{\Sigma_T}^2 \\ &= \mathbb{E}\left\|\left(\Phi(X)^\top\Phi(X)\right)^{-1}\Phi(X)^\top\varepsilon\right\|_{\Sigma_T}^2 \\ &= \mathbb{E}\,\mathrm{Trace}\left(\varepsilon^\top\Phi(X)\left(\Phi(X)^\top\Phi(X)\right)^{-1}\Sigma_T\left(\Phi(X)^\top\Phi(X)\right)^{-1}\Phi(X)^\top\varepsilon\right) \\ &= \sigma^2\mathrm{Trace}\left(\left(\Phi(X)^\top\Phi(X)\right)^{-1}\Sigma_T\right) \\ &= \frac{\sigma^2}{n}\mathrm{Trace}\left(\Sigma_S^{-1}\Sigma_T\right), \end{aligned}$$

where the second to last inequality uses the cyclic invariance of the trace.

A.2.2 Proof of Illustration 6

Let $\bar{\Sigma}_S = \mathbb{E}\Sigma_S$ and $\bar{\Sigma}_T = \mathbb{E}\Sigma_T$. We can upper bound

$$\begin{aligned} \mathrm{Trace}\left(\Sigma_S^{-1}\Sigma_T\right) \le{}& \underbrace{\mathrm{Trace}\left(\bar{\Sigma}_S^{-1}\bar{\Sigma}_T\right)}_{Q_1} + \underbrace{\mathrm{Trace}\left(\left|\Sigma_S^{-1}-\bar{\Sigma}_S^{-1}\right|\left|\Sigma_T-\bar{\Sigma}_T\right|\right)}_{Q_2} \\ &+ \underbrace{\mathrm{Trace}\left(\left|\Sigma_S^{-1}-\bar{\Sigma}_S^{-1}\right|\bar{\Sigma}_T\right)}_{Q_3} + \underbrace{\mathrm{Trace}\left(\bar{\Sigma}_S^{-1}\left|\Sigma_T-\bar{\Sigma}_T\right|\right)}_{Q_4} \end{aligned} \tag{A.2}$$

where we use $|A|$ to denote the matrix with component-wise absolute values of A. To bound each one of these terms we will use the following lemma.

Lemma A.2.1. *Let $n = |S|$ and fix $\delta \in (0, 1/6)$, then with probability at least $1-4\exp(-\delta n)$*

we have

$$\left|\Sigma_S^{-1} - \bar{\Sigma}_S^{-1}\right|_{11} \leq 2\delta, \quad \left|\Sigma_S^{-1} - \bar{\Sigma}_S^{-1}\right|_{12} \leq 2\sqrt{\frac{\delta}{d}}, \quad \textit{and} \quad \left|\Sigma_S^{-1} - \bar{\Sigma}_S^{-1}\right|_{22} \leq 10\frac{\delta}{d}.$$

Similarly, let $m = |T|$ *and fix* $\delta \in (0, \infty)$*, with probability at least* $1 - 4\exp(-\delta m)$ *we have that*

$$\left|\Sigma_T - \bar{\Sigma}_T\right|_{11} = 0, \quad \left|\Sigma_T - \bar{\Sigma}_T\right|_{12} \leq \sqrt{\delta\bar{d}}, \quad \textit{and} \quad \left|\Sigma_T - \bar{\Sigma}_T\right|_{22} \leq 4\delta\bar{d}.$$

Before proving this lemma, let us show how it proves the conclusion from Illustration 6. Assume that the bounds in the Lemma hold with a fixed $\delta \in (0, 1/6)$. It is immediate that $Q_1 = 1 + \frac{\bar{d}}{d}$. Furthermore,

$$\begin{aligned}
Q_2 &\leq 4\delta \cdot \frac{\bar{d}}{d} + 40\delta^2 \cdot \frac{\bar{d}}{d} \leq 44\delta \cdot \left(1 + \frac{\bar{d}}{d}\right) \\
Q_3 &\leq 2\delta + 10\delta\frac{\bar{d}}{d} \leq 10\delta \cdot \left(1 + \frac{\bar{d}}{d}\right) \\
Q_4 &\leq 4\delta \cdot \frac{\bar{d}}{d} \leq 4\delta \cdot \left((1 + \frac{\bar{d}}{d}\right)
\end{aligned}$$

Combining all of these bounds yields

$$\text{Trace}\left(\Sigma_S^{-1}\Sigma_T\right) \leq (1 + 58\delta) \cdot \left(1 + \frac{\bar{d}}{d}\right),$$

setting $\Delta = \delta/58$ proves the promised bound. We now proof Lemma A.2.1.

