On the Expressive Power of Kernel Methods and the Efficiency of Kernel Learning by Association Schemes

Pravesh k. Kothari KOTPRAVESH@GMAIL.COM
Carnegie Melon University, USA

Roi Livni RLIVNI@TAUEX.TAU.AC.IL
Tel Aviv University, Israel

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

We study the expressive power of kernel methods and the algorithmic feasibility of multiple kernel learning for a special rich class of kernels.

Specifically, we define *Euclidean kernels*, a diverse class that includes most, if not all, families of kernels studied in literature such as polynomial kernels and radial basis functions. We then describe the geometric and spectral structure of this family of kernels over the hypercube (and to some extent for any compact domain). Our structural results allow us to prove meaningful limitations on the expressive power of the class as well as derive several efficient algorithms for learning kernels over different domains.

1. Introduction

Kernel methods have been a focal point of research in both theory and practice of machine learning yielding fast, practical, non-linear and easy to implement algorithms for a plethora of important problems (Cortes and Vapnik, 1995; Mika et al., 1998; Yang, 2002; Shalev-Shwartz et al., 2011; Hazan et al., 2015).

Kernels allow learning highly non linear target functions by first embedding the domain $\mathcal{X}$ into a high dimensional Hilbert space via an embedding and then learning a linear classifier in the ambient Hilbert space. Ultimately the procedure outputs a classifier of the form $x \to \langle \mathbf{w}, \phi(x) \rangle$, where ϕ captures the non-linearities and $\mathbf{w} \in \mathcal{H}$ is a linear classifier to be learnt.

The power of the method arises from the fact that while $\mathcal{H}$ could be high or even infinite dimensional, the task can be performed efficiently so long as **a)** We are given access to an efficiently computable *kernel function* k such that $k(x, y) = \langle \phi(x), \phi(y) \rangle$ and **b)** The *large margin assumption holds*: Namely, we assume a bound on the norm of the classifier to be learnt. Then, classical results for kernel methods imply an efficient learning algorithm in terms of the dimension and margin.

This opens the crucial question of designing kernels and constructing an RKHS for a given task so that the large-margin assumption holds. While there's a large body of work that gives a prescription for a good kernel in various learning settings (Shalev-Shwartz et al., 2011; Kowalczyk et al., 2001; Sadohara, 2001; Hazan et al., 2015; Heinemann et al., 2016; Cho and Saul, 2009), the task of choosing a kernel for the application at hand typically involves creative choice and guesswork.

A natural extension of kernel methods is then by allowing *Multiple Kernel Learning* (MKL). In MKL, instead of fixing a kernel, we automatically learn not only the classifier but also the embedding or kernel function.

In general, learning an optimal kernel for specific task can be ill-posed. For e.g., given a binary classification task, an optimal kernel is given by the one-dimensional embedding $\mathbf{x} \to f(\mathbf{x})$ where f is the unknown Bayes optimal hypothesis. Thus, without further qualifications, the task of learning an optimal kernel is equivalent to the task of learning an arbitrary Boolean function. A natural compromise then is to find an optimal kernel (or equivalently, an RKHS embedding) from within some rich enough class of kernels.

In this work we consider a class of kernels that contain most, if not all, explicit kernels used in practice that satisfy a simple property and we term them *Euclidean* kernels. We deter a rigorous definition to later sections, but in a nutshell, a kernel is Euclidean if it depends on the scalar product and the norm of its input. The class of Euclidean kernels capture almost all the instances of kernels considered in prior works (see, for instance Scholkopf and Smola (2001)). For example, polynomial kernels, Gaussian kernels along with Laplacian, Exponential and Sobolev space kernels (and all of their sums and products) are Euclidean.

As a class, the family of functions that can be expressed in a Euclidean kernel space, is a highly expressive and powerful class. Indeed these include, in particular, all polynomials and can thus approximate any target function to arbitrary close precision. However, standard generalization bounds and learning guarantees rely on the large margin assumption. Thus, the objective of this work is to analyze the class of functions that can be expressed through Euclidean kernels under norm constraints.

The main result of this paper shows that the class of *all* such large margin linear classifiers, over the hypercube, is learnable. In fact it can be expressed using a single specific Euclidean kernel up to some scalable deterioration in the margin. Namely, there exists a universal Euclidean kernel such that any classifier in an arbitrary Euclidean kernel belongs to the Hilbert space defined by the universal kernel, with perhaps a slightly larger norm. As a corollary we obtain both a simple and efficient algorithm to learn the class of all Euclidean kernels, as well as a useful characterization of the expressive power of Euclidean kernels which are often used in practice.

These results are then extended in two ways. First, we extend the result from the hypercube and show that, under certain further mild restrictions on the kernels, the results can be generalized from the hypercube to arbitrary compact domains in $\mathbb{R}^n$. Second, we also show that using convex relaxations and methods from MKL introduced in Lanckriet et al. (2004); Cortes et al. (2010a) one can improve the statistical sample complexity and achieve tighter generalization bounds in terms of the dimension.

Our main technical method for learning optimal Euclidean kernels is derived from our new characterization of the spectral structure of Euclidean kernels. Key to this characterization are classical results describing the spectrum of matrices of Johnson Association Scheme studied in algebraic combinatorics. Our proofs, given this connection to association schemes, are short and simple and we consider it as a feature of this work. In retrospect, the use of association schemes seems natural in studying kernels and we consider this the main technical contribution of this paper.

Studying Euclidean kernels over the hypercube may seem restrictive, as these kernels are often applied on real input features. However, as we next summarize, this course of study leads to important insights on the applicability of kernel methods:

First, these results can be extended to real inputs under some mild restrictions over the kernels to be learnt (namely, Lipschitness and no dependence on the norm of the input). Moreover, we believe that the technical tools we develop here, that is – analyzing the spectral structure of the kernel family through tools from Association Scheme and Algebraic Combinatorics, are potentially powerful for any further study of MKL in various domains.

Second, characterizing the efficiency of kernel learning also allows us to better understand the expressive power of kernel methods. Our efficient algorithm that learns the class of Euclidean kernels rules out the possibility of a general reduction from learning to the design of a Euclidean kernel (as is possible, for example, in the more general case of arbitrary kernels). Thus, we obtain that Euclidean kernels with large margin cannot express intersection of halfspaces, deep neural networks etc... Currently, hardness results demonstrate limitations for each fixed kernels, and they also demonstrate that constructing or choosing a kernel might be in general hard. In contrast, our result demonstrates lack of expressive power. Namely, that for Euclidean kernels, hardness stems not from the design of the kernel but from a deficiency in expressivness.

Moreover, as a technical contribution, our results allow an immediate transfer of lower bounds from a single fixed kernel, to a joint uniform lower bound over the whole class of Euclidean kernels. As an example we consider the problem of learning conjunctions over the hypercube – Building upon the work of Klivans and Sherstov (2007), we can show that using a single fixed kernel one cannot improve over state of the art results for agnostic learning of conjunctions. The existence of a universal kernel immediately imply that these results are true even if we allow the learner to choose the kernel in a task specific manner. Thus kernel methods, equipped with Multiple Kernel Learning techniques are still not powerful to achieve any improvement over state of the art results as long as we are restricted to Euclidean kernels.

1.1. Related Work

Kernel methods have been widely used for supervised machine learning tasks beginning with the early works of Aizerman (1964); Boser et al. (1992) and later in the context of support vector machines Cortes and Vapnik (1995). Several authors have suggested new specially designed kernels (in fact Euclidean kernels) for multiple learning tasks. For example, learning Boolean function classes such as DNFs, and decision trees Sadohara (2001); Kowalczyk et al. (2001). Also, several recent papers suggested and designed new Euclidean kernels in an attempt to mimic the computation in large, multilayer networks Cho and Saul (2009); Heinemann et al. (2016).

Limitations on the success of kernel methods and embeddings in linear half spaces have also been studied. specific kernels were studied in, Khardon and Servedio (2005), as well and more general results were obtained in Warmuth and Vishwanathan (2005); Ben-David et al. (2002). The limitations for kernel methods we are concerned with aim to capture *kernel learning*, where the the kernel is distribution dependent.

Beginning with the work of Lanckriet et al. (2004), the problem of efficiently learning a kernel has been investigated within the framework of *Multiple Kernel Learning* (MKL), where various papers have been concerned with obtaining generalization bounds (Srebro and Ben-David (2006); Cortes et al. (2010a); Ying and Campbell (2009)) as well as fast algorithms. (e.g. Sonnenburg et al. (2006); Kloft et al. (2008, 2011); Rakotomamonjy et al. (2008)). Approaches beyond learning positive sums of base kernels include *centered alignment* Cortes et al. (2010b, 2012)) and some non-linear methods Bach (2009); Cortes et al. (2009).

In contrast with most existing work, the class we study (Euclidean kernels) is not explictly described as a non-negative sum of finite base kernels and instead it is defined by properties shared by the existing explicit kernels proposed in literature. Applied directly to learning Euclidean kernels, the framework of Lanckriet et al. will lead to solving an SDP of exponential size in the underlying dimension.

2. Problem Setup and Notations

We recall the standard setting for learning with respect to arbitrary convex loss functions. We consider a concept class $\mathcal{F}$ to be learned over a bounded domain $\mathcal{X}$. In general, we will be concerned with either the hypercube $\mathcal{X}_n = \{0,1\}^n$, or the positive unit cube $\mathbb{B}_n = [0,1]^n \subseteq \mathbb{R}^n$. We will also work with individual layers of the hypercube and denote by $S_{p,n}$, the p-th layer of the hypercube i.e. $S_{p,n} = \{\mathbf{x} \in \{0,1\}^n : \sum \mathbf{x}_i = p\}$.

Given a loss function ℓ, a distribution $\mathcal{D}$ over example-label pairs from $\mathcal{X} \times \mathcal{Y}$, samples $S = \{(\mathbf{x}^{(i)}, y_i)\}_{i \leq m}$ and any hypothesis f, we denote by

$$\mathcal{L}_{\mathcal{D}}(f) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(f(\mathbf{x}), y)] \qquad \mathcal{L}_S(f) = \frac{1}{m}\sum_{i \leq m}[\ell(f(\mathbf{x}^{(i)}), y_i)]$$

the *generalization error* of f and the empirical error of f respectively. Similarly, we set $\mathrm{opt}(\mathcal{F}) := \inf_{f\in\mathcal{F}} \mathcal{L}_{\mathcal{D}}(f)$, and $\mathrm{opt}_S(\mathcal{F}) = \inf_{f\in\mathcal{F}} \mathcal{L}_S(f)$ for the optimal error on the distribution and on the sample, respectively, of the hypothesis class $\mathcal{F}$.

For convex losses, we will make the standard assumption that ℓ is L-Lipschitz w.r.t its first argument, and we will assume that ℓ is bounded by 1 at 0, namely $|\ell(0, y)| < 1$. Given a distribution $\mathcal{D}$ over example-label pairs $\mathcal{X} \times \mathcal{Y}$, the algorithm's objective is to return a hypothesis h such that $\mathcal{L}_{\mathcal{D}}(h) \leq \mathrm{opt}_{\mathcal{D}}(\mathcal{H}) + \epsilon$ with probability at least $2/3$ (the confidence can be boosted in standard ways, but we prefer not to carry extra notation.)

Euclidean RKHS Embeddings Our main result is an efficient algorithm for learning a Euclidean RKHS embedding and a linear classifier in the associated Hilbert space.

Definition 1 (Euclidean Kernel) *A kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be Euclidean if k depends solely on the norms of the input and their scalar product. Namely, there exists a function $g : \mathbb{R}^3 \to \mathbb{R}$ such that*

$$k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = g\left(\|\mathbf{x}^{(1)}\|, \|\mathbf{x}^{(2)}\|, \langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)}\rangle\right),$$

and for all $\mathbf{x} \in \mathcal{X}$ we assume that $k(\mathbf{x}, \mathbf{x}) \leq 1$.

We expand on the definition of Euclidean kernels and define Euclidean RKHS.

Definition 2 (Euclidean RKHS) *For a Hilbert space H and an embedding $\phi : \mathcal{X} \to H$, we say that (H, ϕ) is a Euclidean RKHS if the associated kernel function k is Euclidean. For a fixed domain $\mathcal{X}$, we denote the set of all Euclidean RKHS for $\mathcal{X}$ by $\mathcal{H}_{\mathcal{J}}(\mathcal{X})$.*

Given a Hilbert space H we will also denote by $H(B) = \{\mathbf{w} \in H \mid \|\mathbf{w}\|_H \leq B\}$. Finally, we define the class which is our focus of interest. This is the class of linear separators in Euclidean RKHS with a margin bound.

Definition 3 (The Class $\mathbb{C}(B)$: Euclidean Linear Separators with a Margin) *Fix the domain $\mathcal{X}$. The class of* Euclidean linear separators *with margin B is defined as the set of all linear functions in* any *Euclidean RKHS with norm at most B:*

$$\mathbb{C}(\mathcal{X}; B) = \{f_{H,\mathbf{w}} : \mathcal{X} \to \mathbb{R} \mid H \in \mathcal{H}_{\mathcal{J}}(\mathcal{X}),\ \mathbf{w} \in H(B)\}$$

where $f_{H,\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_H$.

