Dimension and Margin Bounds for Reflection-invariant Kernels *

Thorsten Doliwa, Michael Kallweit, and Hans Ulrich Simon
Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany
{thorsten.doliwa,michael.kallweit,hans.simon}@rub.de

Abstract

A kernel over the Boolean domain is said to be reflection-invariant, if its value does not change when we flip the same bit in both arguments. (Many popular kernels have this property.) We study the geometric margins that can be achieved when we represent a specific Boolean function f by a classifier that employs a reflection-invariant kernel. It turns out $\|\hat{f}\|_\infty$ is an upper bound on the average margin. Furthermore, $\|\hat{f}\|_\infty^{-1}$ is a lower bound on the smallest dimension of a feature space associated with a reflection-invariant kernel that allows for a correct representation of f. This is, to the best of our knowledge, the first paper that exhibits margin and dimension bounds for *specific functions* (as opposed to *function families*). Before this, we present our main results in a more general setting, namely for arbitrary finite domains and other notions of invariance, instead of unnecessarily restricting to the above mentioned special case.

1 Introduction

There has been much interest in margin and dimension bounds during the last decade. The simplest way to cast (most of) the existing results in this direction is offered by the notion of margin and dimension complexity associated with a given sign matrix $A \in \{-1,1\}^{m\times n}$. A linear arrangement, given by unit vectors $u_1,\ldots,u_m; v_1,\ldots,v_n$ (taken from an inner product space), is said to represent A if, for all $i = 1,\ldots,m$ and $j = 1,\ldots,n$, $A_{i,j} = \mathrm{sign}(\langle u_i, v_j\rangle)$. The dimension complexity of A is the smallest dimension of an inner product space that allows for such a representation. The margin complexity is obtained similarly by looking for the linear arrangement that leads to the maximum average margin (or, alternatively, to the maximum margin that can be guaranteed for all choices of i and j). Applying counting arguments, Ben-David, Eiron, and Simon [1] have shown that, loosely speaking, an overwhelming majority of sign matrices of small VC-dimension do not allow for a linear arrangement whose margin or dimension is significantly better than what can be guaranteed in a trivial fashion. Starting with Forster's celebrated exponential lower bound on the dimension complexity of the Hadamard-matrix [4], there has been a series of papers [5, 6, 10, 7, 13, 15] presenting (increasingly powerful) techniques for deriving upper margin bounds or lower dimension bounds on the complexity of sign matrices.

Note that a sign matrix represents a *family* of Boolean functions, one Boolean function per column say. The lack of non-trivial margin or dimension bounds for a *specific* Boolean function has a simple explanation: a specific function $f(x)$ can always trivially be represented in a 1-dimensional space with geometric margin 1 by mapping an instance $x \in \{-1,1\}^n$ to $f(x) \in \{-1,1\}$. The corresponding kernel would map a pair (x,x') of instances to 1 if $f(x) = f(x')$, and to -1 otherwise. Clearly, the 1-dimensional "linear arrangement" for f does not say much about the ability of kernel-based large margin classifier systems to "learn" f because we would need to know f perfectly prior to the choice of the kernel. (If we had this knowledge, there would be nothing to learn anymore.) Nevertheless, this discussion shows that one cannot expect non-trivial margin or dimension bounds for *specific functions* that hold *uniformly for all kernels*.

In this paper, we introduce the concept of reflection-invariant distributed functions over the Boolean domain, a special case of which are the reflection-invariant kernels. It is easy to see that many pop-

*This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. This work was furthermore supported by the Deutsche Forschungsgemeinschaft Grant SI 498/8-1.

ular kernels actually are reflection-invariant. We then derive non-trivial margin and dimension bounds for specific Boolean functions that are valid for all linear arrangements resulting from reflection-invariant kernels. Interestingly, the bounds for a function f can be expressed in terms of f's Fourier-spectrum. As always, $\|\hat{f}\|_\infty$ denotes the largest absolute value found in the spectrum of f's Fourier-coefficients. We show that $\|\hat{f}\|_\infty$ is an upper bound on the largest possible average margin, and $\|\hat{f}\|_\infty^{-1}$ is a lower bound on the smallest possible dimension. Also note that there is an efficient randomized algorithm which will determine (with high probability) all Fourier-coefficients above a threshold $\theta > 0$, namely the KM-algorithm [20]. So an estimation to our bounds can be precomputed for a function before the actual learning process.

The remainder of this paper is structured as follows. In Section 2, we introduce some notation and recall some facts about the Fourier-expansion of real-valued functions over the Boolean domain and kernel-bases classification. In Section 3, we present our main results for arbitrary finite domains and a general notions of invariance. Besides we mention a connection to a recent paper by Haasdonk and Burkhardt [8]. In Section 4 we introduce the concept of rotation-invariance and mention some connections between the Fourier-expansion over an arbitrary finite Abelian group and the spectral decomposition of such functions. In Section 5, we consider distributed functions over the Boolean domain and the concept of reflection-invariance, which is simply rotation-invariance over boolean domain. From a practical point of view reflection-invariance is the more important notion because many popular kernels have this property. Section 6 is devoted to the analysis of reflection-invariant kernels. Some open problems are finally mentioned in Section 7.

2 Definitions and Notations

First of all we assume some familiarity with basics in matrix and learning theory. For example, notions like

- singular values, eigenvalues, spectral norm
- kernels, feature map, Reproducing Kernel Hilbert Space

are assumed as known (although we shall occasionally refresh the readers memory in the course of the paper). Some central definitions and facts concerning

- linear arrangements representing a given Boolean function,
- margin and dimension associated with such a linear arrangement,
- and Fourier expansion over arbitrary finite Abelian groups

will be given later in the paper at the place where it is required. In the following we want to declare some notations and recall the Fourier-expansion of real-valued functions as well as the notion of margin in kernel-based classification.

2.1 Preliminaries and Fourier-expansion

Throughout the paper, δ denotes the Kronecker-symbol, i.e., $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ otherwise. For two n-dimensional vectors x, y, we define $x \circ y$ to be the vector obtained by multiplying x and y componentwise, i.e., $(x \circ y)_i := x_i y_i$ for $i = 1, \ldots, n$. The n-dimensional "all-ones vector" is given by

$$\vec{e} = (1, \ldots, 1) \ .$$

The vector with 1 in component k and zeros elsewhere is denoted as $\vec{e_k}$. The n-dimensional "reflection-vector" is given by

$$\vec{r} = (-1, \ldots, -1) \ .$$

The vector with -1 in component k and ones elsewhere is denoted as $\vec{r_k}$. Note that, for every $x \in \mathbb{R}^n$, $x \circ \vec{r} = -x$, whereas vector $x \circ \vec{r_k}$ coincides with x in all coordinates except for coordinate k which equals $-x_k$. We consider real-valued functions over a finite domain D, i.e., functions of the form $f : D \to \mathbb{R}$. These functions form a $|D|$-dimensional vector space over the reals.