Proof of Lemma A.2.1. Let us start with the statement for S. Enumerate the samples in S as $(x_1, y_1), \ldots, (x_n, y_n)$. For each x_i, let $z_i = \sum_{j=1}^{d}(x_i)_j$ be the sum of its coordinates. Let $\bar{z}_1 = \frac{1}{n}\sum_{i=1}^{n} z_i$ and $\bar{z}_2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2$. Then, we have

$$\Sigma_S = \begin{pmatrix} 1 & \bar{z}_1 \\ \bar{z}_1 & \bar{z}_2 \end{pmatrix} \quad \text{and} \quad \Sigma_S^{-1} = \frac{1}{\bar{z}_2 - \bar{z}_1^2}\begin{pmatrix} \bar{z}_2 & -\bar{z}_1 \\ -\bar{z}_1 & 1 \end{pmatrix}. \tag{A.3}$$

Thus,

$$\Sigma_S^{-1} - \bar{\Sigma}_S^{-1} = \begin{pmatrix} \frac{\bar{z}_2}{\bar{z}_2 - \bar{z}_1^2} - 1 & -\frac{\bar{z}_1}{\bar{z}_2 - \bar{z}_1^2} \\ -\frac{\bar{z}_1}{\bar{z}_2 - \bar{z}_1^2} & \frac{1}{\bar{z}_2 - \bar{z}_1^2} - \frac{1}{d} \end{pmatrix}.$$

The random variables $z_i \sim N(0, d)$ and so $\bar{z}_1 \sim N(0, \frac{d}{n^2})$ and $\bar{z}_2 = \frac{d}{n}w$ with $w \sim \chi_n^2$, a chi-squared distribution with n degrees of freedom.

Fact 1 ([103])**.** Let $z \sim N(0, \sigma^2)$, then for any $t \geq 0$ we have $\mathbb{P}(|z| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Fact 2 ([54])**.** Let $w \sim \chi^2_n$, then for any $t \geq 0$ we have that

$$\max\left\{\mathbb{P}(2\sqrt{nt} + 2t \leq w - n), \mathbb{P}(w - n \leq -2\sqrt{nt})\right\} \leq \exp(-t).$$

Equipped with these two facts, it is easy to derive that for any fixed $\delta > 0$, we have $\mathbb{P}(|\bar{z}_1| \geq \sqrt{\delta d}) \leq 2\exp\left(-\delta n^2\right)$ and $\max\left\{\mathbb{P}(2(\delta + \delta^2)d \leq \bar{z}_2 - d), \mathbb{P}(\bar{z}_2 - d \leq -2\delta d)\right\} \leq \exp(-\delta n)$. Consider the event

$$\mathcal{E} = \left\{|\bar{z}_1| \leq \sqrt{\delta d} \quad \text{and} \quad (1 - 2\delta)d \leq \bar{z}_2 \leq (1 + 4\delta)d\right\}.$$

A union bound argument yields $\mathbb{P}(\mathcal{E}) \geq 1 - 4\exp\left(-\delta n\right)$, which matches the stated probability. For the rest of the proof assume that $\mathcal{E}$ holds with $\delta \in (0, 1/6)$.

We are finally ready to prove the three stated bounds for S. We will repeatedly use that $\delta/(1 - 3\delta) \leq 2\delta$ for $\delta < 1/6$. Let us start with the first statement:

$$\left|\frac{\bar{z}_2}{\bar{z}_2 - \bar{z}_1^2} - 1\right| = \left|\frac{\bar{z}_1^2}{\bar{z}_2 - \bar{z}_1^2}\right| \leq \frac{|\bar{z}_1^2|}{|\bar{z}_2| - |\bar{z}_1^2|} \leq \frac{\delta}{(1 - 3\delta)} \leq 2\delta.$$

Similarly,

$$\left|\frac{\bar{z}_1}{\bar{z}_2 - \bar{z}_1^2}\right| = \left|\frac{\bar{z}_1}{\bar{z}_2 - \bar{z}_1^2}\right| \leq \frac{|\bar{z}_1|}{|\bar{z}_2| - |\bar{z}_1^2|} \leq \frac{\sqrt{\delta d}}{(1 - 3\delta)d} \leq 2\sqrt{\frac{\delta}{d}}$$

where the last inequality follows for any $\delta < 1/6$ and $d \geq 1$. Finally,

$$\left|\frac{1}{\bar{z}_2 - \bar{z}_1^2} - \frac{1}{d}\right| = \left|\frac{d - \bar{z}_2 + \bar{z}_1^2}{(\bar{z}_2 - \bar{z}_1^2)d}\right| \leq \frac{|d - \bar{z}_2| + \bar{z}_1^2}{(|\bar{z}_2| - |\bar{z}_1^2|)d} \leq \frac{5\delta}{(1 - 3\delta)d} \leq 10\frac{\delta}{d}.$$

This completes the statement involving S.