For brevity of notation we will denote $\mathbb{C}_n(B) = \mathbb{C}(\mathcal{X}_n, B)$, and $\mathbb{C}_{p,n}(B) = \mathbb{C}(S_{p,n}, B)$, and similarly $H_{\mathbb{C}_n}$ and $\mathcal{H}_{\mathbb{C}_{p,n}}$.

Another class that will be technically useful in our proofs consists of all Euclidean kernels that can be written as direct sum of kernels over the hypercube layers:

Definition 4 (The class $\mathcal{H}_{\mathbb{C}_{\oplus_n}}$) *The class $\mathcal{H}_{\mathbb{C}_{\oplus_n}} \subseteq \mathcal{H}_{\mathbb{C}_n}$ consists of all Euclidean kernels over the hypercube that are associated with RKHS (H, ϕ) such that $H = H_1 \oplus H_2 \oplus \cdots \oplus H_n$, where each H_p is an RKHS with embedding ϕ_p such that $(H_p, \phi_p) \in \mathcal{H}_{\mathbb{C}_{p,n}}$ and such that for every $p = 1, \ldots, n$:*

$$\phi(\mathbf{x}) = (0, 0, \ldots, \underbrace{\phi_p(\mathbf{x})}_{\text{p}^{\text{th}}\text{ coordinate}}, 0, 0, \ldots, 0),\ \forall \mathbf{x} \in S_{p,n},$$

Similarly we define $\mathbb{C}_{\oplus_n}(B) = \{f_{H,\mathbf{w}} : \mathcal{X}_n \to \mathbb{R} \mid H \in \mathcal{H}_{\mathbb{C}_{\oplus_n}},\ \mathbf{w} \in H(B)\}$.

3. Main Results

We are now ready to state our main results. Our first result is concerned with the case that the domain is $\mathcal{X}_n$, the n-dimensional hypercube. We then proceed to improve on this result and give an analogue statement for $\mathbb{B}_n$, improve sample complexity in terms of dimension and derive limitations for kernel methods. The proof of theorem 5 is given in appendix A

Theorem 5 *Let $\mathcal{X}_n = \{0,1\}^n$ denote the n-th hypercube. The class of Euclidean Linear separators with a margin is learnable.*

Formally, for every $B \geq 0$ the class $\mathbb{C}_n(B)$ is efficiently learnable over $\{0,1\}^n$ w.r.t. any convex L-Lipschitz loss function ℓ with sample complexity $O\left(L\frac{n^3B^2}{\epsilon^2}\right)$.

In fact, there exists a universal Euclidean RKHS U_n, with an efficiently computable associated kernel k such that

$$\mathbb{C}_n(B) \subseteq U(n^{3/2}B).$$

The kernel k may be computed using a preprocess procedure with complexity $O(n^4)$, and querying at each iteration the value $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ for every $\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \in \mathcal{X}_n$ takes linear time in n.

3.1. Corollaries and Improvements

3.1.1. IMPROVING SAMPLE COMPLEXITY THROUGH MKL

Theorem 5 suggests an efficient algorithm for learning the class $\mathbb{C}_n(B)$ through the output of a classifier from a universal Hilbert space U_n. Since U_n need not be the optimal Hilbert space (in terms of margin) the result may lead to suboptimal guarantees.

One natural direction to improve over our result is by optimizing over the kernel of choice, as is done in the framework of MKL. In the next result, we follow the footsteps of Lanckriet et al. (2004)

and describe an algorithm that performs kernel learning, and we achieve improvement in terms of the dependency of the sample complexity in the dimension. On the other hand, the involved optimization task lead to some deterioration in the efficiency of the algorithm and dependence on accuracy. Proof for theorem 6 is provided in appendix B.

Theorem 6 *Let $\mathcal{X}_n = \{0,1\}^n$ denote the n-th hypercube. For every $B \geq 0$ the class $\mathbb{C}_n(B)$ is efficiently learnable over $\{0,1\}^n$ w.r.t. any convex L-Lipschitz loss function ℓ that is bounded by 1 at zero (i.e. $|\ell(0,y)| < 1$), with sample complexity given by $O\left(L\frac{nB^2}{\epsilon^3}\log n\right)$.*

3.1.2. LEARNING OVER REAL INPUT FEATURES

Theorem 5 shows that we can learn a Euclidean kernel over the domain $\mathcal{X}_n = \{0,1\}^n$. Kernel methods are often used in practice over real input features, therefore we give a certain extension of the aforementioned result to real input features domain. For this we need to futher restrict the family of kernels we allow to learn:

Definition 7 (Strongly Euclidean Kernels) *A Euclidean kernel k that is a kernel over $\mathbb{B}_n$ for any $n \geq 1$, is said to be L-Strongly Euclidean if $k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ can be written as:*

$$k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = g(\langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)} \rangle) \tag{1}$$

and g is L-Lipschitz over the domain $[0, n]$.

Polynomial kernels (normalized) are an example for 1-Strongly Euclidean kernels, Of course also exponential kernels and other proposed kernels that have been found useful in theory (Shalev-Shwartz et al. (2011)) are captured by this definition. Analogue to Definition 3 we define the class of strongly Euclidean separators with margin and denote them by $\mathcal{J}^s(\mathcal{X}, B)$.

Our next result state that analogously to the hypercube we can learn strongly Euclidean kernels over a compact domain. This is done through discretization and reduction to the hypercube case. A proof is provided in appendix C.

Theorem 8 *For every $B \geq 0$ the class $\mathcal{J}^s(\mathbb{B}_n, B)$ is efficiently learnable w.r.t. any convex L-Lipschitz loss function, bounded by 1 at 0 (i.e. $|\ell(0,y)| < 1$).*

3.1.3. LIMITATIONS ON THE EXPRESSIVE POWER OF EUCLIDEAN KERNELS

In this section we derive lower bounds for the expressive power of kernel methods. We consider as a test bed for our result the problem of agnostic conjunction learning. Arguably the simplest special case of the problem of agnostic learning halfspaces, is captured by the task of agnostically learning conjunctions. The state of the art algorithm for agnostic learning of conjunctions over arbitrary distributions over the hypercube is based on the work of Paturi (1992) who showed that for every conjunction (equivalently, disjunctions) over the Boolean hypercube in n dimensions, there is a polynomial of degree $\tilde{O}(\sqrt{n}\log(1/\epsilon))$ that approximates the conjunction everywhere within an error of at most ϵ. Combined with the ℓ_1-regression algorithm of Kalai et al. (2008), this yields a $2^{\tilde{O}(\sqrt{n}\log(1/\epsilon))}$-time algorithm for agnostically learning conjunctions. More recently, Gottlieb et al. (2018) obtained improved bounds for learning intersection of halfspaces with runtime is $\tilde{O}((t/\gamma^2)^{\tilde{O}(1/\gamma^2)}$ where γ is a margin parameter and t is the number of half-spaces.

One can easily show that this algorithm is easily captured via learning a *Euclidean* linear separator and thus fits into our framework (see appendix D.2). However, our next result shows that somewhat disappointingly, kernel methods cannot yield an improvement over state of the art result. This is true even if we allow the learner to choose the kernel in a distribution dependent manner. We refer the reader to appendix D for a full proof.

Theorem 9 *There exists a distribution D on $\mathcal{X}_n \subseteq \{0,1\}^n$ and a conjunction $c_I \in C_\wedge$ such that for* every *Euclidean RKHS H and $\mathbf{w} \in H$: for all $\mathbf{w}$ such that $\|\mathbf{w}\|_H = 2^{\tilde{o}(\sqrt{n})}$, we have that*

$$\mathbb{E}\left[|\langle \mathbf{w}, \phi_H(\mathbf{x})\rangle - c(\mathbf{x})|\right] > \frac{1}{6}.$$

4. Technical Overview

We next give a brief overview at a high level of our techniques:

Reduction to the hypercube layer We first observe that in order to show that the class is efficiently learnable over the hypercube, it is enough to restrict attention to the setting where the input distribution $\mathcal{D}$ is supported on $S_{p,n}$ where $S_{p,n} = \{\mathbf{x} \in \{0,1\}^n \mid \sum \mathbf{x}_i = p\}$ - the p^{th} layer of the hypercube.

Our reduction to the hypercube layer involves two steps. First we observe that the class $\mathbb{C}_n(B)$ is contained in $\mathbb{C}_{\oplus_n}(\sqrt{n}B)$. Namely we can replace every RKHS with an RKHS that can be presented as the Cartesian product over the different layers and lose at most factor $\sqrt{n}$ in term of margin. Thus, instead of learning Euclidean kernels, we restrict our attention to $\mathcal{H}_{\mathbb{C}_{\oplus_n}}$ which is expressive enough. This relaxation is exploited in both theorems 5 and 6, hence both sample complexity result carry at least a linear factor dependence on the dimensionality in terms of sample complexity.

Working in $\mathcal{H}_{\mathbb{C}_{\oplus_n}}$ simplifies our objective. Since each RKHS in $\mathcal{H}_{\mathbb{C}_{\oplus_n}}$ is a direct sum of n RKHS-s on each hypercube layer, we can focus on learning each component separately, and we derive efficient algorithms for learning $\mathbb{C}_{p,n}(B)$ for every $p = 1, \ldots, n$. Thus, in theorem 5 we construct a universal kernel over each hypercube layer. Meaning, we construct a Hilbert space U_n^p such that $\mathbb{C}_{p,n}(B)$ is contained in $U_n^p((n+1)B)$. Finally, we sum up the universal kernels to construct a universal kernel over the Cartesian product of the layers.

The approach suggested offers a simple method to learn $\mathbb{C}_n(B)$. The contruction of a universal kernel, though, causes a deterioration of an additional $O(n^2)$ factor in sample complexity. Our second approach in theorem 6 suggests an efficient algorithm for learning the optimal RKHS in each hypercube layer $\mathbb{C}_{p,n}(B)$ directly and avoid a second relaxation.

Both the results, the existence of a universal kernel and the feasibility of learning the optimal RKHS rely on the special structure of kernels in $\mathcal{H}_{\mathbb{C}_{p,n}}$ which we next describe:

Characterizing Euclidean kernel through Johnson Scheme In this part we discuss what is arguably the technical heart of our paper. Namely the application of classical results about the spectra of Johnson scheme matrices for the analysis of Euclidean kernels.

Consider any kernel over any layer $S_{p,n} = \{\mathbf{x} \in \{0,1\}^n \mid \sum \mathbf{x}_i = p\}$ of the n-hypercube - these are characterized by psd matrices indexed by elements of $S_{p,n}$ on the rows and columns. Searching over the class of all kernels thus involves searching over the space of all positive semidefinite matrices in $\binom{n}{p} \times \binom{n}{p}$ dimensions and is prohibitive in cost for $p = \omega(1)$.

The main observation behind our algorithm is that while the assumption of Euclidean kernel allows us to capture almost all the kernels used in practice, it also allows for an efficient characterization of psd matrices defining them. In particular, recall that a Euclidean kernel matrix over $S_{p,n}$ is a matrix with any (x, y)-entry being a function solely of the inner product $\langle x, y\rangle$. Such matrices are called *set-symmetric matrices* and form a *commutative algebra* called the *Johnson association scheme*: the space of such matrices is closed under addition and matrix multiplication and any two matrices in the space commute w.r.t matrix multiplication. We provide more background on the Johnson scheme in Section 5.1.

Standard linear algebra shows that a commutative algebra of matrices must share common eigenspaces. More interestingly, for our setting, the eigenspaces of set symmetric matrices have been completely figured out in the study of Johnson scheme. In particular, despite the matrices themselves being of dimension $\binom{n}{p} \times \binom{n}{p}$, they can have at most $p+1$ distinct eigenvalues! Further, there's a positive semidefinite basis of $p+1$ matrices $\{P_{p,\ell} \mid \ell \leq p\}$ for the linear space of Johnson scheme matrices with tractable expressions for eigenvalues in the $p+1$ different eigenspaces.

Construction of a universal Hilbert Space Equipped with an explicit basis of set symmetric matrices, and having a diagonalized representation for the matrices, we can explicitly construct $p+1$ kernel matrices $K_1, \ldots, K_{p+1}$ whose convex hull spans all set symmetric, positive definite and bounded by 1 matrices. These are the matrices that correspond to a Euclidean kernel. Thus we obtain an explicit characterization of the polytope of Euclidean kernels in terms of $p+1$ vertices where each kernel is a convex combination of the vertices.

Using the above construction we finally consider the direct sum Hilbert space $U_n^p = H_1 \oplus H_2, \ldots, \oplus H_{p+1}$, where the H's correspond to the kernel vertices. A direct corollary of the above characterization is that any target function in $\mathbb{C}_{n,p}(B)$ may be written in the form of $f_{H',\mathbf{w}}(x) = \sum \lambda_i \mathbf{w}_i \cdot \phi_i(x)$ where $\mathbf{w}_i \in H_i$. Standard linear algebra then show we can bound the norm of the above target function in terms of $\|\cdot\|_U$ by losing a factor of at most $\sqrt{n}$. Thus $\mathbb{C}_{n,p}(B)$ is a subset of all $\sqrt{n}B$ bounded norm vectors in U_n^p.