We are particularly interested in functions over the Boolean domain $D = \{-1, 1\}^n$ so that $d = |D| = 2^n$. In this case, the parity functions

$$\chi_z(x) := \prod_{z_i = -1} x_i \ , \ z \in D$$

form an orthonormal basis for the vector space of boolean functions over D, that can be equipped with the inner product

$$\langle f, g \rangle := 2^{-n} \cdot \sum_{x \in D} f(x) \cdot g(x) \ .$$

Thus, every function $f : D \to \mathbb{R}$ can be written in the form

$$f(x) = \sum_{z \in D} \hat{f}(z) \cdot \chi_z(x) \qquad (1)$$

where

$$\hat{f}(z) := \langle f, \chi_z \rangle = 2^{-n} \cdot \sum_{y \in D} f(y) \cdot \chi_z(y) \ .$$

As usual, equation (1) is referred to as the *Fourier expansion of* f, and $\hat{f}(z)$ is called the *Fourier-coefficient of* f *at* z. We briefly note that parity functions satisfy the equation

$$\chi_z(x) = (-1)^{|\{i : x_i = -1, z_i = -1\}|} = \chi_x(z) \ .$$

The matrix $H = (H_{x,z})_{x,z \in D}$ given by $H_{x,z} = \chi_z(x)$ is called the $(2^n \times 2^n)$-*Walsh-matrix*. The afore-mentioned properties of the parity functions imply that H is symmetric and satisfies

$$H \cdot H^\top = H \cdot H = 2^n \cdot I$$

where I denotes the identity matrix.

2.2 Kernel-based Classification

Let K be a valid kernel-function, Φ_K the feature map and $\langle \cdot, \cdot \rangle_K$ the inner product that represent K in the Reproducing Kernel Hilbert Space, and let $\| \cdot \|_K$ be the norm induced by $\langle \cdot, \cdot \rangle_K$.[1] Then, Φ satisfies

$$\forall x, y \in D: \; K(x,y) = \langle \Phi(x), \Phi(y) \rangle \ . \qquad (2)$$

With every "dual vector" $\alpha : D \to \mathbb{R}$, we associate the "weight vector"

$$w(\alpha) := \sum_{x \in D} \alpha(x)\Phi(x) \ . \qquad (3)$$

In the context of "large margin classification", α is considered as a classifier that assigns the label $\mathrm{sign}(\langle w(\alpha), \Phi(x) \rangle)$ to input x. Consider f as a target function for a binary classification task. Then, a negative sign of $f(x) \cdot \langle w(\alpha), \Phi(x) \rangle$ indicates a "classification error" on x. So this expression should be *positive* and it is intuitively even better when is leads to a *large* positive value. Thus, the following number, called the *(geometric) margin achieved by α on x w.r.t. target function f and kernel K*, is of interest:

$$\mu_K(f|\alpha, x) := \frac{f(x) \cdot \langle w(\alpha), \Phi(x) \rangle}{\|w(\alpha)\| \cdot \|\Phi(x)\|} \qquad (4)$$

By averaging over all $x \in D$, we obtain the function

$$\overline{\mu}_K(f|\alpha) := 2^{-n} \sum_{x \in D} \mu_K(f|\alpha, x) \ .$$

Focusing on the margin that is guaranteed for every $x \in D$, we should consider the function

$$\mu_K(f|\alpha) := \min_{x \in D} \mu_K(f|\alpha, x) \ .$$

By taking the supremum over all $\alpha : D \to \mathbb{R}$, we get the respective parameters of a large margin classifier employing kernel function K:

$$\begin{aligned} \overline{\mu}_K(f) &:= \sup_{\alpha : D \to \mathbb{R}} \overline{\mu}_K(f|\alpha) \\ \mu_K(f) &:= \sup_{\alpha : D \to \mathbb{R}} \mu_K(f|\alpha) \end{aligned}$$

Finally, taking the supremum ranging over all kernels of a given specific class $\mathcal{C}$, we get the respective parameters of a best possible large margin classifier among $\mathcal{C}$:

$$\begin{aligned} \overline{\mu}_{\mathcal{C}}(f) &:= \sup_K \overline{\mu}_K(f) \\ \mu_{\mathcal{C}}(f) &:= \sup_K \mu_K(f) \end{aligned}$$

We briefly note that, obviously, the guaranteed margin is upper bounded by the average margin:

$$\begin{aligned} \mu_K(f|\alpha) &\le \overline{\mu}_K(f|\alpha) \\ \mu_K(f) &\le \overline{\mu}_K(f) \\ \mu_{\mathcal{C}}(f) &\le \overline{\mu}_{\mathcal{C}}(f) \end{aligned}$$

[1] In the sequel, we drop index K unless we would like to stress the dependence on K.

3 A General Notion of Invariance

Throughout this section, D denotes an arbitrary finite domain, $\mathcal{S}(D)$ is the group of permutations over D, and $\mathcal{G} \le \mathcal{S}(D)$ is an arbitrary but fixed subgroup. The considerations of the later sections will correspond to the special case where $D = \{-1, 1\}^n$ and $\mathcal{G}$ contains all permutations (transpositions actually) of the form $x \mapsto x \circ a$ for $a \in \{-1, 1\}^n$.

A distributed function $f(x, y)$ is said to be $\mathcal{G}$-*invariant* if, for all $x, y \in D$ and every $\sigma \in \mathcal{G}$, the following holds:

$$f(\sigma(x), \sigma(y)) = f(x, y)$$

Clearly, a function of the form $g(f_1, \ldots, f_d)$ for $\mathcal{G}$-invariant functions $f_1, \ldots, f_d$ is $\mathcal{G}$-invariant itself. More interesting is the the following result:

Lemma 1 *$\mathcal{G}$-invariant distributed functions over a finite domain D are closed under the usual matrix product and under the tensor-product of matrices. More precisely, let $F(x, y)$ and $G(x, y)$ be two $\mathcal{G}$-invariant distributed functions (here viewed as matrices). Then, the functions $(F \cdot G)(x, y)$ is $\mathcal{G}$–invariant and the function $(F \otimes G)[(u, x), (v, y)]$ is invariant over $\mathcal{G} \times \mathcal{G}$ (as subgroup of $\mathcal{S}(D) \times \mathcal{S}(D)$).*

Proof: Consider first the function $(F \cdot G)(x, y)$. Let $x, y \in D$ and $\sigma \in \mathcal{G}$ be arbitrary but fixed. The following calculation shows that it is $\mathcal{G}$-invariant:

$$\begin{aligned} (F \cdot G)_{\sigma(x), \sigma(y)} &= \sum_{z \in D} F_{\sigma(x), z} \cdot G_{z, \sigma(y)} \\ &= \sum_{z \in D} F_{x, \sigma^{-1}(z)} \cdot G_{\sigma^{-1}(z), y} \\ &= \sum_{z \in D} F_{x,z} \cdot G_{z,y} \\ &= (F \cdot G)_{x,y} \end{aligned}$$

Now consider the tensor-product $(F \otimes G)[(u, x), (v, y)]$, which is a distributed function over $D \times D$, i.e., a function over domain $(D \times D) \times (D \times D)$. The following calculation shows that it is $(\mathcal{G} \times \mathcal{G})$-invariant:

$$\begin{aligned} &(F \otimes G)[(\sigma(u), \tau(x)), (\sigma(v), \tau(y))] = \\ &F(\sigma(u), \sigma(v)) \cdot G(\tau(x), \tau(y)) = \\ &F(u, v) \cdot G(x, y) = \\ &(F \otimes G)[(u, x), (v, y)] \end{aligned}$$

$\blacksquare$

Our general notion of invariance has been considered before by Haasdonk and Burkhardt [8], where it is named “simultaneous invariance”. These authors were interested in the construction of kernels whose invariance properties reflect symmetries within the data of a particular application. In this section, we look at invariant kernels from a different angle. We shall reveal non-trivial margin and dimension bounds for *specific* functions that hold uniformly for a family of invariant kernels. More precisely, if $f : D \to \{-1, 1\}$ is a function on domain D and $\mathcal{G}$ is a subgroup of $\mathcal{S}(D)$, then the largest average (or largest guaranteed, resp.) margin that can be obtained when f is represented by a $\mathcal{G}$-invariant kernel is upper-bounded by the largest average (or largest guaranteed, resp.) margin that can be obtained for the family

$$\mathcal{G}_f := \{f_\sigma : \ \sigma \in \mathcal{G}\}$$

where

$$f_\sigma(x) := f(\sigma(x)) \ .$$

Since there are classical margin bounds that apply to the family $\mathcal{G}_f$, we obtain corresponding bounds that apply to the single function f. An analogous remark holds for dimension bounds. Details follow.