To show the statement for T we use a very similar strategy. Notice that Σ_T can also be written as in (A.3) for a $\bar{z}_1$ and $\bar{z}_2$ with an analogous distribution where we substitute d with $\bar{d}$. Then, assuming again that we are in $\mathcal{E}$—defined with the the new $\bar{z}_1$ and $\bar{z}_2$ and with $\bar{d}$ in place of d—yields the stated probability bound. Hence, $\left(\Sigma_T - \bar{\Sigma}_T\right)_{11} = 0$, while

$$\left|\left(\Sigma_T - \bar{\Sigma}_T\right)_{12}\right| = |z_1| \leq \sqrt{\delta d}, \quad \text{and} \quad \left|\left(\Sigma_T - \bar{\Sigma}_T\right)_{22}\right| = |\bar{z}_2 - d| \leq 4\delta d,$$

which completes the proof. □

A.2.3 Proof of Theorem 2.6.2

The proof follows along similar lines as the argument in [73], also summarized in [8, Section 3.7]. Define $n = |S|$. We have already proved an upper bound that matches (2.21), so we focus on proving a lower bound. We can lower bound the supremum using a prior on $\alpha \sim N\left(0, \frac{\sigma^2}{\lambda n} \cdot I\right)$ where $\lambda > 0$ is any fixed positive scalar. Let Y denote the random vector of new observations $Y_i = \langle \varphi(x_i), \alpha \rangle + \epsilon_i$ with $x_i \in S$. In particular,

$$\begin{aligned}
\inf_{\widehat{\alpha}} \sup_{\alpha} \left\{ R_T^{(\alpha)}(\widehat{\alpha}(Y)) - R_T^{(\alpha)}(\alpha) \right\} &= \inf_{\widehat{\alpha}} \sup_{\alpha} \mathbb{E}_Y \|\widehat{\alpha} - \alpha\|_{\Sigma_T}^2 \\
&\geq \inf_{\widehat{\alpha}} \mathbb{E}_\alpha \mathbb{E}_Y \left[\|\widehat{\alpha}(Y) - \alpha\|_{\Sigma_T}^2 \,\middle|\, \alpha \right] \\
&= \inf_{\widehat{\alpha}} \mathbb{E}_Y \mathbb{E}_\alpha \left[\|\widehat{\alpha}(Y) - \alpha\|_{\Sigma_T}^2 \middle| Y \right] \\
&\geq \mathbb{E}_Y \inf_{\widehat{\alpha}} \mathbb{E}_\alpha \left[\|\widehat{\alpha}(Y) - \alpha\|_{\Sigma_T}^2 \middle| Y \right],
\end{aligned} \tag{A.4}$$

where the second line uses the prior, and the fourth line uses Jensen's inequality. We can recognize a minimizer for the inner infimum via first-order optimality conditions as the conditional mean $\widehat{\alpha}_\star = \mathbb{E}[\alpha \mid Y]$. In order to explicitly compute $\widehat{\alpha}_\star$, observe that $(\alpha, Y) = (\alpha, \Phi(X)\alpha + \epsilon)$ is a jointly Gaussian vector:

$$(\alpha, Y) \sim N\left(0, \frac{\sigma^2}{\lambda n} \begin{bmatrix} I & \Phi(X)^\top \\ \Phi(X) & \Phi(X)\Phi(X)^\top + \lambda n \cdot I \end{bmatrix}\right).$$

Next, we will use the following two standard lemmas, whose proofs are elementary.

Lemma A.2.2. *Let* $(X, Y) \sim N\left(0, \begin{bmatrix} \Sigma_{xx} & \Sigma xy \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$, *then,* $\mathbb{E}(X \mid Y) = \Sigma_{xy}\Sigma_{yy}^{-1} Y$.

Lemma A.2.3. *Let* $Q \in \mathbb{R}^{d \times r}$ *be a rank* r *matrix, then,* $Q^\top \left(QQ^\top - I_d\right)^{-1} = \left(Q^\top Q - I_r\right)^{-1} Q^\top$.

Thus using Lemma A.2.2, we compute

$$\begin{aligned}
\widehat{\alpha}_\star &= \frac{\sigma^2}{\lambda n} \Phi(X)^\top \left(\frac{\sigma^2}{\lambda n} \Phi(X)\Phi(X)^\top + \sigma^2 I \right)^{-1} Y \\
&= \frac{\sigma^2}{\lambda n} \left(\frac{\sigma^2}{\lambda n} \Phi(X)^\top \Phi(X) + \sigma^2 I \right)^{-1} \Phi(X)^\top Y \\
&= \left(\Phi(X)^\top \Phi(X) + \lambda n I \right)^{-1} \Phi(X)^\top Y
\end{aligned}$$

where the second line follows from Lemma A.2.3. Substituting this expression into (A.4)