Improving sample complexity through MKL The above results allow us to efficiently learn the class $\mathbb{C}_n(B)$ however it may lead to suboptimal result in sample complexity. As we next discuss this can be improved by optimizing over the choice of kernel as is done in MKL.

First, as discussed before, *any* matrix of the Johnson scheme can be specified by describing the $p+1$ coefficients over the basis - and one can write down explicit expressions in these coefficients for the eigenvalues. Thus, checking PSDness reduces to just verifying $p+1$ different linear inequalities.

The above observation allows us to take the standard ℓ_2-regularized kernel SVM convex formulation and add an additional minimization over the space of coefficients that describe a Euclidean kernel. We show that the resulting modified program is convex in all its variables. Similar observations on the convexity of such programs have been made in previous works starting with the work of Lanckriet et al. (2004). Together with the tractable representation of the constraint system we obtained above, we get an efficiently solvable convex program (see theorem 23)

To achieve generalization bound, we appeal to the surprisingly strong bounds on Rademacher complexity of non-negative linear combinations of q *base kernels* due to Cortes et al. (2010a). Our generalization bounds follow a certain strenghening of the aforementioned result to $\mathbb{C}_{\oplus_n}(B)$. In turn, using the fact that the polytope of Euclidean kernels has exactly $p+1$ vertices, we can derive strong sample complexity upper bound that grows only logarithmically in the dimension n.

Limitations for Learning Conjunctions Our final application for learning kernels helps in proving bounds on the expressive power of the family of large margin linear classifiers in Euclidean RKHS. Our crucial observation relies on a result by Klivans and Sherstov (2007) who showed that for every collection of $2^{o(\sqrt{n})}$ basis functions $\eta_1, \eta_2, \ldots, \eta_M$, there's a distribution D on $\mathcal{X}_n \subseteq \{0,1\}^n$ and a conjunction c such that $\inf_{\alpha_1,\alpha_2,\ldots,\alpha_M} \mathbb{E}_{x\sim D}[|c(x) - \sum_{i\leq M} \alpha_i \eta_i(x)|] > \frac{1}{3}$. Such a result rules out any set of fixed basis functions that can linearly approximate *all* conjunctions.

Our first step in the proof translates the aforementioned result to showing that for any fixed RKHS, there's a conjunction that will require a $2^{\Omega(\sqrt{n})}$-norm linear classifier. We do that by showing that any fixed kernel with a separator with large margin will yield, via Johnson-Lindenstrauss, a small class of basis functions that can approximate any conjunction. However, this technique alone does not capture the possibility of learning the kernel Hilbert space and *then* approximate it via a linear functional in this space. In other words, while the aforementioned result restrict the expressive power of each specific kernel, it does not put limitations over $\mathbb{C}_n(B)$.

However, the existence of a universal Hilbert space demonstrates that the power of Euclidean kernels cannot exceed any limitation over a fixed kernel. Thus, building upon Klivans and Sherstov (2007) we obtain a uniform lower bound for the expressive power of $\mathbb{C}_n(B)$ and in particular $\mathbb{C}_n(\mathbb{B}_n, B)$:

5. Background

5.1. Johnson Scheme

In this section, we describe the *Johnson Scheme* (or set-symmetric) matrices that are an instance of *association schemes*, a fundamental notion in algebraic combinatorics and coding theory. We will need the classical result about the eigendecompositions of such matrices in this work. We refer the reader to the textbook and lecture notes by Godsil for further background Godsil (1993).

Definition 10 (Johnson Scheme) *Fix positive integers t, n for $t < n/2$. The Johnson scheme with parameters t, n, denoted by $\mathcal{J}_{n,t}$ is a collection of matrices with rows and columns indexed by subsets of $[n]$ of size exactly t such that for any $M \in \mathcal{J}_{n,t}$ and any $S, T \subseteq [n]$ of size t, $M(S,T)$ depends only on $|S \cap T|$.*

That is, any entry of a matrix $M \in \mathcal{J}_{n,t}$ depends only on the size of the intersection of the subsets indexing the corresponding row and column. Equivalently, we can think of the matrices in the Johnson scheme as indexed by elements of $\{0,1\}^n$ of Hamming weight exactly t with (x,y)th entry a function of the inner product $\langle x, y\rangle$. The symmetric group on n elements $\mathbb{S}_n$ acts on subsets of size t of $[n]$ by the natural renaming action and further, $|S \cap T| = |\sigma(S) \cap \sigma(T)|$ for any permutation $\sigma \in \mathbb{S}_n$. Thus, M is invariant under the action of $\mathbb{S}_n$ that renames its rows and columns as above.

It is not hard to verify that $\mathcal{J}_{n,t}$ forms a commutative algebra of matrices. A basic fact in linear algebra then says that the matrices in $\mathcal{J}_{n,t}$ must share common eigen-decomposition. The natural action of $\mathbb{S}_n$ associated above makes the task of pinning down a useful description of this eigenspaces tractable - these form classical results in algebraic combinatorics. This description of eigenspaces of the Johnson scheme will come in handy for us and in the following, we will describe the known results in a form applicable to us.

It is convenient to develop two different bases for writing the matrices in $\mathcal{J}_{n,t}$.

Definition 11 (D Basis) *For $0 \leq \ell \leq t < n$, we define the matrix $D_{n,t,\ell} \in \mathbb{R}^{\binom{[n]}{t} \times \binom{[n]}{t}}$ by: $D_{n,t,\ell}(S,T) = 1$ if $|S \cap T| = \ell$ and 0 otherwise.*

It is easy to see that every matrix in $\mathcal{J}_{n,t}$ can be written as a linear combination of the $D_{n,t,\ell}$ matrices for $0 \leq \ell \leq t$. Further, it's easy to check that any pair of D matrices commute with each other and thus so does every pair of matrices from the Johnson scheme.

While the D basis is convenient to express any matrix in the Johnson scheme, it's not particularly convenient to uncover the spectrum of the matrices. For this, we adopt a different basis, called as the P basis.

Definition 12 (P Basis) *For $0 \leq t \leq n$, let $P_{t,p} \in \mathbb{R}^{\binom{[n]}{t} \times \binom{[n]}{t}}$ be the matrix defined by: $P_{t,p}(S,T) = \binom{|S \cap T|}{\ell}$ where we think of $\binom{r}{\ell}$ for $r < \ell$ as 0. It is easy to check that $P_{t,p}$ is positive semidefinite for all t and linearly spans $\mathcal{J}_{n,t}$.*

The following translation between the P and the D basis is easy to verify.

Fact 1 (Basis Change) *Fix $r \leq t < n$. Then,*

1. $P_{p,r} = \sum_{\ell=r}^{t} \binom{\ell}{r} D_\ell$.
2. *For $0 \leq \ell \leq t$, $D_\ell = \sum_{r \geq \ell} (-1)^{r-\ell} \binom{r}{\ell} P_{p,r}$.*

The P basis helps us write down a simple expression to compute the eigenvalue of any matrix in the Johnson scheme, given that we know how to write it as a linear combination of the $P_{p,t}$ matrices. The following result is what makes this possible.

Fact 2 (Eigendecomposition of the Johnson Scheme, Eigenvalues of $P_{p,t}$) *Fix $n, t < n/2$. There are subspaces $V_0, V_1, \ldots, V_t$ such that $\mathbb{R}^{\binom{[n]}{t}} = \oplus_{i \leq t} V_i$ satisfying:*

1. *$V_0, V_1, \ldots, V_t$ are the eigenspaces of every matrix in the Johnson scheme $\mathcal{J}_{n,t}$.*
2. *For $0 \leq j \leq t$, V_j is of dimension $\binom{n}{j} - \binom{n}{j-1}$ (where we define $\binom{n}{-1} = 0$.)*
3. *Let $\lambda_j(Q)$ for $0 \leq j \leq t$ denote the eigenvalue of $Q \in \mathcal{J}_{n,t}$ on the eigenspace V_j. Then,*

$$\lambda_j(P_{p,\ell}) = \begin{cases} \binom{n-\ell-j}{t-\ell} \cdot \binom{t-j}{\ell-j} & \text{if } j \leq \ell \\ 0 & \text{otherwise.} \end{cases}$$

5.2. Kernel Method: Learning Linear Classifiers in RKHS

We now recall the standard framework for agnostically learning linear classifiers in a RKHS (see Scholkopf and Smola (2001) for a detailed overview).

Definition 13 (RKHS for $\mathcal{X}$ and Kernels) *Let H be a Hilbert space with an inner product $\langle \cdot, \cdot \rangle_H$ and the corresponding norm $\| \cdot \|_H$) along with an embedding $\phi : \mathcal{X} \to H$. H, together with the embedding ϕ is said to be an RKHS for $\mathcal{X}$.*

For any RKHS (H, ϕ), there's a unique *kernel function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is defined by the inner products of any two elements of $\mathcal{X}$ embedded in H: i.e., $k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \langle \phi(\mathbf{x}^{(1)}), \phi(\mathbf{x}^{(2)}) \rangle_H$. When $\mathcal{X}$ is finite, k is completely described by the $|\mathcal{X}| \times |\mathcal{X}|$ *kernel matrix* whose rows and columns are indexed by elements of $\mathcal{X}$ and any $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ entry being given by $k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$. A classical result in kernel theory, namely Mercer condition, states that a function k is the kernel function on an RKHS if and only if the corresponding kernel matrix K is psd. In applications, we'd also want the function k to be efficiently computable (w.r.t the natural parameters of the problem).

Consider the class of all functions of the form $f_{\mathbf{w}} : \mathcal{X} \to \mathbb{R}$ defined by $f_{\mathbf{w}}(x) = \langle \mathbf{w}, \phi(x) \rangle$ (where we suppress the subscript H when there is no room for confusion). These are linear functions in the Hilbert space extended to $\mathcal{X}$ via the embedding ϕ. The key observation underlying kernel methods is that the class of all such linear functions where the coefficient vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_H < B$ is efficiently learnable. Via standard primal-dual analysis (encapsulated by the "representer theorem"). Thus, one can show that the solution to the following convex program can be written as $h(\mathbf{x}) = \sum_{i \leq m} \alpha_i k(\mathbf{x}^{(i)}, \mathbf{x})$ for the kernel function k associated with H:

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}_S(f_{\mathbf{w}}) \\ \text{s.t.} \quad & \|\mathbf{w}\|_H < B \end{aligned}$$

The following theorem captures the error and generalization bounds one can show for solving the above convex minimization program. The sample complexity analysis is based on the SGD based method to approximately solve the convex program above presented in Shalev-Shwartz et al. (2007).

Fact 3 (See Shalev-Shwartz et al. (2007) for instance, for a proof) *Let $\mathcal{X}, \mathcal{Y}, \phi, H, k$ be as defined above. There exists an algorithm that takes as input an i.i.d sample S of size $m = m(n, \epsilon, \delta)$ and with probability at least $2/3$ over the sample, outputs a hypothesis $h : \mathcal{X} \to \mathbb{R}$ defined as $h(\mathbf{x}) = \sum_{i \leq t} \alpha_i k(\mathbf{x}^{(i)}, \mathbf{x})$, for scalars α_i satisfying $\sum_{i \leq t} |\alpha_i| \leq \frac{B^2}{\epsilon}$ that satisfies: $\mathcal{L}_D(h) \leq opt_D(H(B)) + \epsilon$. The running time and the sample complexity of the algorithm is $O(\frac{B^2}{\epsilon^2})$.*

Ackgnoweledgments The authors would like to thank the anonymous reviewers for constructive improvements and to Amit Daniely for helpful discussions and suggestions.

References

Mark A Aizerman. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

Rosa I Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 616–623. IEEE, 1999.

Francis R Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in neural information processing systems*, pages 105–112, 2009.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. Limitations of learning via embeddings in euclidean half spaces. *Journal of Machine Learning Research*, 3(Nov):441–461, 2002.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning non-linear combinations of kernels. In *Advances in neural information processing systems*, pages 396–404, 2009.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 247–254, 2010a.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Two-stage learning kernel algorithms. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 239–246, 2010b.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828, 2012.

C. D. Godsil. *Algebraic combinatorics*. Chapman and Hall Mathematics Series. Chapman & Hall, New York, 1993. ISBN 0-412-04131-6.

Lee-Ad Gottlieb, Eran Kaufman, Aryeh Kontorovich, and Gabriel Nivasch. Learning convex polytopes with margin. In *Advances in Neural Information Processing Systems*, pages 5706–5716, 2018.

Elad Hazan, Roi Livni, and Yishay Mansour. Classification with low rank and missing data. In *ICML*, pages 257–266, 2015.

Uri Heinemann, Roi Livni, Elad Eban, Gal Elidan, and Amir Globerson. Improper deep kernels. In *Artificial Intelligence and Statistics*, pages 1159–1167, 2016.

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

Roni Khardon and Rocco A Servedio. Maximum margin algorithms with boolean kernels. *Journal of Machine Learning Research*, 6(Sep):1405–1429, 2005.

Adam R Klivans and Alexander A Sherstov. A lower bound for agnostically learning disjunctions. In *International Conference on Computational Learning Theory*, pages 409–423. Springer, 2007.