Assume that $K(x, y)$ is a kernel that is $\mathcal{G}$-invariant and consider the feature map $\Phi = \Phi_K$ that represents K in the Reproducing Kernel Hilbert Space. Then, for all $x, y \in D$ and every $\sigma \in \mathcal{G}$, Φ satisfies

$$\langle \Phi(\sigma(x), \Phi(\sigma(y)\rangle = \langle \Phi(x), \Phi(y)\rangle \ . \qquad (5)$$

Lemma 2 *If kernel K is $\mathcal{G}$-invariant, then the following holds for every $x \in D$ and every $\sigma \in \mathcal{G}$:*

$$\begin{aligned} \|\Phi_K(\sigma(x))\|_K &= \|\Phi_K(x)\|_K \\ \|w(\alpha)\|_K &= \|w(\alpha_\sigma)\|_K \end{aligned}$$

In other words, the norm $\|\cdot\|_K$ is constant on feature vectors of instances taken from the same orbit

$$x^{\mathcal{G}} := \{\sigma(x) : \sigma \in \mathcal{G}\}$$

and it assigns the same value to all dual vectors from the set

$$\{w(\alpha_\sigma) : \ \sigma \in \mathcal{G}\} \ .$$

Proof: Let $\Phi = \Phi_K$, $\|\cdot\| = \|\cdot\|_K$, and $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_K$. Clearly, $\|\Phi(\sigma(x))\| = \|\Phi(x)\|$ because of

$$\begin{aligned} \|\Phi(\sigma(x))\|^2 &= \langle \Phi(\sigma(x)), \Phi(\sigma(x))\rangle \\ &\overset{(5)}{=} \langle \Phi(x), \Phi(x)\rangle \\ &= \|\Phi(x)\|^2 \ . \end{aligned}$$

As for the second statement, see the following calculation:

$$\begin{aligned} \|w(\alpha_\sigma)\|^2 &= \langle w(\alpha_\sigma), w(\alpha_\sigma)\rangle \\ &\overset{(3)}{=} \left\langle \sum_{x\in D} \alpha_\sigma(x)\Phi(x), \sum_{y\in D} \alpha_\sigma(y)\Phi(y) \right\rangle \\ &= \sum_{x,y\in D} \alpha(\sigma(x))\alpha(\sigma(y))\langle \Phi(x), \Phi(y)\rangle \\ &= \sum_{x,y\in D} \alpha(x)\alpha(y)\big\langle \Phi(\sigma^{-1}(x)), \Phi(\sigma^{-1}(y))\big\rangle \\ &\overset{(5)}{=} \sum_{x,y\in D} \alpha(x)\alpha(y)\langle \Phi(x), \Phi(y)\rangle \\ &= \|w(\alpha)\|^2 \end{aligned}$$

$\blacksquare$

Lemma 3 *For every $\mathcal{G}$-invariant kernel K, and every choice of $f : D \to \{-1, 1\}$, $x \in D$, $\sigma \in \mathcal{G}$, and $\alpha : D \to \mathbb{R}$, the following holds:*

$$\mu_K(f_\sigma|\alpha_\sigma, x) = \mu_K(f|\alpha, \sigma(x))$$

Proof: The proof starts as follows:

$$\begin{aligned} & f_\sigma(x) \cdot \langle w(\alpha_\sigma), \Phi(x)\rangle \overset{(3)}{=} \\ & \quad f_\sigma(x)\left\langle \sum_{y\in D} \alpha_\sigma(y)\Phi(y), \Phi(x) \right\rangle = \\ & \quad f(\sigma(x)) \sum_{y\in D} \alpha(\sigma(y))\langle \Phi(y), \Phi(x)\rangle \overset{(5)}{=} \\ & \quad f(\sigma(x)) \sum_{y\in D} \alpha(\sigma(y))\langle \Phi(\sigma(y)), \Phi(\sigma(x))\rangle = \\ & \quad f(\sigma(x))\left\langle \sum_{y\in D} \alpha(\sigma(y))\Phi(\sigma(y)), \Phi(\sigma(x)) \right\rangle = \\ & \quad f(\sigma(x))\left\langle \sum_{y\in D} \alpha(y)\Phi(y), \Phi(\sigma(x)) \right\rangle = \\ & \quad f(\sigma(x))\langle w(\alpha), \Phi(\sigma(x))\rangle \end{aligned}$$

Using this calculation in combination with Lemma 2, the proof is easy to accomplish:

$$\begin{aligned} \mu_K(f_\sigma|\alpha_\sigma, x) &\overset{(4)}{=} \frac{f_\sigma(x) \cdot \langle w(\alpha_\sigma), \Phi(x)\rangle}{\|w(\alpha_\sigma)\| \cdot \|\Phi(x)\|} \\ &= \frac{f(\sigma(x)) \cdot \langle w(\alpha), \Phi(\sigma(x))\rangle}{\|w(\alpha)\| \cdot \|\Phi(\sigma(x))\|} \\ &\overset{(4)}{=} \mu_K(f|\alpha, \sigma(x)) \end{aligned}$$

$\blacksquare$

Corollary 4 *For every $\mathcal{G}$-invariant kernel K, and every choice of $f : D \to \{-1,1\}$, $\sigma \in \mathcal{G}$, and $\alpha : D \to \mathbb{R}$, the following holds:*

$$\begin{aligned}\overline{\mu}_K(f_\sigma|\alpha_\sigma) &= \overline{\mu}_K(f|\alpha)\\ \mu_K(f_\sigma|\alpha_\sigma) &= \mu_K(f|\alpha)\\ \overline{\mu}_K(f_\sigma) &= \overline{\mu}_K(f)\\ \mu_K(f_\sigma) &= \mu_K(f)\\ \overline{\mu}_{\mathcal{G}}(f_\sigma) &= \overline{\mu}_{\mathcal{G}}(f)\\ \mu_{\mathcal{G}}(f_\sigma) &= \mu_{\mathcal{G}}(f)\end{aligned}$$

Note that the last two equations in Corollary 4 basically say that the largest (average or guaranteed) margin that can be achieved for a function f by a large margin classifier is invariant under $\mathcal{G}$ (provided that the underlying kernel is $\mathcal{G}$ -invariant).