yields

$$
\begin{aligned}
&\inf_{\widehat{\alpha}} \sup_{\alpha} \left\{ R_T^{(\alpha)}(\widehat{\alpha}) - R_T^{(\alpha)}(\alpha) \right\} \\
&\quad \geq \mathbb{E}_Y \mathbb{E}_{\alpha|Y} \left\| \left(\Phi(X)^\top \Phi(X) + \lambda n I\right)^{-1} \Phi(X)^\top Y - \alpha \right\|_{\Sigma_T}^2 \\
&\quad = \mathbb{E}_{\alpha,\varepsilon} \left\| \left(\Phi(X)^\top \Phi(X) + \lambda n I\right)^{-1} \Phi(X)^\top \left(\Phi(X)\alpha + \varepsilon\right) - \alpha \right\|_{\Sigma_T}^2 \\
&\quad = \mathbb{E}_{\alpha,\varepsilon} \left\| \left(\Phi(X)^\top \Phi(X) + \lambda n I\right)^{-1} \Phi(X)^\top \left(\Phi(X)\alpha - \left(\Phi(X) + \lambda n I\right)\alpha - \varepsilon\right) - \alpha \right\|_{\Sigma_T}^2 \\
&\quad = \mathbb{E}_{\alpha,\varepsilon} \left\| \left(\Phi(X)^\top \Phi(X) + \lambda n I\right)^{-1} \Phi(X)^\top \left(\varepsilon - n\lambda\alpha\right) \right\|_{\Sigma_T}^2 \\
&\quad = \mathbb{E}_{\alpha} \left\| \lambda \cdot \left(\Sigma_S + \lambda \cdot I\right)^{-1} \alpha \right\|_{\Sigma_T}^2 + \mathbb{E}_{\varepsilon} \left\| \left(\Sigma_S + \lambda I\right)^{-1} \Phi\left(X\right)^\top \varepsilon \right\|_{\Sigma_T}^2 \\
&\quad = \frac{\lambda\sigma^2}{n} \mathrm{Trace}\left(\left(\Sigma_S + \lambda I\right)^{-2} \Sigma_T \right) + \frac{\sigma^2}{n} \mathrm{Trace}\left(\left(\Sigma_S + \lambda I\right)^{-2} \Sigma_S \Sigma_T \right) \\
&\quad = \frac{\sigma^2}{n} \mathrm{Trace}\left(\left(\Sigma_S + \lambda I\right)^{-2} \left(\Sigma_S + \lambda I\right) \Sigma_T \right) \\
&\quad = \frac{\sigma^2}{n} \mathrm{Trace}\left(\left(\Sigma_S + \lambda I\right)^{-1} \Sigma_T \right).
\end{aligned}
$$

Taking $\lambda \to 0$ yields a lower bound of $\frac{\sigma^2}{n} \mathrm{Trace}\left(\left(\Sigma_S\right)^{-1} \Sigma_T \right)$, establishing the result.

BIBLIOGRAPHY

[1] Sameer Agarwal et al. "An atlas for the pinhole camera". In: *Found. Comput. Math.* 24.1 (2024), pp. 227–277. ISSN: 1615-3375,1615-3383. DOI: 10.1007/s10208-022-09592-6. URL: https://doi.org/10.1007/s10208-022-09592-6.

[2] Chris Aholt, Bernd Sturmfels, and Rekha Thomas. "A Hilbert Scheme in Computer Vision". In: *Canadian Journal of Mathematics* 65.5 (2013), pp. 961–988. ISSN: 1496-4279. DOI: 10.4153/cjm-2012-023-2. URL: http://dx.doi.org/10.4153/CJM-2012-023-2.

[3] Yulia Alexandr, Joe Kileel, and Bernd Sturmfels. "Moment varieties for mixtures of products". In: *Proceedings of the 2023 International Symposium on Symbolic and Algebraic Computation*. 2023, pp. 53–60.

[4] Jason M. Altschuler and Pablo A. Parrilo. "Kernel approximation on algebraic varieties". In: *SIAM J. Appl. Algebra Geom.* 7.1 (2023), pp. 1–28. ISSN: 2470-6566. DOI: 10.1137/21M1425050. URL: https://doi.org/10.1137/21M1425050.

[5] N. Aronszajn. "Theory of reproducing kernels". In: *Trans. Amer. Math. Soc.* 68 (1950), pp. 337–404. ISSN: 0002-9947,1088-6850. DOI: 10.2307/1990404. URL: https://doi.org/10.2307/1990404.

[6] Sanjeev Arora, Wei Hu, and Pravesh K Kothari. "An analysis of the t-sne algorithm for data visualization". In: *Conference on learning theory*. PMLR. 2018, pp. 1455–1462.

[7] Haim Avron, Kenneth L Clarkson, and David P Woodruff. "Faster kernel ridge regression using sketching and preconditioning". In: *SIAM Journal on Matrix Analysis and Applications* 38.4 (2017), pp. 1116–1138.

[8] Francis Bach. *Learning theory from first principles*. MIT press, 2024.

[9] Mikhail Belkin and Partha Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation". In: *Neural computation* 15.6 (2003), pp. 1373–1396.

[10] Clément Berenfeld and Marc Hoffmann. *Density estimation on an unknown submanifold*. 2020. arXiv: 1910.08477 [math.ST]. URL: https://arxiv.org/abs/1910.08477.

[11] Tyrus Berry and Timothy Sauer. "Density estimation on manifolds with boundary". In: *Comput. Statist. Data Anal.* 107 (2017), pp. 1–17. ISSN: 0167-9473,1872-7352. DOI: 10.1016/j.csda.2016.09.011. URL: https://doi.org/10.1016/j.csda.2016.09.011.