Marius Kloft, Ulf Brefeld, Pavel Laskov, and Sören Sonnenburg. Non-sparse multiple kernel learning. In *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, volume 4, 2008.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12(Mar):953–997, 2011.

Adam Kowalczyk, Alexander J Smola, Robert C Williamson, et al. Kernel machines and boolean functions. In *NIPS*, pages 439–446, 2001.

Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5(Jan):27–72, 2004.

Sebastian Mika, Bernhard Schölkopf, Alexander J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *NIPS*, volume 11, pages 536–542, 1998.

Ramamohan Paturi. On the degree of polynomials that approximate symmetric boolean functions (preliminary version). In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 468–474. ACM, 1992.

Alain Rakotomamonjy, Francis R Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9(Nov):2491–2521, 2008.

Ken Sadohara. Learning of boolean functions using support vector machines. In *International Conference on Algorithmic Learning Theory*, pages 106–118. Springer, 2001.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, pages 807–814. ACM, 2007.

Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2011.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7(Jul):1531–1565, 2006.

Nathan Srebro and Shai Ben-David. Learning bounds for support vector machines with learned kernels. In *International Conference on Computational Learning Theory*, pages 169–183. Springer, 2006.

Manfred K Warmuth and SVN Vishwanathan. Leaving the span. In *International Conference on Computational Learning Theory*, pages 366–381. Springer, 2005.

Ming-Hsuan Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *Fgr*, volume 2, page 215, 2002.

Yiming Ying and Colin Campbell. Generalization bounds for learning the kernel. 2009.

Appendix A. Proof of theorem 5

This section is devoted to prove theorem 5 which we now restate.

Theorem 5 *Let $\mathcal{X}_n = \{0,1\}^n$ denote the n-th hypercube. The class of Euclidean Linear separators with a margin is learnable.*

Formally, for every $B \geq 0$ the class $\mathbb{C}_n(B)$ is efficiently learnable over $\{0,1\}^n$ w.r.t. any convex L-Lipschitz loss function ℓ with sample complexity $O\left(L\frac{n^3B^2}{\epsilon^2}\right)$.

In fact, there exists a universal Euclidean RKHS U_n, with an efficiently computable associated kernel k such that

$$\mathbb{C}_n(B) \subseteq U(n^{3/2}B).$$

The kernel k may be computed using a preprocess procedure with complexity $O(n^4)$, and querying at each iteration the value $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ for every $\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \in \mathcal{X}_n$ takes linear time in n.

The proof involves two stages. First we show a reduction from the hypercube case to the hypercube layer. Namely, we show that if we can construct a universal kernel for each layer, then we can also construct a universal kernel for the hypercube. Then we proceed to construct a universal kernel for each layer. Finally, we give a full proof at the final section appendix A.3.

A.1. Reduction to the hypercube layer $S_{p,n}$

Our first step will be to reduce the problem of constructing a universal kernel over the hypercube, to the construction of universal kernels over the hypercube layers. For this we first recall the subclass of Euclidean kernels of all kernels that can be decomposed to a Cartesian product over the layers – $\mathbb{C}_{\oplus_n}$. We next show that $\mathbb{C}(B) \subseteq \mathbb{C}_{\oplus_n}(\sqrt{n}B)$ as a corollary, constructing a universal Hilbert space for $H_{\mathbb{C}_{\oplus_n}}$ is sufficient. We then show that if we can construct a universal Hilbert space on each layer, by taking their Cartesian sum, we can construct a universal Hilbert space over $\mathbb{C}_{\oplus_n}$.

Lemma 14 *For every n we have the following inclusion:*

$$\mathbb{C}(B) \subseteq \mathbb{C}_{\oplus_n}(\sqrt{n}B).$$

Proof Let H be a Euclidean RKHS and let $H_1, \ldots, H_n$ be the projections of H onto $\left(\text{span}(\mathbf{x})_{\mathbf{x}\in S_{1,n}}, \ldots, \text{span}(\mathbf{x})_{\mathbf{x}\in S_{n,n}}\right)$ respectively. We then take the space $\bar{H} = H_1 \oplus H_2, \ldots, \oplus H_n$ and the embedding $\bar{\phi}(\mathbf{x}) = (0, 0, \ldots, \underbrace{\phi(\mathbf{x})}_{\sum \mathbf{x}_i = p}, 0, \ldots, 0)$, with associated kernel $\bar{k}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \begin{cases} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) & \|\mathbf{x}^{(i)}\| = \|\mathbf{x}^{(j)}\| \\ 0 & \text{else} \end{cases}$. Then one can show that $f_{H,\mathbf{w}} = f_{\bar{H},(\mathbf{w}_1,\ldots,\mathbf{w}_n)}$ where $\mathbf{w}_p$ is the projection of $\mathbf{w}$ onto H_p. Overall we have that

$$\|(\mathbf{w}_1, \ldots, \mathbf{w}_n)\|_{\bar{H}} = \sqrt{\sum \|\mathbf{w}_i\|_{H_i}^2} = \sqrt{\sum \|\mathbf{w}_i\|_H^2} \leq \sqrt{\sum \|\mathbf{w}\|_H^2} \leq \sqrt{n}B$$

■

Lemma 15 (Learning Euclidean Linear Separators over the Hypercube) *Fix n and let $k_1, \ldots, k_n$ be kernels associated with universal RKHS $((U_n^1, \phi_1), \ldots, (U_n^n, \phi_n))$ such that for all $B > 0$ and $p = 1, \ldots, n$:*

$$\mathbb{C}_{p,n}(B) \subseteq U_n^p(\alpha B).$$

Let k be the Euclidean kernel associated with the Hilbert space $U = U_n^1 \oplus, \ldots, \oplus U_n^n$ together with the embedding

$$\phi(x) = (0, 0, \ldots, \underbrace{\phi_t(\mathbf{x})}_{t\text{-th coordinate}}, \ldots, 0), \quad \forall \sum \mathbf{x}_i = t.$$

Then:

$$\mathbb{C}_{\oplus_n}(B) \subseteq U(\alpha B),$$

and Computing k takes $O(n + T(n))$ where $T(n)$ is the time complexity for the kernels $k_1, \ldots, k_n$.

Proof Choose $f_{H,\mathbf{w}} \in \mathbb{C}_{\oplus_n}(B)$ for some $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n) \in H_1 \oplus H_2, \ldots, \oplus H_n$. It is easy to see that for every $\mathbf{x} \in S_{p,n}$ we have that $f_{H,\mathbf{w}}(\mathbf{x}) = f_{H_p,\mathbf{w}_p}(\mathbf{x})$. Next, for each $\mathbf{w}_p$ there exists $\mathbf{v}_p \in U_n^p(\alpha\|\mathbf{w}\|_p)$ such that $f_{H_p,\mathbf{w}_p} = f_{U_n^p,\mathbf{v}_p}$. Overall we get that $f_{H,\mathbf{w}} = f_{U_n,\mathbf{v}}$ where $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$. It remains to bound the norm of $\mathbf{v}$:

$$\|\mathbf{v}\|_U = \sqrt{\sum \|\mathbf{v}_p\|_{U_n^p}^2} \leq \sqrt{\sum \alpha \|\mathbf{w}_i\|_{H_p}^2} \leq \alpha \|\mathbf{w}\|_H \leq \alpha B$$

■

A.2. Learning over $S_{p,n}$

The main result of this section shows that there exists a universal Hilbert space over a single layer of the hypercube.

Lemma 16 (Learning Euclidean Linear Separators over a Single Layer)
For fixed n and every $B \geq 0$ the class $\mathbb{C}_{p,n}(B)$ is efficiently learnable w.r.t. any convex L-Lipschitz loss function ℓ in with sample complexity $O(p^2 B^2/\epsilon^2)$.

Specifically, for every p there exists an efficiently computable kernel associated with a universal Euclidean RKHS U_n^p, such that

$$\mathbb{C}_{p,n}(B) \subseteq U_n^p\left((p+1)B\right).$$

The computation of k involves a preprocessing stage of $O(p^3)$, and then the computation of each entry $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is done in time $O(p)$.

To prove theorem 16 we begin with a direct application of classical results on eigenspaces of the matrices of the Johnson scheme to obtain a useful characterization of Euclidean kernels over a single layer of the Boolean hypercube.

Let $\eta^p \in \mathbb{R}^{p+1}$ be defined by $\eta^p_\ell = \binom{p}{\ell-1}$ for every ℓ and define $\Delta^p \in \mathbb{R}^{p+1\times p+1}$ by

$$\Delta^p_{j,\ell} = \begin{cases} \binom{n-\ell-j}{p-\ell} \cdot \binom{p-j}{\ell-j} & 0 \le j \le \ell \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

For fixed p, corresponding to the P-basis of positive definite matrices in Definition 12 we will denote by $\bar{k}_t$, the kernel over $S_{p,n}$ that is given by

$$\bar{k}_t(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \binom{\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}}{t-1}.$$

Following the discussion in section 5.1 and noting that the kernel matrix $K_{p,t} \in \mathbb{R}^{\binom{p}{t}\times\binom{p}{t}}$ equals a non-negative scaling of P_t and is thus PSD, $\bar{k}_t$ is a kernel over $S_{p,n}$.

Lemma 17 (Characterizing Euclidean Kernels) *Fix $p \le n/2$, and let $\{\mathbf{x}^{(1)}, \ldots \mathbf{x}^{(\binom{n}{p})}\} = S_{p,n}$. For a Euclidean kernel function over $S_{p,n}$ there exists an RKHS (H, ϕ) with associated kernel function k if and only if there is a vector $\beta \in \mathbb{R}^{p+1}$ such that $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sum_{t\le p+1} \beta_t \cdot \bar{k}_t(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ satisfying:*

1. $\Delta^p \beta \ge 0$ *for* $j = 0, \ldots, p$.
2. $\langle \eta^p, \beta \rangle \le 1$

Proof This is a direct application of Fact 2. Let $K \in \mathbb{R}^{S_{p,n}\times S_{p,n}}$ is defined by $K_{i,j} = k(\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle)$ for some kernel function k. Observe that K is a kernel matrix and corresponds to an RKHS (H, ϕ) if and only if K is positive semidefinite, further we have that $\|\phi(\mathbf{x}^{(i)})\| \le 1$ if and only if $K(i,i) \le 1$.

Since K is a kernel matrix of a Euclidean kernel, in particular, it is set-symmetric (an entry only depends on the inner products of the row and column index vectors) and thus, shares eigenspaces with all the matrices in the Johnson scheme and in particular with the matrices $P_{p,\ell}$ that span the space. The β_i are thus the coefficients in the P-basis for K and allow us to write down the eigenvalues of K as linear functions in β and the fixed constant eigenvalues of $P_{p,\ell}$. By Fact 2 and eq. (2), the first condition is then just the statements that all eigenvalues of K be non-negative. The second condition checks that $K(x,x)$ is bounded by one ■

The simple lemma above is surprisingly powerful. Even though the matrix K is huge (of dimensions $n^p \times n^p$ roughly), verifying that it's PSD is easy and corresponds to just checking $p+1$ different linear inequalities in $p+1$ variables. A simple corollary of theorem 17 is that we can by change of variable, describe the set of Euclidean kernels as a polytope with p vertices corresponding to kernels

Corollary 18 *For each i set $\beta^{(i)} \in \mathbb{R}^{p+1}$ such that*

$$(\Delta^p)\, \bar{\beta}^{(i)} = e_i, \quad \beta^{(i)} = \frac{\bar{\beta}^{(i)}}{\langle \eta^p, \bar{\beta}^{(i)} \rangle} \tag{3}$$

Then the kernel function $k_{p,i} = \sum \beta^{(i)}_t \bar{k}_{p,t}$ is indeed a kernel. Moreover every Euclidean kernel associated to an RKHS can be written as $k = \sum \lambda_i k_{p,i}$ where $\lambda_i \ge 0$ and $\sum \lambda_i \le 1$.

Proof Let $\mathcal{K} \in \mathbb{R}^{p+1}$ be the set of all vectors β such that $\sum \beta_t \bar{k}_t$ is a Euclidean kernel associated with an RKHS. By theorem 17, this set is convex and also $\beta^{(i)} \in \mathcal{K}$.

We next wish to show that $\beta^{(1)}, \ldots, \beta^{(p+1)}$ contain all the vertices of the set $\mathcal{K}$. Indeed, recall that invertible affine transformations preserve the set of vertices. Set $\xi^{(p)} = (\Delta^p)^{-\top} \eta^p$ and consider the following set:

$$\begin{aligned}
\Delta^p \mathcal{K} &= \{\Delta^p \beta : \Delta^p \beta \geq 0,\ \langle \eta^p, \beta \rangle \leq 1\} \\
&= \{v : v \geq 0,\ \langle \eta^p, (\Delta^p)^{-1} v \rangle \leq 1\} \\
&= \{v : v \geq 0,\ \langle (\Delta^p)^{-\top} \eta^p, v \rangle \leq 1\} \\
&= \{v : v \geq 0,\ \langle \xi^{(p)}, v \rangle \leq 1\}.
\end{aligned}$$

One can then observe that the set of vertices of the set $\Delta^p \mathcal{K}$ is given by $\{\frac{1}{\xi^{(p)}_i} e_i\}_{i=1}^{p+1}$. Finally observe that

$$\xi_i^{(p)} = \left((\Delta^p)^{-\top} \eta^p\right)_i = \langle (\Delta^p)^{-\top} \eta^p, e_i \rangle = \langle \eta^p, (\Delta^p)^{-1} e_i \rangle = \langle \eta^p, \bar{\beta}^{(i)} \rangle$$

Taking the reverse image we obtain that the set of vertices of the set $\mathcal{K}$ are given indeed by $\beta^{(1)}, \ldots, \beta^{(p+1)}$. By definition of $\mathcal{K}$ we obtain the desired result. ■

Finally we are ready to prove theorem 16.