Let $M \in \{-1,1\}^{r\times s}$ be a sign matrix. Consider a linear arrangement $\mathcal{A}$ given by unit vectors $u_1,\ldots,u_r; v_1,\ldots,v_s \in \mathbb{R}^d$. The *average margin achieved by this arrangement for sign matrix M* is defined as follows:

$$\overline{\mu}(M|\mathcal{A}) := \frac{1}{rs}\cdot\sum_{i=1}^{r}\sum_{j=1}^{s} M_{i,j}\langle u_i, v_j\rangle$$

The *largest average margin that can be achieved for sign matrix M by any linear arrangement* is then given by

$$\overline{\mu}(M) := \sup_{\mathcal{A}} \overline{\mu}(M|\mathcal{A}) \ ,$$

where the supremum ranges over all linear arrangements $\mathcal{A}$ for M. Forster and Simon [7] have shown that, for every $M \in \mathbb{R}^{r\times s}$, every $d \geq 1$, and every choice of unit vectors $u_1,\ldots,u_r; v_1,\ldots,v_s$ in a real inner-product space, the following holds:

$$\sum_{i=1}^{r}\sum_{j=1}^{s} M_{i,j}\langle u_i, v_j\rangle \leq \sqrt{rs}\|M\| \ . \tag{6}$$

From that, we conclude that

$$\overline{\mu}(M) \leq \frac{\|M\|}{\sqrt{rs}} \ .$$

Consider the sign matrix $M^{f,\mathcal{G}}$ given by

$$M^{f,\mathcal{G}}_{x,\sigma} := f_\sigma(x) \ .$$

In combination with Corollary 4, we arrive at the following

Theorem 5 *Let D be a finite domain, and let $\mathcal{G}$ be a subgroup of $\mathcal{S}(D)$. Then, every function $f : D \to \{-1,1\}$ satisfies*

$$\overline{\mu}_{\mathcal{G}}(f) \leq \frac{\|M^{f,\mathcal{G}}\|}{\sqrt{|D|\cdot|\mathcal{G}|}} \ .$$

In other words, no large margin classifier that employs a $\mathcal{G}$-invariant kernel can achieve an average margin for f which exceeds $\frac{\|M^{f,\mathcal{G}}\|}{\sqrt{|D|\cdot|\mathcal{G}|}}$.

As our input space D is finite, we can assume without loss of generality that the Reproducing Kernel Hilbert Space for a kernel K on D coincides with $\mathbb{R}^{d(K)}$ for some suitable $1 \leq d(K) \leq |D|$. We say that $\alpha : D \to \mathbb{R}$ *represents target function f correctly w.r.t. kernel K* if

$$\forall x \in D : \mu_K(f|\alpha, x) > 0 \ .$$

Corollary 6 *Let $d_{\mathcal{G}}(f)$ denote the smallest dimension of a feature space associated with a $\mathcal{G}$-invariant kernel K that allows for a correct representation of f. Then,*

$$d_{\mathcal{G}}(f) \geq \frac{\sqrt{|D|\cdot|\mathcal{G}|}}{\|M^{f,\mathcal{G}}\|} \ .$$

Proof: According to Lemma 3, a kernel that allows for a correct representation of f allows also for a correct representation of all f_σ. According to a result by Forster [4], the corresponding feature space must have dimension at least $\sqrt{|D|\cdot|\mathcal{G}|}/\|M^{f,\mathcal{G}}\|$. ■

Corollary 6 can be strengthened slightly:

Corollary 7 *Let σ_i denote the i-th singular value of $M^{f,\mathcal{G}}$, where $\sigma_1,\sigma_2,\ldots$ are in decreasing order. Then, $d_{\mathcal{G}}(f)$ satisfies the following lower bound:*

$$d_{\mathcal{G}}(f)\cdot\sum_{i=1}^{d_{\mathcal{G}}(f)}\sigma_i^2 \geq 1 \tag{7}$$

Proof: Let $A \in \{-1,1\}^{r\times s}$ be a matrix whose columns are viewed as binary functions $f_1,\ldots,f_s$. It has been shown by Forster and Simon [7] that the dimension d of a feature space which allows for a correct representation of $f_1,\ldots,f_s$ satisfies

$$d\cdot\sum_{i=1}^{d}\sigma_i^2(A) \geq rs \ . \tag{8}$$

This trivially implies (7). ■

4 Rotation-invariant Functions

We recall the general notion of a Fourier-expansion over a finite Abelian group in Section 4.1. Some facts about distributed functions over a finite Abelian group are derived in Section 4.2. In Section 4.3, we tie everything together and state the resulting margin and dimension bounds obtained in this general setting.

4.1 Fourier-expansions over Finite Abelian Groups

Let $(D,+)$ be a finite Abelian group of size $d = |D|$. A function $\chi : D \to \mathbb{C}$ is called a *character* over D if, for every $x, y \in D$,

$$\chi(x+y) = \chi(x)\cdot\chi(y) \ .$$

It is well-known that there are exactly d characters, and they form an orthonormal basis of the vector space $\mathbb{C}^D$ with respect to the inner product

$$\langle f, g\rangle := \frac{1}{d} \cdot \sum_{x \in D} f(x) \cdot \overline{g(x)} \ . \tag{9}$$

We may fix a bijection between D and the set of characters and write χ_z for the character that corresponds to $z \in D$. Every function $f : D \to \mathbb{C}$ can be written in the form

$$f(x) = \sum_{z \in D} \hat{f}(z) \cdot \chi_z(x) \tag{10}$$

where

$$\hat{f}(z) := \langle f, \chi_z\rangle = \frac{1}{d} \cdot \sum_{y \in D} f(y) \cdot \overline{\chi_z(y)} \ .$$

Equation (10) is referred to as the *Fourier expansion of* f, and $\hat{f}(z)$ is called the *Fourier-coefficient of* f *at* z.

According to the "Fundamental Theorem for Finitely Generated Abelian Groups", every finite Abelian group is, up to isomorphism, of the form

$$D = \mathbb{Z}_{q_1} \times \cdots \times \mathbb{Z}_{q_n} \tag{11}$$

for some sequence $q_1, \ldots, q_n$ of prime powers. Equation (11) is assumed henceforth so that

$$d = |D| = \prod_{k=1}^{n} q_k \ .$$

It is well-known that the characters over $\mathbb{Z}_m$ are given by

$$\chi_k^{(m)}(j) = \omega_m^{jk} \ ,$$

where

$$\omega_m = \exp\left(\frac{2\pi i}{m}\right)$$

is the primitive root of unity of order m. The characters over D are then given by

$$\chi_z(x) = \prod_{k=1}^{n} \chi_{z_k}^{(q_k)}(x_k) \ .$$

Consider now the matrix $H = (H_{x,z})_{x,z \in D}$ given by $H_{x,z} = \chi_z(x)$. It is obvious that H is symmetric. By the orthonormality of the characters with respect to the inner product in (9), it follows that

$$H^* \cdot H = H \cdot H^* = d \cdot I \ ,$$

where I denotes the identity matrix.

4.2 Distributed Functions over Finite Abelian Groups

We are interested in distributed functions $f : D \times D \to \mathbb{C}$ and arrange the d^2 Fourier-coefficients of such a function as a matrix as follows:

$$\begin{aligned}
\widehat{F}_{a,b} &= \widehat{f}(a, -b) && (12)\\
&= d^{-2} \sum_{(x,y) \in D \times D} f(x,y)\overline{\chi_{(a,-b)}(x,y)} && (13)\\
&= d^{-2} \cdot \sum_{x \in D} \sum_{y \in D} f(x,y)\overline{\chi_a(x)}\chi_b(y) && (14)
\end{aligned}$$

In matrix notation, this reads as

$$\widehat{F} = d^{-2} \cdot H^* \cdot F \cdot H \ . \tag{15}$$

A distributed function $f(x,y)$ over D is said to be *rotation-invariant* if, for all $x, y, a \in D$, the following holds:

$$f(x+a, y+a) = f(x,y)$$

Note that rotation-invariance collapses to reflection-invariance when the underlying Abelian group is of the form $(\mathbb{Z}_2^n, +)$ or, written as group with a multiplicative structure, of the form $(\{-1,1\}^n, \cdot)$.