[12] Alberto Bietti, Luca Venturi, and Joan Bruna. "On the sample complexity of learning under geometric stability". In: *Advances in neural information processing systems* 34 (2021), pp. 18673–18684.

[13] Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. "Attraction-repulsion spectrum in neighbor embeddings". In: *Journal of Machine Learning Research* 23.95 (2022), pp. 1–32.

[14] Joan Bruna et al. "Spectral networks and locally connected networks on graphs". In: *arXiv preprint arXiv:1312.6203* (2013).

[15] Theophilos Cacoullos. "Estimation of a multivariate density". In: *Annals of the Institute of Statistical Mathematics* 18 (1966), pp. 179–189.

[16] T Tony Cai and Rong Ma. "Theoretical foundations of t-sne for visualizing high-dimensional clustered data". In: *Journal of Machine Learning Research* 23.301 (2022), pp. 1–54.

[17] Yifan Chen et al. "Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations". In: *arXiv preprint arXiv:2207.06503* (2022).

[18] Tin Sum Cheng et al. "A Theoretical Analysis of the Test Error of Finite-Rank Kernel Ridge Regression". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 4767–4798. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/0f580c1ace3b857a390575ca42de7938-Paper-Conference.pdf.

[19] Thomas Church, Jordan Ellenberg, and Benson Farb. "Representation stability in cohomology and asymptotics for families of varieties over finite fields". In: *Contemporary Mathematics* 620 (2014), pp. 1–54.

[20] Thomas Church and Benson Farb. "Representation theory and homological stability". In: *Advances in Mathematics* 245 (2013), pp. 250–314.

[21] Ronald R. Coifman and Stéphane Lafon. "Diffusion maps". In: *Appl. Comput. Harmon. Anal.* 21.1 (2006), pp. 5–30. ISSN: 1063-5203,1096-603X. DOI: 10.1016/j.acha.2006.04.006. URL: https://doi.org/10.1016/j.acha.2006.04.006.

[22] Aldo Conca et al. “Noetherianity up to symmetry”. In: *Combinatorial Algebraic Geometry: Levico Terme, Italy 2013, Editors: Sandra Di Rocco, Bernd Sturmfels* (2014), pp. 33–61.

[23] Erin Connelly, Timothy Duff, and Jessie Loucks-Tavitas. “Algebra and geometry of camera resectioning”. In: *Mathematics of Computation* 94.355 (2025), pp. 2613–2643.

[24] Austin Conner, Kangjin Han, and Mateusz Michalek. “Sullivant-Talaska ideal of the cyclic Gaussian Graphical Model”. In: *arXiv preprint arXiv:2308.05561* (2023).

[25] Kurt Cutajar et al. “Preconditioning kernel matrices”. In: *International conference on machine learning*. PMLR. 2016, pp. 2529–2538.

[26] Mateo Díaz et al. *Invariant Kernels: Rank Stabilization and Generalization Across Dimensions*. 2025. arXiv: 2502.01886 [math.OC]. URL: https://arxiv.org/abs/2502.01886.

[27] Mateo Díaz et al. “Robust, randomized preconditioning for kernel ridge regression”. In: *arXiv preprint arXiv:2304.12465* (2023).

[28] Timothy Duff, Jack Kendrick, and Rekha R. Thomas. *Universal Gröbner Bases of (Universal) Multiview Ideals*. 2025. arXiv: 2509.12376 [math.AC]. URL: https://arxiv.org/abs/2509.12376.

[29] David Eisenbud. *Commutative Algebra*. Vol. 150. Graduate Texts in Mathematics. With a view toward algebraic geometry. Springer-Verlag, New York, 1995, pp. xvi+785. ISBN: 0-387-94268-8; 0-387-94269-6. DOI: 10.1007/978-1-4612-5350-1. URL: https://doi.org/10.1007/978-1-4612-5350-1.

[30] Bryn Elesedy. “Provably Strict Generalisation Benefit for Invariance in Kernel Methods”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 17273–17283. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/8fe04df45a22b63156ebabbb064fcd5e-Paper.pdf.

[31] “Entry A007717”. In: *The On-Line Encyclopedia of Integer Sequences* (). URL: https://oeis.org/A007717.

[32] P. Fleischmann. “On invariant theory of finite groups”. In: *Invariant theory in all characteristics*. Vol. 35. CRM Proc. Lecture Notes. Amer. Math. Soc., Providence, RI, 2004, pp. 43–69. ISBN: 0-8218-3244-1. DOI: 10.1090/crmp/035/04. URL: https://doi.org/10.1090/crmp/035/04.

[33] Gerald B. Folland. *A course in abstract harmonic analysis*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1995, pp. x+276. ISBN: 0-8493-8490-7.

[34] William Fulton and Joe Harris. *Representation theory*. Vol. 129. Graduate Texts in Mathematics. A first course, Readings in Mathematics. Springer-Verlag, New York, 1991, pp. xvi+551. ISBN: 0-387-97527-6; 0-387-97495-4. DOI: 10.1007/978-1-4612-0979-9. URL: https://doi.org/10.1007/978-1-4612-0979-9.