Proof of theorem 16 First, without loss of generality we may assume $p \leq \frac{n}{2}$. If $p > \frac{n}{2}$ then we simply map $S_{p,n}$ into $S_{n-p,n}$ by having $\mathbf{x}_i \to (1 - \mathbf{x}_i)$. Note that a Euclidean kernel remains a Euclidean kernel under this mapping.

Set $k_{p,1}, \ldots, k_{p,p+1}$ be as in Corollary 18, and define for each kernel its associated RKHS $(H_1^p, \phi_1^p), \ldots, (H_{p+1}^p, \phi_{p+1}^p)$.

We next define our candidate for a universal Hilbert space and consider the Hilbert Space

$$U_n^p = H_1^p \oplus H_2^p, \cdots, \oplus H_{p+1}^p.$$

Then it is not hard to see that U_n^p forms an RKHS with the natural embedding and kernel

$$\begin{aligned}
\phi^u(\mathbf{x}) &:= \frac{1}{p+1}(\phi_1^p(\mathbf{x}), \ldots, \phi_{p+1}^p(\mathbf{x})) \\
k_p &:= \frac{1}{p+1} \sum_{i=1}^{p+1} k_{p,i}.
\end{aligned}$$

Fix $f_{H,\mathbf{w}} \in \mathbb{C}_{p,n}(B)$, the we need to show that $f_{H,\mathbf{w}} \in U_n^p((p+1)B)$.

Denote by H_S the projection of H onto $\operatorname{span}\{\phi(\mathbf{x}^{(i)})\}_{\{\mathbf{x}^{(i)} \in S_{p,n}\}}$, where ϕ is the embedding onto H. Without loss of generality we can assume that $\mathbf{w} \in H_S$. , Indeed, let $\mathbf{w}' \in H_S$ be the projection of $\mathbf{w}$ on H_S then $\|\mathbf{w}'\| < \|\mathbf{w}\| \leq B$ and we have that for all $\mathbf{x} \in S_{p,n}$: $f_{H,\mathbf{w}}(x) = \langle \mathbf{w}, \phi(x) \rangle = \langle \mathbf{w}', \phi(x) \rangle = f_{\mathbf{w}',H}(x)$.

Since $f_{H,\mathbf{w}} \in H_S$, we may write for some vector α,

$$f_{H,\mathbf{w}}(x) = \sum_{i=1}^{\binom{n}{p}} \alpha_i k(\mathbf{x}^{(i)}, x),$$

and we obtain$\|\mathbf{w}\|^2 = \sum \alpha_i \alpha_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

By theorem 18 there is a convex sum $\lambda_1, \ldots \lambda_{p+1}$ such that $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sum \lambda_t k_{t,p}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. For each $t \leq p$ set $\mathbf{w}_t = \sum \alpha_i \phi_{t,p}(\mathbf{x}^{(i)}) \in H_t^p$, and define $\mathbf{v} \in U_n^p$ to be

$$\mathbf{v} = (\lambda_1(p+1)\mathbf{w}_1, \ldots, \lambda_{p+1}(p+1)\mathbf{w}_{p+1}).$$

Our proof is done if we can show that $\|\mathbf{v}\| \leq (p+1)B$ and $f_{H,\mathbf{w}} = f_{U_n^p,\mathbf{v}}$. First we show that $f_{H,\mathbf{w}} = f_{U_n^p,\mathbf{v}}$:

$$\begin{aligned}
f_{H,\mathbf{w}}(x) &= \sum \alpha_i \sum \lambda_t k_{t,p}(\mathbf{x}^{(i)}, x) \\
&= \sum \lambda_t \sum \alpha_i k_{t,p}(\mathbf{x}^{(i)}, x) \\
&= \sum \langle \lambda_t(p+1) \cdot \mathbf{w}_t, \frac{1}{p+1}\phi_t(x) \rangle \\
&= \langle \mathbf{v}, \phi^u(x) \rangle_{U_n} \\
&= f_{\mathbf{U_n^p},\mathbf{v}}(x).
\end{aligned}$$

It remains to bound the norm of $\mathbf{v}$ by $(p+1)B$. First we obtain

$$\begin{aligned}
B^2 \geq \|\mathbf{w}\|^2 &= \sum \alpha_i \alpha_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \\
&= \sum \alpha_i \alpha_j \sum \lambda_t k_{t,p}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \\
&= \sum \lambda_t \sum \alpha_i \alpha_j k_{t,p}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \\
&= \sum \lambda_t \sum \|\mathbf{w}_t\|^2 \qquad (4)
\end{aligned}$$

Next, using eq. (4) we have that

$$\begin{aligned}
\|(\lambda_1(p+1)\mathbf{w}_1, \ldots, \lambda_{p+1}(p+1)\mathbf{w}_{p+1})\|_{U_n^p} &= \sqrt{\sum \|\lambda_t(p+1)\mathbf{w}_t\|_{H_t}^2} \\
&\leq (p+1)\sqrt{\sum \lambda_t \|\mathbf{w}_t\|^2} \leq (p+1)B.
\end{aligned}$$

Finally, we address the computational issue of computing k. Note that to describe k we need to solve the linear equations depicted in eq. (3) and solve the linear equations $\Delta\beta^{(i)} = e_i$. These equations can be solved in time $O(p^3)$. Once $\beta^{(i)}$ are known we can compute (once) the function $g(k) = \sum \beta^{(i)} \binom{k}{i}$ to compute $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = g(\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle)$.

A.3. Putting it all together

By theorem 16 there are RKHS $k_1, \ldots, k_n$ associated with RKHS $((U_n^1, \phi_1^n), \ldots, (U_n^n, \phi_n^n)$, such that $\mathbb{C}_{p,n}(B) \subseteq U_n^p((n+1)B)$. Each kernel can be computed using a preprocess stage of $O(n^3)$, overall we can compute the whole class of kernels in time $O(n^4)$, then the computation of each entry of the kernel take times $T = O(n)$. theorem 15 then says that there exists a universal RKHS (U_n, ϕ_n) such that $\mathbb{C}_{\oplus_n}(B) \subseteq U_n(nB)$.

Finally we obtain by theorem 14 that

$$\mathbb{C}_n(B) \subseteq \mathbb{C}_{\oplus_n}(\sqrt{n}B) \subseteq U_n((n+1)\sqrt{n}B).$$

The computation of each entry in the kernel is than given by $O(n + T(n)) = O(n)$.

Appendix B. Proof of theorem 6

We next restate theorem 6 which we prove in this section.

Theorem 6 *Let $\mathcal{X}_n = \{0,1\}^n$ denote the n-th hypercube. For every $B \geq 0$ the class $\mathbb{C}_n(B)$ is efficiently learnable over $\{0,1\}^n$ w.r.t. any convex L-Lipschitz loss function ℓ that is bounded by 1 at zero (i.e. $|\ell(0,y)| < 1$), with sample complexity given by $O\left(L\frac{nB^2}{\epsilon^3}\log n\right)$.*

Similar to theorem 5 our idea is to return a function $f_{H,\mathbf{w}} \in \mathbb{C}_{\oplus_n}(\sqrt{n}B) \subseteq \mathbb{C}(\sqrt{n}B)$ and reduce the problem to the single hyper cube layers. Unlike theorem 5, to learn over the layers, we will not construct a universal kernel, but instead we will apply the tools from Multiple Kernel Learning, to output a target function $f_{H,\mathbf{w}} \in \mathcal{H}_{\mathbb{C}_{p,n}}$ that optimizes over the regulerized objective. This is the procedure that helps us in shaving off a factor n in the sample complexity. Concretely, we will develop an efficient algorithm for the following optimization problem:

$$\underset{\{(\mathbf{w},H)|H\in\mathbb{C}_{p,n},\mathbf{w}\in H\}}{\text{minimize}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \mathcal{L}_S(f_{H,\mathbf{w}}) \tag{5}$$

We will then proceed to derive generalization bounds for the class $\mathbb{C}_{\oplus_n}(B)$. The final details of the proof are then summed up in appendix B.5.

B.1. Reduction to the hypercube layer $S_{p,n}$

We next set out to learn a regulerized objective over $\mathcal{H}_{\mathbb{C}_{\oplus_n}}$:

Lemma 19 *For every n, let $S = \{(\mathbf{x}^{(i)}, y_i)\}_{i=1}^m$ be a sample from $\mathcal{X}_n$. Suppose that for every sample $S \subseteq S_{p,n}$ there exists an efficient algorithm that runs in time $T(n,|S|,1/\epsilon)$ and solves the optimization problem in eq. (5) up to ϵ error. Then the following optimization problem can be solved efficiently in time $nT(n,|S|,n/\epsilon)$ to ϵ accuracy.*

$$\underset{\{(\mathbf{w},H)|H\in\mathcal{H}_{\mathbb{C}_{\oplus_n}},\mathbf{w}\in H\}}{\text{minimize}} \frac{\lambda}{2}\|\mathbf{w}\|_H^2 + \mathcal{L}_S(f_{H,\mathbf{w}}) \tag{6}$$

Proof By the structure of $\mathcal{H}_{\mathbb{C}_{\oplus_n}}$ we can write

$$\begin{aligned}
&\min_{\{(\mathbf{w},H)|H\in\mathcal{H}_{\mathbb{C}_{\oplus_n}},\mathbf{w}\in H\}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^m \ell(\langle \mathbf{w}, \phi(\mathbf{x}^{(i)})\rangle_H, y_i) \\
&= \min_{\{(\mathbf{w}_1,\ldots,\mathbf{w}_n),H_1\oplus H_2\oplus\cdots\oplus H_n)|H_p\in H_{\mathbb{C}_{p,n}},\mathbf{w}_p\in H_p\}} \frac{\lambda}{2}\sum_{p=1}^n \|\mathbf{w}_p\|_{H_p}^2 + \sum_{p=1}^n \sum_{\mathbf{x}^{(i)}\in S_{p,n}} \ell(\langle \mathbf{w}_p, \phi(\mathbf{x}^{(i)})\rangle_H, y_i) \\
&= \sum_{p=1}^n \min_{\{(\mathbf{w}_p),H_p)|H_p\in\mathcal{H}_{\mathbb{C}_{p,n}},\mathbf{w}_p\in H_p\}} \frac{\lambda}{2}\|\mathbf{w}_p\|_{H_p}^2 + \sum_{\mathbf{x}^{(i)}\in S_{p,n}} \ell(\langle \mathbf{w}_p, \phi(\mathbf{x}^{(i)})\rangle_H, y_i)
\end{aligned}$$

By assumption we can now solve each n optimization problems in the summands efficiently to obtain an optimal $\mathbf{w} = (\mathbf{w}_1,\ldots,\mathbf{w}_n)$ and an RKHS $H = H_1 \oplus \ldots \oplus H_n$. ■

B.2. Efficient algorithm for learning $\mathcal{H}_{\mathbb{C}_{p,n}}$

Our next step in the proof relies on proposing an efficient optimization algorithm over class $\mathcal{H}_{\mathbb{C}_{p,n}}$. In contrast with previous section, we will not relax the task of learning $\mathbb{C}_{p,n}(B)$ and propose an improper formulation. Instead we directly optimize over the kernel and linear separator using tools from MKL. The main result for this section is the following Lemma, which is proved at the end.

Lemma 20 *For every p, let $S = \{(\mathbf{x}^{(i)}, y_i)\}_{i=1}^m$ be a sample from $S_{p,n}$. The optimization problem in eq. (5) can be solved efficiently in time $\mathrm{poly}(\frac{1}{\lambda}, 1/\epsilon, m)$ to ϵ accuracy.*

The proof utilizes the convexity of the program that can be demonstrated by duality– this observation has been made and exploited for MKL in Lanckriet et al. (2004) and followups. The second ingredient of the proof uses the nice structure of the class of Euclidean kernels over $S_{p,n}$ which are defined by $(p+1)$ linear constraints. For a general class of kernel matrices, MKL may involve adding a semi-positiveness constraint which may turn the problem into a non-scalable SDP. Here however, the nice structure of Euclidean kernels, gives us a tractable representation over a convex sum of few base kernels.

To describe the algorithm we add further notations: First let us denote by

$$\mathcal{B}(p+1) = \{\beta \in \mathbb{R}^{p+1} \mid \beta \geq 0 \sum \beta_i \leq 1\}$$

the $p+1$ dimensional simplex and for each $\beta \in \mathcal{B}(p+1)$ we write k_β to denote the kernel $k_\beta = \sum \beta_i k_{p,i}$, where k_i are as given in theorem 18. Note that k is a kernel if and only if $k = k_\beta$ for some $\beta \in \mathcal{B}(p+1)$. We will similarly denote by (H_β, ϕ_β) the associated RKHS. We next describe the algorithm for solving eq. (7)

.