Here are some examples for rotation-invariant functions:

- A distributed function of the form $f(x,y) = g(x-y)$ is obviously rotation-invariant. Conversely, any rotation-invariant function $f(x,y)$ can be written in this form by setting $g(x) := f(x,0)$ because rotation-invariance implies that
$$f(x,y) = f(x-y, 0) = g(x-y) \ .$$
- Because of the obvious identity
$$\chi_z(x-y) = \chi_z(x) \cdot \overline{\chi_z(y)} \ ,$$
the distributed function $\chi_z(x)\cdot\overline{\chi_z(y)}$ is rotation-invariant too.

The fact that $f(x,y) = g(x-y)$ is a rotation-invariant function can be restated as follows: any function $f(x,y)$ that can be cast as a function in $x_1 - y_1 \bmod q_1, \ldots, x_n - y_n \bmod q_n$ is rotation-invariant.

In terms of the matrix of Fourier-coefficients, $\hat{F}$, rotation-invariant functions over D can be characterized as follows:

Lemma 8 *A distributed function $f(x,y)$ over D is rotation-invariant iff $\hat{F}$ is a diagonal matrix.*

Proof: Assume first that $f(x,y)$ is rotation-invariant. Consider a Fourier-coefficient in $\hat{F}$ outside the main diagonal, say $\widehat{F}_{a,b}$ so that $a_k \neq b_k$. Every pair (x,y) can be put into the equivalence class

$$\{(x + j\vec{e_k}, y + j\vec{e_k}) :\ j = 0, \ldots, q_k - 1\} \ .$$

We show that every equivalence class contributes 0 to (14):

$$\sum_{j=0}^{q_k-1} f(x + j\vec{e_k}, y + j\vec{e_k})\overline{\chi_a(x + j\vec{e_k})} \cdot \chi_b(y + j\vec{e_k}) =$$

$$f(x,y)\overline{\chi_a(x)} \cdot \chi_b(y) \sum_{j=0}^{q_k-1} \overline{\chi_{a_k}^{(q_k)}(j)}\chi_{b_k}^{(q_k)}(j)$$

The latter sum vanishes because it equals

$$\sum_{j=0}^{q_k-1} \omega_{q_k}^{(b_k-a_k)j} \ .$$

Recall that δ denotes the Kronecker symbol and it is well-known that

$$\sum_{j=0}^{m-1} \omega_m^{(l'-l)j} = m \cdot \delta_{l,l'} \ .$$

This shows that $\widehat{F}_{a,b} = 0$.
Now assume that $\hat{F}$ is a diagonal matrix. We conclude from (15) that

$$F = H \cdot \hat{F} \cdot H^* \ , \tag{16}$$

which implies that

$$F_{x,y} = \sum_{z \in D} \widehat{F}_{z,z} \cdot \chi_x(z) \cdot \overline{\chi_y(z)} \ .$$

Rotation-invariance is now easily obtained:

$$\begin{aligned} f(x+a, y+a) &= \sum_{z \in D} \widehat{F}_{z,z} \cdot \chi_{x+a}(z) \cdot \overline{\chi_{y+a}(z)} \\ &= \sum_{z \in D} \widehat{F}_{z,z} \cdot \chi_z(x+a) \cdot \overline{\chi_z(y+a)} \\ &= \sum_{z \in D} \widehat{F}_{z,z} \cdot \chi_z(x) \cdot \overline{\chi_z(y)} \\ &= f(x,y) \end{aligned}$$

In the second-last equation, we used the rotation-invariance of $\chi_z(x) \cdot \overline{\chi_z(y)}$. ∎

Corollary 9 *Assume that $f(x,y)$ is a rotation-invariant distributed function over D. Then the (complex) eigenvalues of $d^{-1} \cdot F$ are found on the main diagonal of $\hat{F}$.*

Proof: Rewrite (16) as

$$d^{-1}F = (d^{-1/2}H) \cdot \hat{F} \cdot (d^{-1/2}H^*)$$

and observe that this is nothing but the spectral decomposition of $d^{-1}F$ (since $\hat{F}$ is a diagonal matrix and $d^{-1/2}H$ is unitary). ∎

We briefly note the following result:

Lemma 10 *Let $\hat{F}$ be the (diagonal) matrix that contains the Fourier-coefficients of the (rotation-invariant) distributed function $f(x-y)$. Then, for every $z \in D$, $\hat{f}(z) = \widehat{F}_{z,z}$.*

Proof: Consider the function $f_y(x) := f(x-y)$. We shall show below that the Fourier coefficients of f and f_y are related as follows:

$$\widehat{f_y}(z) = \hat{f}(z) \cdot \overline{\chi_y(z)} \ . \tag{17}$$

The proof is now obtained by the following calculation:

$$\begin{aligned} \widehat{F}_{z,z} &= d^{-2} \cdot \sum_{x,y \in D} f(x-y) \cdot \overline{\chi_z(x)} \cdot \chi_z(y) \\ &= d^{-1} \cdot \sum_{y \in D} \left(d^{-1} \cdot \sum_{x \in D} f_y(x) \overline{\chi_z(x)} \right) \chi_z(y) \\ &= d^{-1} \cdot \sum_{y \in D} \widehat{f_y}(z) \cdot \chi_z(y) \\ &\overset{(17)}{=} \hat{f}(z) \cdot d^{-1} \cdot \sum_{y \in D} \underbrace{\overline{\chi_y(z)} \chi_z(y)}_{=1} \\ &= \hat{f}(z) \end{aligned}$$

The following calculation verifies (17):

$$\begin{aligned} \widehat{f_y}(z) &= d^{-1} \cdot \sum_{x \in D} f(x-y) \cdot \overline{\chi_x(x)} \\ &= d^{-1} \cdot \sum_{x \in D} \sum_{w \in D} \hat{f}(w) \cdot \chi_w(x-y) \cdot \overline{\chi_z(x)} \\ &= d^{-1} \cdot \sum_{x \in D} \sum_{w \in D} \hat{f}(w) \cdot \chi_w(x) \cdot \overline{\chi_w(y)} \cdot \overline{\chi_z(x)} \\ &= d^{-1} \cdot \sum_{w \in D} \underbrace{\left(\sum_{x \in D} \chi_w(x) \cdot \overline{\chi_z(x)} \right)}_{=d \cdot \delta_{w,z}} \hat{f}(w) \cdot \overline{\chi_w(y)} \\ &= \hat{f}(z) \cdot \overline{\chi_z(y)} \end{aligned}$$

∎

Corollary 9 and Lemma 10 yield the following

Corollary 11 *Let F denote the matrix with entries $F_{x,y} = f(x-y)$. Then the spectrum of (complex) eigenvalues of $d^{-1} \cdot F$ coincides with the spectrum of (complex) Fourier-coefficients of f.*

4.3 Margin and Dimension Bounds for Rotation-invariant Kernels

Let D be a finite Abelian group, and let $\mathcal{G}$ be the subgroup of $\mathcal{S}(D)$ that contains all permutations of the form $x \mapsto x + a$. Note that $|\mathcal{G}| = |D|$. For every function $f : D \to \{-1, 1\}$,

$$\overline{\mu}_{rot}(f) := \overline{\mu}_{\mathcal{G}}(f)$$

denotes the largest possible average margin that can be achieved by a linear arrangement for f resulting from a rotation-invariant kernel. As for the smallest possible dimension, parameter $d_{rot}(f)$ is understood analogously.

Corollary 12 *Let D be a finite Abelian group of size d. Every function $f : D \to \{-1, 1\}$ satisfies*

$$\overline{\mu}_{rot}(f) \le \|\hat{f}\|_\infty \ . \tag{18}$$

In other words, no large margin classifier that employs a rotation-invariant kernel can achieve an average margin for f which exceeds $\|\hat{f}\|_\infty$.