[35] David Ginsbourger, Olivier Roustant, and Nicolas Durrande. "On degeneracy and invariances of random fields paths with applications in Gaussian process modelling". In: *J. Statist. Plann. Inference* 170 (2016), pp. 117–128. ISSN: 0378-3758,1873-1171. DOI: 10.1016/j.jspi.2015.10.002. URL: https://doi.org/10.1016/j.jspi.2015.10.002.

[36] David Ginsbourger et al. "Argumentwise invariant kernels for the approximation of invariant functions". In: *Ann. Fac. Sci. Toulouse Math. (6)* 21.3 (2012), pp. 501–527. ISSN: 0240-2963,2258-7519. DOI: 10.5802/afst.1343. URL: https://doi.org/10.5802/afst.1343.

[37] Marco Gori, Gabriele Monfardini, and Franco Scarselli. "A new model for learning in graph domains". In: *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.* Vol. 2. IEEE. 2005, pp. 729–734.

[38] Daniel R. Grayson and Michael E. Stillman. *Macaulay2, a software system for research in algebraic geometry*. Available at http://www2.macaulay2.com.

[39] P. Greengard et al. *Factor Clustering with T-SNE*. SSRN, 2020. URL: https://books.google.com/books?id=Jxb2zgEACAAJ.

[40] G. H. Hardy and S. Ramanujan. "Asymptotic formulæ in combinatory analysis [Proc. London Math. Soc. (2) **17** (1918), 75–115]". In: *Collected papers of Srinivasa Ramanujan*. AMS Chelsea Publ., Providence, RI, 2000, pp. 276–309. ISBN: 0-8218-2076-1.

[41] Harrie Hendriks. "Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions". In: *Ann. Statist.* 18.2 (1990), pp. 832–849. ISSN: 0090-5364,2168-8966. DOI: 10.1214/aos/1176347628. URL: https://doi.org/10.1214/aos/1176347628.

[42] Geoffrey E Hinton and Sam Roweis. "Stochastic Neighbor Embedding". In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press, 2002. URL: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

[43] Daoji Huang and Matt Larson. "Fine multidegrees, universal Gröbner bases, and matrix Schubert varieties". In: *arXiv preprint arXiv:2410.02135* (2024).

[44] Mathieu Jacomy et al. "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software". In: *PloS one* 9.6 (2014), e98679.

[45] Jack Kendrick. *Density Estimation on Rectifiable Sets*. 2025. arXiv: 2505.23023 [math.ST]. URL: https://arxiv.org/abs/2505.23023.

[46] Jack Kendrick. *The Tree-SNE Tree Exists*. 2025. arXiv: 2510.15014 [stat.ML]. URL: https://arxiv.org/abs/2510.15014.

[47] Joe Kileel and Kathlén Kohn. *Snapshot of Algebraic Vision*. 2023. arXiv: 2210.11443 [math.AG]. URL: https://arxiv.org/abs/2210.11443.

[48] Yoon Tae Kim and Hyun Suk Park. "Geometric structures arising from kernel density estimation on Riemannian manifolds". In: *J. Multivariate Anal.* 114 (2013), pp. 112–126. ISSN: 0047-259X,1095-7243. DOI: 10.1016/j.jmva.2012.07.006. URL: https://doi.org/10.1016/j.jmva.2012.07.006.

[49] Stefan Klus et al. "Symmetric and antisymmetric kernels for machine learning problems in quantum physics and chemistry". In: *Machine Learning: Science and Technology* 2.4 (Aug. 2021), p. 045016. ISSN: 2632-2153. DOI: 10.1088/2632-2153/ac14ad. URL: http://dx.doi.org/10.1088/2632-2153/ac14ad.

[50] Allen Knutson and Ezra Miller. "Gröbner geometry of Schubert polynomials". In: *Ann. of Math. (2)* 161.3 (2005), pp. 1245–1318. ISSN: 0003-486X,1939-8980. DOI: 10.4007/annals.2005.161.1245. URL: https://doi.org/10.4007/annals.2005.161.1245.

[51] Dmitry Kobak and George C Linderman. "UMAP does not preserve global structure any better than t-SNE when using the same initialization". In: *BioRxiv* (2019), pp. 2019–12.

[52] Dmitry Kobak et al. “Heavy-Tailed Kernels Reveal a Finer Cluster Structure in t-SNE Visualisations”. In: *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2020, pp. 124–139. ISBN: 9783030461508. DOI: 10.1007/978-3-030-46150-8_8. URL: http://dx.doi.org/10.1007/978-3-030-46150-8_8.

[53] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[54] Beatrice Laurent and Pascal Massart. “Adaptive estimation of a quadratic functional by model selection”. In: *Annals of statistics* (2000), pp. 1302–1338.

[55] John M. Lee. *Introduction to smooth manifolds*. Second. Vol. 218. Graduate Texts in Mathematics. Springer, New York, 2013, pp. xvi+708. ISBN: 978-1-4419-9981-8.