Algorithm

Input: For $m =$, m i.i.d. samples from $\mathcal{D}$ supported on $S_{p,n} \times \mathcal{Y}$: $\{(\mathbf{x}^{(i)}, y_i)\}_{i \leq m}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, convex and 1-Lipschitz.

Output: $\alpha \in \mathbb{R}^m$, $\beta \in \mathcal{B}(p+1)$ defining the linear classifier $\sum_{i=1}^m \alpha_i K_\beta(\mathbf{x}^{(i)}, \mathbf{x})$ in the Hilbert space associated with the kernel matrix K_β defined by $K_\beta = \sum_{0 \leq t \leq p} \beta_t k_{p,t}$. where $k_{p,t}$ are given by theorem 18.

Operation:

1. Let ℓ^* be the Fenchel conjugate of the loss function ℓ: $\ell^*(a,b) = \sup_x \langle a, x\rangle - \ell(x,b)$ for any a, b.
2. For $0 \leq t \leq p$ set $K_{S,t} \in \mathbb{R}^{m \times m}$, be such that $K_{S,t}(i,j) = k_{p,t}(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)})$.
3. Define $G_{S,\lambda}(\alpha, \beta) = -\frac{\lambda}{2} \sum_{0 \leq t \leq p} \beta_t (\alpha^\top K_{S,t} \alpha) - \frac{1}{m} \sum_{i=1}^m \ell^*(\alpha_i/m, y_i)$.
4. Solve
$$\inf_{\beta \in \mathcal{B}(p+1)} \sup_{\alpha \in \mathbb{R}^m} G_{S,\lambda}(\alpha, \beta).$$
5. Output α, β.

B.3. Analysis: Running Time and Correctness

We analyze the running time and correctness of the algorithm in this section.

The analysis of the algorithm is based on combining the analysis of the standard ℓ_2-regularized SVM algorithm with Lemma 18. We provide the details next.

For the running time upper bound, we only need to verify that Step 4 can be implemented efficiently. We show this next.

Lemma 21 *There is an algorithm to compute* $\inf_{\beta\in\mathcal{B}(p+1)} \sup_{\alpha\in\mathbb{R}^m} G_{S,\lambda}(\alpha,\beta)$ *in time* $\mathrm{poly}(m,n)\log(B/\epsilon)$.

Proof $G_{S,\lambda}$ is linear (and thus convex) in β for any fixed α. We will write

$$G_{S,\lambda}(\beta) = \sup_{\alpha\in\mathbb{R}^p} G_{S,\lambda}(\alpha,\beta).$$

Then $G_{S,\lambda}(\beta)$ is a supremum of convex functions and is thus convex in β. At any β, one can efficiently compute $G_{S,\lambda}(\beta)$ by solving the concave program. Thus, it is enough to minimize $G_{S,\lambda}(\beta)$ as a function of β.

To run any off-the-shelf convex minimization algorithm, we only need to verify that we can also compute a subgradient of $G_{S,\lambda}(\beta)$ at any β efficiently. It is a standard fact that if at any β the supremum of a set of convex functions is achieved by one of the constituent functions, say, $G_{S,\lambda}(\beta)$ then, any subgradient of this constituent function is a subgradient of $G_{S,\lambda}$ at β. The latter is easy to compute given the explicit expression for $G_{S,\lambda}(\alpha,\beta)$ evaluated at the fixed β and the optimizer α_1 of $G_{S,\lambda}(\beta)$ at β. ■

Next, we show why minimizing $G_{S,\lambda}$ corresponds to learning the optimal linear classifier in any regular RKHS for $S_{p,n}$.

Remark 22 *Similar facts have been observed before in the literature beginning with the influential work of Lanckriet et. al. Lanckriet et al. (2004) (See Proposition 15).*

Lemma 23 *Let* $\beta\in\mathcal{B}(p+1)$ *define a RKHS* H_β *for* $S_{p,n}$ *and given a sample* $S\subseteq S_{p,n}$ *consider the following minimization program:*

$$F_{S,\lambda}(\beta) = \inf_{\mathbf{w}\in H_\beta} \frac{\lambda}{2}\|\mathbf{w}\|^2_{H_\beta} + \frac{1}{m}\cdot\sum_{i\le m}\ell(\langle\mathbf{w},\phi_{H_\beta}(\mathbf{x}^{(i)})\rangle, y_i). \tag{7}$$

Then $F_{S,\lambda}$ *is a convex function of* β. *In fact,* $F_{S,\lambda}(\beta) = G_{S,\lambda}(\beta)$, *and if* β^*, α^* *are the solution to* $\inf\sup G_{S,\lambda}(\alpha,\beta)$ *then the* $\mathbf{w}^*$ *that minimizes the internal program in* $F_{S,\lambda}(\beta^*)$ *is given by*

$$\mathbf{w}^* = \sum \alpha_i^*\phi_{\beta^*}(\mathbf{x}^{(i)}).$$

Proof Given a sample S and fixed $\beta\in\mathcal{B}(p+1)$ denote by $K_{S,\beta}$ the kernel matrix obtained by $K_{S,\beta}(i,j) = k_\beta(\mathbf{x}^{(i)},\mathbf{x}^{(j)})$. Recall that we have similarly defined $K_{S,t}$ for $0\le t\le p+1$ in **??**. For a fixed β by Fenchel's duality we can write

$$\min_{\mathbf{w}\in H_\beta}\frac{\lambda}{2}\|\mathbf{w}\|^2_{H_\beta} + \frac{1}{m}\sum_{i=1}^m \ell(\langle\mathbf{w},\phi_{H_\beta}(\mathbf{x}^{(i)})\rangle, y_i) = \max_{\alpha\in\mathbb{R}^m} -\frac{\lambda}{2}\alpha^\top K_{S,\beta}\alpha - \frac{1}{m}\sum_{i=1}^m \ell^*(\frac{\alpha_i}{m}, y_i),$$

where $\ell^*(\alpha, y_i) = \max \alpha \cdot x - \ell(x, y_i)$ is the convex conjugate of $\frac{1}{m}\ell(x, y_i)$. Expanding $K_{S,\beta} = \sum \beta_t K_{S,t}$ we obtain:

$$F_{S,\lambda}(\beta) = \sup_{\alpha} -\frac{\lambda}{2} \sum \beta_t \left(\alpha^\top K_{S,t} \alpha\right) - \frac{1}{m} \sum_{i=1}^{m} \ell^*(\frac{\alpha_i}{m}, y_i) = \sup_{\alpha} G_{S,\lambda}(\alpha, \beta) = G_{S,\lambda}(\beta)$$

This establishes that convex program in Step 4 has the same optimum as the program in (7). Let α^*, β^* be an optimum solution to $\inf_{\beta \in \mathcal{B}(p)} \sup_{\alpha \in \mathbb{R}^m} G_{p,\lambda}$, by standard methods in SVM analysis (see Shalev-Shwartz and Ben-David (2014) for example), one can in fact express $\langle \mathbf{w}^*, \phi(\mathbf{x}) \rangle$, the optimal linear classifier yielded by the primal program in terms of α^* and β^* as: $\sum_{i \leq m} \alpha_i^* \phi_\beta(\mathbf{x}^{(i)})$. ■

Proof of theorem 20 The proof is an immediate corollary of theorem 23 and the structure of $\mathcal{H}_{\mathbb{C}_{p,n}}$ depicted in theorem 18.

B.4. Generalization bounds for the class $\mathbb{C}_{\oplus_n}(B)$

We next set out to prove the following generalization bound for learning the class $\mathbb{C}_{\oplus_n}(B)$

Lemma 24 *Let ℓ be a Lipschitz convex loss function. Given an IID sample S of size m from an unknown distribution $\mathcal{D}$ supported over $\mathcal{X}_n \times \mathcal{Y}$. With probability $2/3$ the following holds for every $f_{H,\mathbf{w}} \in \mathbb{C}_{\oplus_n}(B)$ (uniformly)*

$$\mathcal{L}_S(f_{H,\mathbf{w}}) \leq \mathcal{L}_D(f_{H,\mathbf{w}}) + O\left(B\sqrt{\frac{\log n}{S}}\right)$$

The proof relies on the following bound on the Rademacher complexity and the following standard generalization bound: Recall that the Rademacher Complexity of a class $\mathcal{H}$ over a sample $S = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ is defined as follows

$$R_m(\mathcal{H}, S) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(\mathbf{x}^{(i)})\right]$$

where $\sigma \in \{-1, 1\}^t$ are i.i.d. Rademacher distributed random variables. The following bound the generalization performance of an empirical risk minimizer with respect to the class $\mathcal{H}$. (e.g. Shalev-Shwartz and Ben-David (2014); Bartlett and Mendelson (2002))

Fact 4 *Let ℓ be a 1-Lipschitz convex loss function with $|\ell(0, y)| \leq 1$. Assume that for all $\mathbf{x}$ and $f \in \mathcal{H}$ we have $|f(\mathbf{x})| < c$. Given an IID sample from $\mathcal{D}$ supported over $\mathcal{X} \times \mathcal{Y}$, for any $f \in \mathcal{H}$ with probability at least $1 - \delta$ (over S):*

$$\mathcal{L}_{S_m}(f) \leq \mathcal{L}_{\mathcal{D}}(f) + 4 \sup_{S_m} R_m(\mathcal{H}, S) + 4c\sqrt{\frac{2 \ln 2/\delta}{m}} \tag{8}$$

Lemma 25 *For the class $\mathbb{C}_{\oplus_n}(B)$, we have the following bound on the Rademacher Complexity*

$$\mathcal{R}(\mathbb{C}_{\oplus_n(B)}, S) \leq \sqrt{\frac{2eB^2 \log n}{|S|}}$$

where e is the natural exponent $e = \lim_{n\to\infty}(1 - \frac{1}{n})^n$.

Proof For each p and sample S denote $S_p = S \cap \{\mathbf{x}^{(i)} \mid \sum \mathbf{x}^{(i)} = p\}$, and recall that for every $f_{\mathbf{w},H} \in \mathbb{C}_{\oplus_n}$ we can write $\mathbf{w} = \mathbf{w}_1 \oplus \mathbf{w}_2 \oplus \cdots \oplus_n$ where $\mathbf{w}_p \in \mathbb{C}_{p,n}(\|\mathbf{w}_p\|)$ and $\sum \|\mathbf{w}_p\|^2 \leq B$.

By definition of the Rademacher Complexity we have the following:

$$\begin{aligned}
|S| \cdot \mathcal{R}(\mathbb{C}_{\oplus_n(B)}, S) &= \mathbb{E}\left[\sup_{f_{\mathbf{w},H}\in\mathbb{C}_{\oplus_n}(B)} \sum_{\phi(\mathbf{x}^{(i)})\in S} \sigma_i f_{\mathbf{w},H}(\phi(\mathbf{x}^{(i)}))\right] \\
&= \mathbb{E}\left[\sup_{f_{\mathbf{w},H}\in\mathbb{C}_{\oplus_n}(B)} \sum_{p=1}^{n} \sum_{\phi(\mathbf{x}^{(i)})\in S_p} \sigma_i f_{\mathbf{w},H}(\phi(\mathbf{x}^{(i)}))\right] \\
&= \mathbb{E}\left[\sup_{\{\sum B_p^2 \leq B\}} \sum_{p=1}^{n} \sup_{f_{\mathbf{w}_p,H_p}\in\mathbb{C}_{p,n}(B_p)} \sum_{\phi(\mathbf{x}^{(i)})\in S_p} \sigma_i f_{\mathbf{w}_p,H_p}(\phi(\mathbf{x}^{(i)}))\right] \\
&= \mathbb{E}\left[\sup_{\{\sum B_p^2 \leq B\}} \sum_{p=1}^{n} \sup_{f_{\mathbf{w}_p,H_p}\in\mathbb{C}_{p,n}(B_p)} \langle \mathbf{w}_p; \sum_{\phi(\mathbf{x}^{(i)})\in S_p} \sigma_i \phi(\mathbf{x}^{(i)})\rangle_{H_p}\right]
\end{aligned}$$

Note that by letting $\mathbf{w}_p = \sum_{\phi(\mathbf{x}^{(i)})\in S_p} \sigma_i \phi(\mathbf{x}^{(i)})$ and by Cauchy Schwartz we have that

$$\sup_{\|\mathbf{w}_p\|\leq B_p} \langle \mathbf{w}_p; \sum_{\phi(\mathbf{x}^{(i)})\in S_p} \sigma_i \phi(\mathbf{x}^{(i)})\rangle_{H_p} = B_p \| \sum_{\phi(\mathbf{x}^{(i)})\in S_p} \sigma_i \phi(\mathbf{x}^{(i)})\|_{H_p}$$

Thus we continue with the derivation and obtain

$$\mathbb{E}\left[\sup_{\{\sum B_p^2 \leq B\}} \sum_{p=1}^{n} \sup_{f_{\mathbf{w}_p,H_p}\in\mathbb{C}_{p,n}(B_p)} \langle \mathbf{w}_p; \sum_{\phi(\mathbf{x}^{(i)})\in S_p} \sigma_i \phi(\mathbf{x}^{(i)})\rangle_{H_p}\right] = \mathbb{E}\left[\sup_{\{\sum B_p^2 \leq B\}} \sum_{p=1}^{n} B_p \sup_{H_p\in H_{\mathbb{C}_{p,n}}} \| \sum \sigma_i \phi(\mathbf{x}^{(i)})\|_{H_p}\right]$$