Proof: According to Theorem 5,

$$\overline{\mu}_{rot}(f) \leq \frac{\|M^{f,\mathcal{G}}\|}{\sqrt{|D| \cdot |\mathcal{G}|}} = \frac{\|M^{f,\mathcal{G}}\|}{d} \ .$$

The matrix $M^{f,\mathcal{G}}$ coincides with the matrix F given by $F_{x,y} = f(x-y)$ (up to permutation of columns). We conclude from this and from Corollary 11 that

$$\|M^{f,\mathcal{G}}\| = \|F\| = d \cdot \|\hat{f}\|_\infty \ ,$$

which leads us to inequality (18). ∎

Corollary 6 and 7 combined with Corollary 11 lead us to the following results:

Corollary 13 *Let $d_{rot}(f)$ denote the smallest dimension of a feature space associated with a rotation-invariant kernel K that allows for a correct representation of f. Then, $d_{rot}(f) \geq \|\hat{f}\|_\infty^{-1}$.*

Corollary 14 *Let $\widehat{f_i}$ denote the (complex) i-th Fourier-coefficient of f, where $|\widehat{f_1}|, \ldots, |\widehat{f_d}|$ are in decreasing order. Then,*

$$d_{rot}(f) \cdot \sum_{i=1}^{d_{rot}(f)} \left|\widehat{f_i}\right|^2 \geq 1$$

5 Reflection-invariant Functions

A *distributed function over D* is a real-valued function over the domain $D \times D$. We will occasionally identify a distributed function f over D with the $(D \times D)$-matrix F given by $F_{x,y} = f(x,y)$.

In this section, we are particularly interested in distributed functions $f(x,y)$ over $D = \{-1,1\}^n$. In this case, f has 2^{2n} Fourier-coefficients that can be nicely arranged as a matrix $\hat{F}$. More precisely:

$$\begin{aligned} \widehat{F}_{a,b} &= \widehat{f}(a,b) \\ &= 2^{-2n} \cdot \sum_{(x,y) \in D \times D} \chi_{(a,b)}(x,y) f(x,y) \\ &= 2^{-2n} \cdot \sum_{x \in D} \sum_{y \in D} \chi_a(x) \chi_b(y) f(x,y) \end{aligned}$$

In matrix notation, this reads as

$$\widehat{F} = 2^{-2n} \cdot H \cdot F \cdot H \ ,$$

where H denotes the $(2^n \times 2^n)$-Walsh-matrix.

A distributed function $f(x,y)$ over $\{-1,1\}^n$ is said to be *reflection-invariant* if, for all $x, y, a \in \{-1,1\}^n$, the following holds:

$$f(x \circ a, y \circ a) = f(x,y) \tag{19}$$

Distributed functions $f(x,y)$ over $\mathbb{R}^n$ that satisfy (19) for all $x, y \in \mathbb{R}^n$ and every $a \in \{-1,1\}^n$ are said to be *reflection-invariant in the Euclidean space.* Here are some examples:

- A distributed function of the form $f(x,y) = g(x \circ y)$ is reflection-invariant (in the Euclidean space provided that the domain is $\mathbb{R}^n$):

  $$g((x \circ a) \circ (y \circ a)) = g(x \circ y \circ (a \circ a)) = g(x \circ y)$$

  Conversely, any reflection-invariant function $f(x,y)$ (over domain $\{-1,1\}^n$) can be written in this form by setting $g(x) := f(x, \vec{e})$ because reflection-invariance implies that

  $$f(x,y) = f(x \circ y, y \circ y) = f(x \circ y, \vec{e}) = g(x \circ y) \ .$$

- Because of the obvious identity

  $$\chi_z(x \circ y) = \chi_z(x) \cdot \chi_z(y) \ ,$$

  the distributed function $\chi_z(x) \cdot \chi_z(y)$ is reflection-invariant too.

- The metric

  $$L_p(x-y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

  induced by the L_p-norm is clearly reflection-invariant in the Euclidean space.

In Section 6, we shall see that many popular kernel functions happen to be reflection-invariant.

The fact that $f(x,y) = g(x \circ y)$ is a reflection-invariant function can be restated as follows: any function $f(x,y)$ that can be cast as a function in $x_1 \cdot y_1, \ldots, x_n \cdot y_n$ is reflection-invariant. Similarly, any function $f(x,y)$ that can be cast as a function in $L_p(x-y)$ is reflection-invariant.

We move on and consider the possibility of making new reflection-invariant functions from functions that are already known to be reflection-invariant. We clearly have the following

Pointwise Closure Property: The pointwise limit of reflection-invariant functions is a reflection-invariant function. Furthermore, if $f_1, \ldots, f_d$ are reflection-invariant functions and $g: \mathbb{R}^d \to \mathbb{R}$ is an arbitrary function, then

$$g(f_1(x,y), \ldots, f_d(x,y))$$

is reflection-invariant too.

Sometimes a new distributed function is constructed from given-ones by means of matrix operations (which, in general, are not performed pointwise). We claim that the following holds:

Lemma 15 *Reflection-invariant distributed functions over $\{-1,1\}^n$ are closed under the usual matrix product and under the tensor-product of matrices. More precisely, let $F(x,y)$ and $G(x,y)$ be two reflection-invariant distributed functions (here viewed as matrices). Then, the functions $(F \cdot G)(x,y)$ and $(F \otimes G)[(u,x),(v,y)]$ are reflection-invariant too.*

Note that Lemma 15 is an immediate corollary to Lemma 1.

The following result does not seem to be new.[2] We have derived a more general result in Section 4.2 (a section concerned with rotation-invariance with reflection-invariance as a special case).[3]

Lemma 16 *1. A distributed function $f(x, y)$ over $\{-1, 1\}^n$ is reflection-invariant iff $\hat{F}$ is a diagonal matrix.*

2. Assume that $F_{x,y} = f(x, y) = g(x \circ y)$ is a reflection-invariant distributed function. Then, matrix F is symmetric, the (real) eigenvalues of $2^{-n} \cdot F$ coincide with the Fourier-coefficients of g, and $\hat{g}(z) = \widehat{F}_{z,z}$.

Note that Lemma 8 applied to $D = \mathbb{Z}_2^n$ yields the first statement in Lemma 16. The second statement in Lemma 16 follows analogously from the subsequent results (from Corollary 9 up to Corollary 11).

6 Reflection-invariant Kernels

In this section, we consider kernel functions $K(x, y)$ over the Boolean or over the Euclidean domain. In other words, $K(x, y)$ is a distributed function over $\{-1, 1\}^n$ or over $\mathbb{R}^n$ with the additional property that every finite principal sub-matrix of K is symmetric and positive semidefinite. In Section 6.1, we demonstrate that the family of reflection-invariant kernels is quite rich and contains many popular kernels. In Section 6.2, we derive some central properties of reflection-invariant kernels and present an upper bound on the margin that can be achieved by a kernel of this type.

6.1 Examples and Closure Properties

Let us start with some examples. The following (quite popular) kernels (over $\mathbb{R}^n$ except for the DNF-Kernel that has a Boolean domain) can be cast as functions in $x_1 \cdot y_1, \ldots, x_n \cdot y_n$ or as functions in $\|x - y\|_2$ and are therefore reflection-invariant:

Polynomial Kernels: $K(x, y) = p(x^\top y)$ for an arbitrary polynomial p with positive coefficients.