[56] Eitan Levin and Venkat Chandrasekaran. “Free descriptions of convex sets”. In: *arXiv preprint arXiv:2307.04230* (2023).

[57] Eitan Levin and Mateo Díaz. “Any-dimensional equivariant neural networks”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 2773–2781.

[58] George C. Linderman and Stefan Steinerberger. *Clustering with t-SNE, provably*. 2017. arXiv: 1706.02582 [cs.LG]. URL: https://arxiv.org/abs/1706.02582.

[59] George C. Linderman et al. “Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data”. In: *Nature Methods* 16.3 (Feb. 2019), pp. 243–245. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0308-4. URL: http://dx.doi.org/10.1038/s41592-018-0308-4.

[60] Siyuan Ma and Mikhail Belkin. “Diving into the shallows: a computational perspective on large-scale shallow learning”. In: *Advances in neural information processing systems* 30 (2017).

[61] Siyuan Ma and Mikhail Belkin. “Kernel machines that adapt to GPUs for effective large batch training”. In: *Proceedings of Machine Learning and Systems* 1 (2019), pp. 360–373.

[62] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[63] Marvin Marcus and Henryk Minc. *A survey of matrix theory and matrix inequalities*. Reprint of the 1969 edition. Dover Publications, Inc., New York, 1992, pp. xii+180. ISBN: 0-486-67102-X.

[64] Sohir Maskey, Ron Levie, and Gitta Kutyniok. "Transferability of graph neural networks: an extended graphon approach". In: *Applied and Computational Harmonic Analysis* 63 (2023), pp. 48–83.

[65] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[66] Giacomo Meanti et al. "Efficient Hyperparameter Tuning for Large Scale Kernel Ridge Regression". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. 2022.

[67] Giacomo Meanti et al. "Kernel methods through the roof: handling billions of points efficiently". In: *Advances in Neural Information Processing Systems 32*. 2020.

[68] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Learning with invariances in random features and kernel models". In: *Conference on Learning Theory*. PMLR. 2021, pp. 3351–3418.

[69] Francesco Mezzadri. "How to generate random matrices from the classical compact groups". In: *Notices Amer. Math. Soc.* 54.5 (2007), pp. 592–604. ISSN: 0002-9920,1088-9477.

[70] Alessio Micheli. "Neural network for graphs: A contextual constructive approach". In: *IEEE Transactions on Neural Networks* 20.3 (2009), pp. 498–511.

[71] Ezra Miller and Bernd Sturmfels. *Combinatorial Commutative Algebra*. Vol. 227. Graduate Texts in Mathematics. Springer-Verlag, New York, 2005, pp. xiv+417. ISBN: 0-387-22356-8.

[72] Jeffrey W. Miller and Matthew T. Harrison. "Exact sampling and counting for fixed-margin matrices". In: *Ann. Statist.* 41.3 (2013), pp. 1569–1592. ISSN: 0090-5364,2168-8966. DOI: 10.1214/13-AOS1131. URL: https://doi.org/10.1214/13-AOS1131.

[73] Jaouad Mourtada. "Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices". In: *The Annals of Statistics* 50.4 (2022), pp. 2157–2178.

[74] Fionn Murtagh and Pedro Contreras. "Algorithms for hierarchical clustering: an overview". In: *WIREs Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97. DOI: https://doi.org/10.1002/widm.53. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.53. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.53.

[75] Frank Nielsen. "Hierarchical Clustering". In: Feb. 2016, pp. 195–211. ISBN: 978-3-319-21902-8. DOI: 10.1007/978-3-319-21903-5_8.

[76] James G. Oxley. *Matroid Theory*. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1992, pp. xii+532. ISBN: 0-19-853563-5.

[77] Arkadas Ozakin and Alexander Gray. "Submanifold density estimation". In: *Advances in Neural Information Processing Systems*. Ed. by Y. Bengio et al. Vol. 22. Curran Associates, Inc., 2009. URL: https://proceedings.neurips.cc/paper_files/paper/2009/file/2ac2406e835bd49c70469acae337d292-Paper.pdf.

[78] Pablo A. Parrilo and Ali Jadbabaie. "Approximation of the joint spectral radius using sum of squares". In: *Linear Algebra Appl.* 428.10 (2008), pp. 2385–2402. ISSN: 0024-3795,1873-1856. DOI: 10.1016/j.laa.2007.12.027. URL: https://doi.org/10.1016/j.laa.2007.12.027.

[79] Emanuel Parzen. "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 1065–1076. ISSN: 00034851. URL: http://www.jstor.org/stable/2237880 (visited on 05/07/2025).

[80] Bruno Pelletier. "Kernel density estimation on Riemannian manifolds". In: *Statistics and Probability Letters* 73.3 (2005), pp. 297–304. ISSN: 0167-7152. DOI: https://doi.org/10.1016/j.spl.2005.04.004. URL: https://www.sciencedirect.com/science/article/pii/S0167715205001239.