Again we apply C.S inequality to choose $B_p \propto \sup_{H_p\in H_{\mathbb{C}_{p,n}}} \| \sum \sigma_i \phi(\mathbf{x}^{(i)})\|_{H_p}$. and obtain

$$\begin{aligned}
|S| \cdot \mathcal{R}(\mathbb{C}_{\oplus_n(B),S}) &= \mathbb{E}\left[\sup_{\{\sum B_p^2 \leq B\}} \sum_{p=1}^{n} B_p \sup_{H_p\in H_{\mathbb{C}_{p,n}}} \| \sum \sigma_i \phi(\mathbf{x}^{(i)})\|_{H_p}\right] \\
&= \mathbb{E}\left[B\sqrt{\sum_{p=1}^{n}\left(\sup_{H_p\in H_{\mathbb{C}_{p,n}}} \| \sum \sigma_i \phi(\mathbf{x}^{(i)})\|_{H_p}\right)^2}\right] \\
&\leq B\sqrt{\sum_{p=1}^{n}\mathbb{E}\left[\left(\sup_{H_p\in H_{\mathbb{C}_{p,n}}} \| \sum \sigma_i \phi(\mathbf{x}^{(i)})\|_{H_p}\right)^2\right]} \qquad \text{Concavity of } \sqrt{}
\end{aligned}$$

We next set out to bound the quantity $\mathbb{E}\left[\left(\sup_{H_p \in H_{\mathbb{C}_{p,n}}} \|\sum \sigma_i \phi(\mathbf{x}^{(i)})\|_{H_p}\right)^2\right]$. At this step our proof follows the foots steps of Cortes et al. (2010a) who bound a similar quantity for achieving their generalization bound. First recall that $H_{\mathcal{J}_{p,n}}$, consists of all Hilbert spaces induced by taking as a kernel the convex hull of the Hilbert spaces that we will denote $H_{p,1}, \ldots, H_{p,p+1}$. One can then show that

$$\begin{aligned}
\sup_{H_p \in H_{\mathbb{C}_{p,n}}} \|\sum \sigma_i \phi(\mathbf{x}^{(i)})\|^2_{H_p} &= \sup_k \|\sum \sigma_i \phi(\mathbf{x}^{(i)})\|^2_{H_{p,k}} \\
&\leq \left(\sum_{k=1}^{p+1} \|\sum \sigma_i \phi(\mathbf{x}^{(i)})\|^{2r}_{H_{p,k}}\right)^{1/r}, \qquad \forall r \geq 1 \\
&= \left(\sum_{k=1}^{p+1} \left(\sigma^\top K_{p,k} \sigma\right)^r\right)^{1/r}
\end{aligned}$$

By concavity we then obtain

$$\mathbb{E}\left[\left(\sup_{H_p \in H_{\mathbb{C}_{p,n}}} \|\sum \sigma_i \phi(\mathbf{x}^{(i)})\|^2_{H_p}\right)\right] \leq \left(\sum_{k=1}^{p+1} \mathbb{E}\left[\left(\sigma^\top K_{p,k} \sigma\right)^r\right]\right)^{1/r}$$

By Lemma 1 in Cortes et al. (2010a), we have the following inequality

$$\mathbb{E}\left[\left(\sigma^\top K_{p,k} \sigma\right)^r\right] \leq (2r \mathrm{Tr}(K_{p,k}))^r$$

Also, since $\mathrm{Tr}(K_{p,k}) \leq |S_p|$ we obtain that for all $r \geq 1$

$$\begin{aligned}
\left(\sum_{k=1}^{p+1} \mathbb{E}\left[\left(\sigma^\top K_{p,k} \sigma\right)^r\right]\right)^{1/r} &\leq (p(2r|S_p|)^r)^{1/r} \qquad \text{Set } r = \log p \\
&= (2e(\log p |S_p|))
\end{aligned}$$

Overall we obtain that

$$\begin{aligned}
|S| \mathcal{R}(\mathbb{C}_{\oplus_n}(B), S) &\leq B \sqrt{\sum_{p=1}^{n} (2e(\log p |S_p|))} \\
&\leq B \log n \sqrt{2e \sum_{p=1}^{n} |S_p| \log n} \\
&= B \sqrt{2e |S| \log n}
\end{aligned}$$

■

B.5. Putting it all together

Consider the optimization problem in eq. (6) with $\lambda = \frac{\epsilon}{nB^2}$. Note that by assumption that ℓ is bounded by 1 at $\mathbf{w} = 0$ we have in particular that the minimizer obtain an objective smaller than 1 (which is the objective obtained by $\mathbf{w} = 0$. In particular if $f_{H^*,\mathbf{w}^*}$ minimizes eq. (6) up to $\frac{\epsilon}{2}$ error then $\|\mathbf{w}\| \leq \sqrt{\frac{n}{\epsilon}}B$, and hence $f_{H^*,\mathbf{w}^*} \in \mathbb{C}_{\oplus_n}(\sqrt{\frac{n}{\epsilon}}B)$. Also for every solution $f_{H,\mathbf{w}} \in C_{\oplus_n}(\sqrt{n}B)$, using the generalization bound in theorem 24 we obtain that w.p. $2/3$, if $S = O(nB\sqrt{\frac{\log n}{m}})$:

$$\begin{aligned}
\mathcal{L}_D(f_{H^*,\mathbf{w}^*}) &\leq \mathcal{L}_S(f_{H^*,\mathbf{w}^*}) + \epsilon \\
&\leq \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{L}_S(f_{H^*,\mathbf{w}^*}) + \epsilon \\
&\leq \min_{f_{H,\mathbf{w}} \in \mathbb{C}_{\oplus_n}(B)} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \mathcal{L}_S(f_{H,\mathbf{w}}) + \epsilon \\
&\leq \min_{f_{H,\mathbf{w}} \in \mathbb{C}_{\oplus_n}(B)} \mathcal{L}_S(f_{H,\mathbf{w}}) + 2\epsilon \\
&\leq \min_{f_{H,\mathbf{w}} \in \mathbb{C}_{\oplus_n}(B)} \mathcal{L}_D(f_{H,\mathbf{w}}) + 3\epsilon
\end{aligned}$$

Appendix C. Proof of theorem 8

Theorem 8 *For every $B \geq 0$ the class $\mathcal{J}^s(\mathbb{B}_n, B)$ is efficiently learnable w.r.t. any convex L-Lipschitz loss function, bounded by 1 at 0 (i.e. $|\ell(0, y)| < 1$).*

In this section, we extend the algorithm from previous sections to arbitrary distributions with marginals supported over the solid hypercube $[0, 1]^n \subseteq \mathbb{R}^n$. This captured kernel learning over any bounded subset of $\mathbb{R}^n$ up to rescaling.

Our idea is essentially discretization of the solid hypercube in order to view it as a hypercube in a somewhat larger dimension. We thus define the following useful object.

Definition 26 (ϵ-Hypercube Embedding) *Fix an $\epsilon > 0$. A pair of functions $\{\Psi_1, \Psi_2\} : [0, 1]^n \to \{0, 1\}^{nt}$ is said to be an ϵ-Hypercube pair embedding of the unit cube in nt dimensions, if for every $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in [0, 1]^n$: $|\langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)} \rangle - \frac{1}{t}\langle \Psi_1(\mathbf{x}^{(1)}), \Psi_2(\mathbf{x}^{(2)}) \rangle| \leq \epsilon$.*

It is easy to construct ϵ-Hypercube pair embeddings of $[0, 1]^n$. We start with an embedding of the unit interval as given by the following lemma.

Lemma 27 (ϵ-Hypercube Embedding of the Unit Interval) *Fix an $\epsilon > 0$. There exists a $t = \Theta(\log \frac{1}{\epsilon}/\epsilon^2)$ and an efficiently computable randomized maps $\psi_i : [0, 1] \to \{0, 1\}^t$ such that for any $x_1, x_2 \in [0, 1]$, $|x_1 x_2 - \frac{1}{t}\langle \psi_1(x_1), \psi_2(x_2) \rangle| \leq 2\epsilon$.*

Proof Let $\bar{x}$ for any $x \in [0, 1]$ denote the value obtained by rounding down to the nearest multiple of $\epsilon/3$. Then, notice that $|x_1 x_2 - \bar{x_1}\bar{x_2}| \leq \epsilon$. Next, for every $\bar{x}$, choose $\psi_i(\bar{x}) \in \{0, 1\}^t$ by setting $\psi_i(\bar{x})_j$ independently with probability $\bar{x}$ to be 1 and 0 otherwise. Then, notice that $\mathbb{E}[\langle \psi_1(\bar{x_1}), \psi_2(\bar{x_2}) \rangle] = t\bar{x_1}\bar{x_2}$. Further, for any fixed $\bar{x_1}, \bar{x_2}$, $\Pr[|\langle \psi_1(\bar{x_1}), \psi_2(\bar{x_2}) \rangle - t\bar{x_1}\bar{x_2}| > t\epsilon] \leq \epsilon^2/100$ for some $t = \Theta(\log \frac{1}{\epsilon}/\epsilon^2)$. By a union bound, for every $\bar{x_1}, \bar{x_2}$ in the discretized interval $[0, 1]$, we have: $|\langle \psi_1(\bar{x_1}), \psi_2(\bar{x_2}) \rangle - t\bar{x_1}\bar{x_2}| \leq t\epsilon$ with probability at least $2/3$ as required. ■

We can now use theorem 27 to obtain an ϵ-Hypercube Embedding of $[0, 1]^n$.

Lemma 28 (ϵ-Hypercube Embedding of the Unit Ball) *For any $\epsilon > 0$, there's an efficiently computable explicit randomized map that with probability at least $2/3$ outputs an ϵ-Hypercube Embedding of $[0,1]^n$, with $t = O(\frac{n^2}{\epsilon^2}\log\frac{n}{\epsilon})$.*

Proof Let ψ_i be a pair of ϵ/n-Hypercube Embedding of the unit interval in t dimensions. Let $\Psi_i : [0,1]^n \to \{0,1\}^{nt}$ be defined as $\Psi_i(\mathbf{x}) = \psi_i^{\otimes n}(\mathbf{x}_1) = (\psi_i(\mathbf{x}_1), \psi_i(\mathbf{x}_2), \ldots, \psi_i(\mathbf{x}_n))$ for every $\mathbf{x}$. Then, we claim that Ψ_i is a pair of ϵ-Hypercube embedding of the unit ball. To verify this, observe that $|\langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)}\rangle - \langle \Psi_1(\mathbf{x}^{(1)}), \Psi_2(\mathbf{x}^{(2)})\rangle| \leq \sum_{i\leq n} |\mathbf{x}_i^{(1)}\mathbf{x}_i^{(2)} - \langle \psi_1(\mathbf{x}_i^{(1)}), \psi_2(\mathbf{x}_i^{(2)})\rangle| \leq n \cdot \epsilon/n = \epsilon$. ■

We can now complete the proof of Theorem 8.

Proof [Proof of Theorem 8] We first describe our algorithm to learn the class of linear classifiers associated with L-Lipschitz continuous Euclidean kernels over the solid cube.