All-subsets Kernel: $K(x, y) = \prod_{i=1}^n (1 + x_i y_i)$.

ANOVA Kernel: Let $1 \le s \le n$ and define

$$K_s(x, y) = \sum_{1 \le i_1 < \cdots < i_s \le n} \prod_{j=1}^{s} x_{i_j} y_{i_j} \ .$$

DNF-Kernel: $K(x, y) = -1 + 2^{-n} \prod_{i=1}^n (x_i y_i + 3)$.

Exponential Kernels: $K(x, y) = e^{p(x^\top y)}$ for an arbitrary polynomial p with positive coefficients.

Gaussian Kernel: $K(x, y) = e^{-\|x-y\|_2^2/\sigma^2}$ for an arbitrary $\sigma > 0$.

These kernels have the usual nice properties like being efficiently evaluable although the number of (implicitly represented) features is exponentially large (or even infinite). Polynomial, Exponential, and Gaussian Kernels (first used in [2]) are found in almost any basic text-book that is relevant to the subject (e.g. [3]). The All-subsets Kernel is found in [18], and the ANOVA Kernel is found in [19]. As for the latter two kernels, see also [17]. The DNF-Kernel has been proposed in [16].[4] The reader interested in more information about these (and other) kernels may consult the relevant literature. Here, we simply point to the fact that all kernels mentioned above are reflection-invariant.

We move on and consider the possibility of making new reflection-invariant kernels from kernels that are already known to be reflection-invariant. To this end, we briefly call into mind some basic closure properties of kernels:

Lemma 17 *Let K, K_1, K_2 be kernels, and let $c > 0$ be a positive constant. Then, the distributed functions*

$$\begin{array}{rcl} K_1(x,y) + K_2(x,y) & , & c \cdot K(x,y) \\ K_1(x,y) \cdot K_2(x,y) & , & (K_1 \otimes K_2)[(u,x),(v,y)] \end{array}$$

are kernels too. Moreover, the pointwise limit of kernels yields a kernel.

The proof of Lemma 17 can be looked-up in [3], for example.

Corollary 18 *If $K_1, \ldots, K_d$ are kernels and $P : \mathbb{R}^d \to \mathbb{R}$ is a polynomial (or a converging power series) with positive coefficients, then $P(K_1(x, y), \ldots, K_d(x, y))$ is a kernel too.*

Note that closure properties of reflection-invariant functions (see the "Pointwise Closure Property" and Lemma 15) are comparably strong so that Lemma 17 and Corollary 18 remain valid (mutatis mutandis) for reflection-invariant kernels.

The following kernels (proposed in [11] and [9], respectively) define a new kernel-matrix K in terms of a given symmetric matrix B (called "similarity matrix" in this context):

[2] The second statement of Lemma 16 is mentioned as well-known in [14] (without providing a pointer to the literature).

[3] The results in Section 4.2 might be known as well, but we are not aware of an appropriate pointer to the literature.

[4] In [16], the kernel is defined over the Boolean domain $\{0, 1\}^n$. Our formula above is obtained from the formula in [16] by plugging in the affine transformation that identifies 1 with -1 and 0 with 1. A similar remark applies to the Monotone DNF-Kernel discussed at the end of this section.

Exponential Diffusion Kernel: For $\lambda \in \mathbb{R}$, define

$$K = e^{\lambda \cdot B} = \sum_{k \geq 0} \frac{\lambda^k}{k!} \cdot B^k \ .$$

von Neumann Diffusion Kernel: For $0 \leq \lambda < \|B\|^{-1}$, define

$$K = (I - \lambda \cdot B)^{-1} = \sum_{k \geq 0} \lambda^k \cdot B^k \ .$$

It follows from the closure properties of reflection-invariant functions that both diffusion kernels would inherit reflection-invariance from the underlying similarity matrix B.

The family of reflection-invariant kernels is quite rich. But here are two kernels (the first-one from [16], and the second-one from [12]) which are counterexamples:

Monotone DNF-Kernel:

$$K(x, y) = -1 + 2^{-2n} \prod_{i=1}^{n} (x_j y_j - x_j - y_j + 5) \ .$$

Spectrum Kernel: Here, $x, y \in \{-1, 1\}^n$ are considered as binary strings. For $1 \leq p \leq n$ and for every substring $u \in \{-1, 1\}^p$,

$$\Phi_v^p(x) = |\{(u, w) : \ x = uvw\}|$$

counts how often v occurs as a substring of x. The p-Spectrum Kernel is then given by

$$K(x, y) = \sum_{v \in \{-1,1\}^p} \Phi_v^p(x) \cdot \Phi_v^p(y) \ .$$

It is easy to see that both kernels are *not* reflection-invariant. More generally, string kernels (measuring similarity between strings) often violate reflection-invariance.

6.2 Properties of Reflection-invariant Kernels

In the sequel, we shall provide an upper bound on the average margin (that, a-fortiori, upper-bounds the guaranteed margin). Recall the definitions of Section 2.2. Since (2) is valid for any kernel-function, the condition of reflection-invariance is equivalent to

$$\forall x, y, a \in D : \ \langle \Phi(x \circ a), \Phi(y \circ a) \rangle = \langle \Phi(x), \Phi(y) \rangle \ .$$

With every function $g : D \to \mathbb{R}$ and every $a \in D$, we associate the function $g_a(x) := g(x \circ a)$ and call it the *a-reflection* of g. With this notation, we get

Lemma 19 *If kernel K is reflection-invariant, then the following holds for all $x, a \in D$:*

$$\begin{aligned} \|\Phi_K(x \circ a)\|_K &= \|\Phi_K(x)\|_K \\ \|w(\alpha)\|_K &= \|w(\alpha_a)\|_K \end{aligned}$$

In other words, the norm $\| \cdot \|_K$ assigns the same length to all feature vectors, and a reflection of the dual vector does not change the norm of the induced weight vector.

Note that Lemma 19 is an immediate corollary to Lemma 2.

Lemma 20 *For every reflection-invariant kernel K, and every choice of $f : D \to \{-1, 1\}$, $x, a \in D$, and $\alpha : D \to \mathbb{R}$, the following holds:*

$$\mu_K(f_a|\alpha_a, x) = \mu_K(f|\alpha, x \circ a)$$

Again this is a direct inference of the general results of Section 3 (see Lemma 3).

Corollary 21 *For every reflection-invariant kernel K, and every choice of $f : D \to \{-1, 1\}$, $a \in D$, and $\alpha : D \to \mathbb{R}$, the following holds:*

$$\begin{aligned} \overline{\mu}_K(f_a|\alpha_a) &= \overline{\mu}_K(f|\alpha) \\ \mu_K(f_a|\alpha_a) &= \mu_K(f|\alpha) \\ \overline{\mu}_K(f_a) &= \overline{\mu}_K(f) \\ \mu_K(f_a) &= \mu_K(f) \\ \overline{\mu}_{inv}(f_a) &= \overline{\mu}_{inv}(f) \\ \mu_{inv}(f_a) &= \mu_{inv}(f) \end{aligned}$$

The punchline of the preceding discussion is the following result:

Theorem 22 *Every Boolean function f satisfies*

$$\overline{\mu}_{inv}(f) \leq \|\hat{f}\|_\infty \ .$$

In other words, no large margin classifier that employs a reflection-invariant kernel can achieve an average margin for f which exceeds $\|\hat{f}\|_\infty$.