[81] Charles R Qi et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[82] Charles Ruizhongtai Qi et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space". In: *Advances in neural information processing systems* 30 (2017).

[83] Isaac Robinson and Emma Pierce-Hoffman. *Tree-SNE: Hierarchical Clustering and Visualization Using t-SNE*. 2020. arXiv: 2002.05687 [cs.LG]. URL: https://arxiv.org/abs/2002.05687.

[84] Zvi Rosen, Jessica Sidman, and Louis Theran. "Algebraic matroids in action". In: *The American Mathematical Monthly* 127.3 (2020), pp. 199–216.

[85] Luana Ruiz, Fernando Gama, and Alejandro Ribeiro. "Graph neural networks: Architectures, stability, and transferability". In: *Proceedings of the IEEE* 109.5 (2021), pp. 660–682.

[86] Steven Sam and Andrew Snowden. "GL-equivariant modules over polynomial rings in infinitely many variables". In: *Transactions of the American Mathematical Society* 368.2 (2016), pp. 1097–1158.

[87] Steven Sam and Andrew Snowden. "Gröbner methods for representations of combinatorial categories". In: *Journal of the American Mathematical Society* 30.1 (2017), pp. 159–203.

[88] Steven V Sam and Andrew Snowden. "Stability patterns in representation theory". In: *Forum of Mathematics, Sigma*. Vol. 3. Cambridge University Press. 2015, e11.

[89] Franco Scarselli et al. "The graph neural network model". In: *IEEE transactions on neural networks* 20.1 (2008), pp. 61–80.

[90] Benjamin Schmidt. "Stable Random Projection: Lightweight, General-Purpose Dimensionality Reduction for Digitized Libraries". In: *Journal of Cultural Analytics* 3.1 (Oct. 2018). DOI: 10.22148/16.025.

[91] B Schölkopf. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2002.

[92] Jean-Pierre Serre. *Linear representations of finite groups*. French. Vol. Vol. 42. Graduate Texts in Mathematics. Springer-Verlag, New York-Heidelberg, 1977, pp. x+170. ISBN: 0-387-90190-6.

[93] J Shawe-Taylor. "Kernel Methods for Pattern Analysis". In: *Cambridge University Press google schola* 2 (2004), pp. 181–201.

[94] Barry Simon. *Representations of finite and compact groups*. Vol. 10. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 1996, pp. xii+266. ISBN: 0-8218-0453-7. DOI: 10.1090/gsm/010. URL: https://doi.org/10.1090/gsm/010.

[95] Leon Simon. *Lectures on geometric measure theory*. Vol. 3. Proceedings of the Centre for Mathematical Analysis, Australian National University. Australian National University, Centre for Mathematical Analysis, Canberra, 1983, pp. vii+272. ISBN: 0-86784-429-9.

[96] Larry Smith. *Polynomial invariants of finite groups*. Vol. 6. Research Notes in Mathematics. A K Peters, Ltd., Wellesley, MA, 1995, pp. xvi+360. ISBN: 1-56881-053-9.

[97] Richard P Stanley. "Invariants of finite groups and their applications to combinatorics". In: *Bulletin of the American Mathematical Society* 1.3 (1979), pp. 475–511.

[98] Richard P. Stanley. *Combinatorics and Commutative Algebra*. Second. Vol. 41. Progress in Mathematics. Birkhäuser Boston, Inc., Boston, MA, 1996, pp. x+164. ISBN: 0-8176-3836-9.

[99] Richard P. Stanley. *Enumerative combinatorics. Vol. 2*. Vol. 62. Cambridge Studies in Advanced Mathematics. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin. Cambridge University Press, Cambridge, 1999, pp. xii+581. ISBN: 0-521-56069-1; 0-521-78987-7. DOI: 10.1017/CBO9780511609589. URL: https://doi.org/10.1017/CBO9780511609589.

[100] Behrooz Tahmasebi and Stefanie Jegelka. "The exact sample complexity gain from invariances for kernel regression". In: *Advances in Neural Information Processing Systems* 36 (2023).

[101] Nathaniel Thomas et al. "Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds". In: *arXiv preprint arXiv:1802.08219* (2018).

[102] Bartel Leendert Van Der Waerden. "On varieties in multiple-projective spaces". In: *Indagationes Mathematicae (Proceedings)*. Vol. 81. 1. Elsevier. 1978, pp. 303–312.

[103] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[104] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.

[105] Hermann Weyl. *The classical groups: their invariants and representations*. 1. Princeton university press, 1946.

[106] Hassler Whitney. "Elementary structure of real algebraic varieties". In: *Ann. of Math. (2)* 66 (1957), pp. 545–556. ISSN: 0003-486X. DOI: 10.2307/1969908. URL: https://doi.org/10.2307/1969908.

[107] Zonghan Wu et al. “A comprehensive survey on graph neural networks”. In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.

[108] Manzil Zaheer et al. “Deep sets”. In: *Advances in neural information processing systems* 30 (2017).