For every distribution $\mathcal{D}$ over $[0,1]^n \times \mathcal{Y}$, via the $\frac{\epsilon^2}{100BL}$-hypercube embedding $\Psi_2 : [0,1]^n \to \{0,1\}^{nt}$. , we obtain a distribution $\mathcal{D}^{\Psi_2}$ over $\{0,1\}^{nt} \times \mathcal{Y}$, where $t = \tilde{O}(\frac{n^2}{\epsilon^4}B^2L^2)$. By definition of $\mathcal{D}^{\Psi_2}$, we can simulate access to i.i.d. samples from $\mathcal{D}^{\Psi_2}$ given access to i.i.d. samples from $\mathcal{D}$ and use 6 to obtain an efficient algorithm with sample complexity $\tilde{O}(\frac{n^3B^4L^2}{\epsilon^7})$ to find a hypothesis h^* that has error at most $\mathrm{opt}_{\mathcal{D}^\Psi}(\mathbb{C}_{nt}(B^2/\epsilon)) + \epsilon$. We will then be done if we can show:

$$\mathrm{opt}_{\mathcal{D}}(\mathbb{C}_{\mathbf{s}_n^L}(B)) \leq \mathrm{opt}_{\mathcal{D}^\Psi}(\mathbb{C}_{nt}(B^2/\epsilon)) + O(\epsilon B^2 L).$$

Then we get the desired result by taking $\epsilon \to \frac{\epsilon}{B^2L}$. First, using fact (3) we know there exists an ϵ-approximate solution h^* such that

$$h^*(\mathbf{x}) = \sum \alpha_i k(\mathbf{x}^{(i)}, \mathbf{x}), \quad \|\alpha\|_1 \leq O(B^2/\epsilon)$$

Note that if $k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = g(\langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)}\rangle)$ is a kernel over $[0,1]^n$ then we can define over the hypercube $\{0,1\}^{nt}$ a Euclidean kernel:

$$\tilde{k}(\bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)}) = g(\frac{1}{t}\langle \bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)}\rangle)).$$

Let $\tilde{h}(\bar{\mathbf{x}}) = \sum \alpha_i \tilde{k}(\Psi_1(\mathbf{x}^{(i)}), \bar{\mathbf{x}})$. Note that $\frac{1}{t}\left\langle \Psi_1(\mathbf{x}^{(1)}), \Psi_2(\mathbf{x}^{(2)})\right\rangle < n$, hence we have by L-Lipschitness of g:

$$\|h^*(\mathbf{x}) - \tilde{h}(\Psi_2(\mathbf{x}))\| \leq \sum |\alpha_i||k(\mathbf{x}^{(i)}, \mathbf{x}) - \tilde{k}(\Psi_1(\mathbf{x}^{(i)}), \Psi_2(\mathbf{x}))| \leq O(\epsilon B^2 L)$$

■

Appendix D. Proof of theorem 9

Theorem 9 *There exists a distribution D on $\mathcal{X}_n \subseteq \{0,1\}^n$ and a conjunction $c_I \in C_\wedge$ such that for* every *Euclidean RKHS H and $\mathbf{w} \in H$: for all $\mathbf{w}$ such that $\|\mathbf{w}\|_H = 2^{\tilde{o}(\sqrt{n})}$, we have that*

$$\mathbb{E}\left[|\langle \mathbf{w}, \phi_H(\mathbf{x})\rangle - c(\mathbf{x})|\right] > \frac{1}{6}.$$

We next set out to show that no *fixed* regular kernel can uniformly approximate conjunctions, this result relies on a similar result by Klivans and Sherstov (2007), who showed that there is no linear subspace of dimension $d = 2^{o(\sqrt{n})}$ whose linear span can uniformly approximate all conjunctions. Using the Johnson Lindenstrauss style low-dimensional embedding, we prove that an existence of a kernel that uniformly approximates all conjunctions immediately implies a low dimensional RKHS embedding with this property. theorem 9 then becomes an immediate corollary of theorem 5. We let $\mathcal{C}_n = \{c_{I'}(\mathbf{x}) : c_{I'}(\mathbf{x}) = \wedge_{i \in I} \mathbf{x}_i \; I \subseteq [n]\}$ denote the class of conjunctions over the hypercube $\mathcal{X}_n$.

Our lower bound works in two steps: First we show that no *fixed* Euclidean kernel can uniformly approximate conjunctions, this result relies on a similar result by Klivans and Sherstov (2007), who showed that there is no linear subspace of dimension $d = 2^{o(\sqrt{n})}$ whose linear span can uniformly approximate all conjunctions. Using the Johnson Lindenstrauss style low-dimensional embedding, we prove that an existence of a kernel that uniformly approximates all conjunctions immediately implies a low dimensional RKHS embedding with this property. As a second step we show, using minmax argument and convexity of $F_{S,\lambda}$, that for some distribution, all Euclidean kernels must fail.

Lemma 29 *For sufficiently large n, there exists a conjunction $c(\mathbf{x}) \in \mathcal{C}_n$ and a layer $S_{p,n} = \{\mathbf{x} \in \{0,1\}^n, \sum \mathbf{x}_i = p\}$ such that for every fixed Euclidean kernels k, if $B_n = 2^{o(\sqrt{n})}$:*

$$\min_{\|\mathbf{w}\| < B_n} \max_{\mathbf{x} \in S_{p,n}} |c(\mathbf{x}) - \langle \mathbf{w}, \phi(\mathbf{x}) \rangle| > \frac{1}{6}$$

Proof Assume to the contrary. Fix p and consider $c_{I'}$ a conjunction with $|I| = v$ for some fixed $v \leq p$. We obtain that for all $\|\mathbf{x}\| = p$, there is some $\|\mathbf{u}_{I'}\| = 2^{o(\sqrt{n})}$ and k, such that:

$$|c_{I'}(\mathbf{x}) - \langle \mathbf{u}, \phi(\mathbf{x}) \rangle| \leq \frac{1}{6}.$$

Since $\|\phi(\mathbf{x})\| < 1$ and $\|\phi(\mathbf{x})\|$ depends only on p we can, by choosing $\mathbf{w}_{I'} = \|\phi(\mathbf{x})\| \cdot \mathbf{u}$, obtain a vector $\mathbf{w}_{I'}$ such that:

$$\left| c_{I'}(\mathbf{x}) - \mathbf{w}_{I'} \cdot \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|} \right| \leq \frac{1}{6}.$$

By the representer theorem, we may assume that $\mathbf{w}_I = \sum_{\|\mathbf{x}^{(i)}\|=p} \beta_i \phi(\mathbf{x}^{(i)})$ for some β. Since the kernel is Euclidean, and thus invariant under permutations, one can show that for every conjunction $c_I(\mathbf{x})$ with $|I| = v$ literals, we have that for some $\mathbf{w}_I$: [1]

$$\left| c_I(\mathbf{x}) - \mathbf{w}_I \cdot \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|} \right| \leq \frac{1}{6}. \tag{9}$$

Next, since $c_I(\mathbf{x}) \in \{-1, 1\}$, we can rewrite (9) as :

$$\frac{5}{6\|\mathbf{w}_I\|} < \frac{c_I(\mathbf{x}) \mathbf{w}_I \cdot \phi(\mathbf{x})}{\|\mathbf{w}_I\| \cdot \|\phi(\mathbf{x})\|} < \frac{7}{6\|\mathbf{w}_I\|}.$$

1. Indeed, let π be a permutation such that $\pi(I) = I'$. Then, $\mathbf{w}_I = \sum_{\|\mathbf{x}^{(i)}\|=s} \beta_i \phi(\pi_{I,I}(\mathbf{x}^{(i)}))$. Further, $\|\mathbf{w}_{I'}\| = \|\mathbf{w}_I\|$ and clearly satisfies (9), for all $\|\mathbf{x}\| = p$.

We can apply JL Lemma (see for example (Arriaga and Vempala (1999) corollary 2), onto the kernel space, to construct a projection $T : H \to \mathbb{R}^d$ where $d = O(\|\mathbf{w}\|^2 \log 1/(\delta))$ such that w.p $(1-\delta)$, a uniformly random sample from the hypercube will satisfy:

$$\frac{1}{3\|\mathbf{w}_I\|} < \frac{5}{12\|\mathbf{w}_I\|} < \frac{c_I(\mathbf{x})T(\mathbf{w}_I)\cdot T(\phi(\mathbf{x}))}{\|T(\mathbf{w}_I)\|\cdot\|T(\phi(\mathbf{x}))\|} < \frac{7}{12\|\mathbf{w}_I\|} < \frac{4}{3\|\mathbf{w}_I\|}.$$

Choosing $\delta = O(2^{-n})$ and applying union bound over all literals of size v, we obtain a subspace $d = O(2^{o(\sqrt{n})}n)$ such that for every $\mathbf{x}$ in the hypercube.

$$|c_I(\mathbf{x}) - \alpha_I \cdot T(\phi(\mathbf{x}))| < \frac{1}{3}$$

Where $\alpha_I = \|\mathbf{w}_I\|\frac{T(\mathbf{w}_I)}{\|T(\mathbf{w}_I)\|}$. Next consider the d mappings $g_i(\mathbf{x}) = (T(\phi(\mathbf{x})))_i$. We've shown that for some linear combination

$$|c_I(\mathbf{x}) - \sum \alpha_{I,i} g_i(\mathbf{x})| < \frac{1}{3}$$

Taken together we have shown that for an arbitrary size p and arbitrary number of literals v there exists a set of mapping $g_1^{(p,v)}, \ldots, g_d^{(p,v)}$ with $d = 2^{o(\sqrt{n})}$ that can approximate within $\epsilon = \frac{1}{3}$ accuracy each conjunction on samples of size p. We can extend each mapping $g^{(p,v)}$ to the whole hypercube by considering

$$g^{(p,v)}(\mathbf{x}) = \begin{cases} g^{(p,v)}(\mathbf{x}) & \|\mathbf{x}\| = p \\ 0 & \text{o.w} \end{cases}$$

Thus, taking a union of all $g^{(p,v)}$ we obtain a set of $O(n^2 2^{o(\sqrt{n})})$ mappings that can approximate each conjunction, uniformly over the hypercube. This contradicts the result of Klivans and Sherstov (2007) such that for every $2^{o(\sqrt{n})}$ dimensional subspace V, there's some conjunction which cannot be approximated by any element of V. ■

Applying a minmax argument we can restate the result as follows

Lemma 30 *For every fixed Euclidean kernel k, there exists a distribution D over $\mathcal{X}_n$ and a conjunction $c(\mathbf{x})$ so that:*

$$\min_{\|\mathbf{w}\|<B} \mathbb{E}\left[|c(\mathbf{x}) - \langle \mathbf{w}, \phi(\mathbf{x})\rangle|\right] < \frac{1}{12}$$

then $B = 2^{(\Omega(\sqrt{n}))}$.

Proof Indeed, the negation of the statement would yield that letting $\mathcal{D}$ be the family of all distributions over $\mathcal{X}_n$, then:

$$\max_{D\sim\mathcal{D}} \min_{\|\mathbf{w}\|<B} \mathbb{E}_{\mathbf{x}\sim D}|c(\mathbf{x}) - \langle \mathbf{w}, \phi(\mathbf{x})\rangle| < \frac{1}{12}$$

Exploiting the convexity of the objective in terms of $\mathbf{w}$ and D we can apply the minimax principle and obtain a contradiction to theorem 29. ■

D.1. Putting it all together

The proof is an immediate corollary of theorem 30 and the existence of a universal kernel as presented in theorem 5

D.2. Learning Conjunctions via Euclidean kernels

Given our lower bound for learning conjunctions through kernels, the first natural question is whether the upper bound $2^{\tilde{O}(\sqrt{n}\log(1/\epsilon))}$ is attainable using Euclidean kernel methods. The L_1 regression algorithm introduced in Kalai et al. (2008) employs an observation of Paturi (1992) that conjunctions can be approximated in monomial space of degree $\tilde{O}(\sqrt{n}\log(1/\epsilon))$ to learn in time $2^{\tilde{O}(\sqrt{n}\log(1/\epsilon))}$. They also make the observation, that the algorithm may be implemented by an SVM-like convex formulation – however their analysis relies on the dimension of the linear classifier being small. We show that using a slightly modified version of the polynomial kernel, standard SVM analysis can achieve the same learnability result. Such an analysis implies, in particular, thatwill succeed in achieving the same performance.

We use a similar analysis to show an improved bound under distributional assumptions. We begin by stating the main fact exploited by all algorithms for learning conjunctions

Fact 5 *Paturi (1992) For every conjunction $c(\mathbf{x})$ over the hypercube $\mathcal{X}_n$ there exists a polynomial $p_I(\mathbf{x}) = \sum_{I\subseteq\{0,1\}^n} \alpha_I \prod_{i\in I} \mathbf{x}^{(i)}$ of degree $O(n^{O(\sqrt{n}\log 1/\epsilon)})$. whose coefficient satisfy $\sum \alpha_I^2 = 2^{\tilde{O}(\sqrt{n}\log(1/\epsilon))}$.*

Theorem 31 *For every layer of the hypercube $S_{p,n}$, There is a Euclidean kernel k and an embedding $\phi : S_{p,n} \to H$ such that for every conjunction $c_I(\mathbf{x})$ there is $\|\mathbf{w}\| = 2^{\tilde{O}(\sqrt{n}\log(1/\epsilon))}$ such that*

$$|c_I(\mathbf{x}) - \langle \mathbf{w}, \phi(\mathbf{x})\rangle| < \epsilon$$

Proof Our choice of kernel is inspired by the basis kernels of the Johnson Scheme. Namely, set $T_n = O(\sqrt{n}\log 1/\epsilon)$. we choose as kernel

$$k(\mathbf{x}^{(i)}\cdot\mathbf{x}^{(j)}) = \frac{1}{N_p}\cdot\sum_{t\le T_n}\binom{\mathbf{x}^{(i)}\cdot\mathbf{x}^{(j)}}{t}$$

where $N_p = \sum_{t\le T_n}\binom{p}{t} = O(n^{\sqrt{n}\log 1/\epsilon})$. One can show that for any two points $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$

$$k(\mathbf{x}^{(i)}\cdot\mathbf{x}^{(j)}) = \frac{1}{N_p}\sum_{|I|\le T_n}\prod_{k\in I}\mathbf{x}_k^{(i)}\cdot\mathbf{x}_k^{(j)}$$

Let H be the associated Hilbert space with the kernel k, then one can observe that the kernel k embeds the sample points in the space of monomials together with the standard scalar product normalized by $\frac{1}{N_p}$. by fact 5, we know that there exists $p\in H$ whose ℓ_2 norm over the coefficient is at most $2^{\tilde{O}(\sqrt{n}\log(1/\epsilon))}$. which in turns implies that $\|p\|_H^2 = \frac{1}{N_p}|\sum\alpha_I^2| = 2^{\tilde{O}(\sqrt{n}\log(1/\epsilon))}$. ■