Proof: Let $\|M\|$ denote the spectral norm of a matrix $M \in \mathbb{R}^{r \times s}$. Recall that the spectral norm of a symmetric matrix coincides with the largest absolute value found in the spectrum of M's eigenvalues. We shall apply (6) to the matrix $F \in \{-1, 1\}^{D \times D}$ with entries $F_{x,y} = f(x \circ y) = f_y(x)$. Here, $r = s = |D| = 2^n$ and, according to Lemma 16, $\|F\| = 2^n \cdot \|\hat{f}\|_\infty$, and (6) now reads as follows:

$$\sum_{x \in D} \sum_{y \in D} f_y(x) \langle u_x, v_y \rangle \leq 2^{2n} \cdot \|\hat{f}\|_\infty \qquad (20)$$

Assume now, for sake of contradiction, that there exists a reflection-invariant kernel K and an $\alpha : D \to \mathbb{R}$ such that the following holds:

$$\begin{aligned} \overline{\mu}_K(f|\alpha) &= 2^{-n} \sum_{x \in D} \frac{f(x) \cdot \langle w(\alpha), \Phi(x) \rangle}{\|w(\alpha)\| \cdot \|\Phi(x)\|} && (21) \\ &> \|\hat{f}\|_\infty && (22) \end{aligned}$$

According to Corollary 21, $\overline{\mu}_K(f|\alpha) = \overline{\mu}_K(f_y|\alpha_y)$ for every $y \in D$. Thus, the inequality $\overline{\mu}_K(f_y|\alpha_y) > \|\hat{f}_y\|_\infty$ holds as well. Note, however, that $\|\hat{f}_y\|_\infty = \|\hat{f}\|_\infty$. Summing over all $y \in D$ and multiplying both hand-sides in (21), (22) by 2^n, we obtain

$$\sum_{x,y \in D} \frac{f_y(x) \cdot \langle w(\alpha_y), \Phi(x) \rangle}{\|w(\alpha_y)\| \cdot \|\Phi(x)\|} > 2^{2n} \cdot \|\hat{f}\|_\infty \ .$$

Setting $u_x := \Phi(x)/\|\Phi(x)\|$ and $v_y := w(\alpha_y)/\|w(\alpha_y)\|$, we arrived at a contradiction to (20). ∎

So by using the algorithm of Kushilevitz and Mansour [20] one can estimate the largest Fourier-coefficient of a specific function f to determine an upper bound for the maximal achievable average margin.

In analogy to Corollaries 6 and 7, we obtain the following results:

Corollary 23 *Let $d_{inv}(f)$ denote the smallest dimension of a feature space associated with a reflection-invariant kernel K that allows for a correct representation of f. Then, $d_{inv}(f) \geq \|\hat{f}\|_\infty^{-1}$.*

Proof: Similarly to Corollary 6 the corresponding feature space for the kernel must have dimension at least $\sqrt{|D| \cdot |\mathcal{G}|}/\|M^{f,\mathcal{G}}\|$. Here, it is $2^n/\|F\|$. Since $\|F\| = 2^n \cdot \|\hat{f}\|_\infty$, we are done. ∎

Corollary 24 *Let $\widehat{f_i}$ denote the i-th Fourier-coefficient of f, where $|\widehat{f_1}|, \ldots, |\widehat{f_{2^n}}|$ are in decreasing order. Then, $d_{inv}(f)$ satisfies the following lower bound:*

$$d_{inv}(f) \cdot \sum_{i=1}^{d_{inv}(f)} \left|\widehat{f_i}\right|^2 \geq 1 \tag{23}$$

Proof: According to (7)

$$d_{\mathcal{G}}(f) \cdot \sum_{i=1}^{d_{\mathcal{G}}(f)} \sigma_i^2 \geq 1 \tag{24}$$

holds. We apply this result to the matrix F definied in the proof of Theorem 22. Here, $r = s = 2^n$. Moreover, $\sigma_i(F) = 2^n|\hat{f}_i|$ according to Lemma 16, because $\sigma_i(A) = |\lambda_i(A)|$ when A is symmetric. (Here, $\lambda_i(A)$ is the sequence of eigenvalues in decreasing order of their absolute values.) Plugging this into (24), we obtain (23). ∎

7 Open Problems

Haasdonk and Burkhardt [8] consider two notions of invariance: "simultaneous invariance" and "total invariance". Simultaneous invariance very much corresponds to the notion of invariance that we discussed in section 3 so that our margin and dimension bounds apply. Total invariance is a stronger notion so that our bounds apply more than ever. But the obvious challenge is to find *stronger* margin and dimension bounds for *totally* invariant kernels.

The basic idea behind our paper is roughly as follows. For a family of kernels (e.g., polynomial kernels), we argue that the existence a "good representation" for a particular target function implies the existence of a "good representation" for a whole family of target functions (so that classical margin and dimension bounds can be brought into play). We think that invariance under a group operation (the notion considered in this paper) is just the first obvious thing one should consider. We would like to develop more versatile techniques that, while following the same basic idea, lead to strong margin and dimension bounds for a wider class of kernels.

References

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

[2] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[3] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[4] Jürgen Forster. A linear lower bound on the unbounded error communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.

[5] Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt, and Hans Ulrich Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *Proceedings of the 21'st Annual Conference on the Foundations of Software Technology and Theoretical Computer Science*, pages 171–182, 2001.

[6] Jürgen Forster, Niels Schmitt, Hans Ulrich Simon, and Thorsten Suttorp. Estimating the optimal margins of embeddings in euclidean half spaces. *Machine Learning*, 51(3):263–281, 2003.

[7] Jürgen Forster and Hans Ulrich Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theoretical Computer Science*, 350(1):40–48, 2006.

[8] Bernard Haasdonk and Hans Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine Learning*, 68(1):35–61, 2007.

[9] Jaz S. Kandola, John Shawe-Taylor, and Nello Cristianini. Learning semantic similarity. In *Advances in Neural Information Processing Systems 15*, pages 657–664. MIT Press, 2003.

[10] Eike Kiltz and Hans Ulrich Simon. Threshold circuit lower bounds on cryptographic functions. *Journal of Computer and System Sciences*, 71(2):185–212, 2005.

[11] Risi I. Kondor and John D. Lafferty. Diffusion kernels on graphs and other discrete input

spaces. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[12] Christina Leslie, Eleazar Eskin, and William S. Noble. The spectrum kernel: A string kernel for SVM protein classification. In *Pacific Symposium on Biocomputing*, pages 564–575, 2002.

[13] Nathan Linial, Shahar Mendelson, Gideon Schechtman, and Adi Shraibman. Complexity measures of sign matrices. Combinatorica. To appear.

[14] Nati Linial and Adi Shraiman. Lower bounds in communication complexity based on factorization norms. In *Proceedings of the 39th Annual Symposium on Theory of Computing*, pages 699–708, 2007.

[15] Alexander A. Razborov and Alexander A. Sherstov. The sign-rank of AC^0. Personal Communication.

[16] Ken Sadohara. Learning of boolean functions using support vector machines. In *Proceedings of the 12th International Conference on Algorithmic Learning Theory*, pages 106–118, 2001.

[17] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[18] Eiji Takimoto and Manfred K. Warmuth. Pathe kernels and multiplicative updates. *Journal of Machine Learning Research*, 4:773–818, 2003.

[19] Vladimir Vapnik, Christopher J. C. Burges, Bernhard Schoelkopf, and R. Lyons. A new method for constructing artificial neural networks. Interim ARPA Technical Report, AT&T Bell Laboratories, 1995.

[20] Eyal Kushilevitz and Yishay Mansour Learning decision trees using the Fourier spectrum In *Proceedings of the 23th Annual Symposium on Theory of Computing*, pages 455–464, 1991.