A Tutorial Review of RKHS Methods in Machine Learning

Thomas Hofmann[*], Bernhard Schölkopf[†], and Alexander J. Smola[‡]

December 6, 2005

1 Introduction

Over the last ten years, estimation and learning methods utilizing positive definite kernels have become rather popular, particularly in machine learning. Since these methods have a stronger mathematical slant than earlier machine learning methods (e.g., neural networks), there is also significant interest in the statistical and mathematical community for these methods. The present review aims to summarize the state of the art on a conceptual level. In doing so, we build on various sources (including the books [Vapnik, 1998, Burges, 1998, Cristianini and Shawe-Taylor, 2000, Herbrich, 2002] and in particular [Schölkopf and Smola, 2002]), but we also add a fair amount of recent material which helps unifying the exposition.

The main idea of all the described methods can be summarized in one paragraph. Traditionally, theory and algorithms of machine learning and statistics has been very well developed for the linear case. Real world data analysis problems, on the other hand, often requires nonlinear methods to detect the kind of dependences that allow successful prediction of properties of interest. By using a positive definite kernel, one can sometimes have the best of both worlds. The kernel corresponds to a dot product in a (usually high dimensional) feature space. In this space, our estimation methods are linear, but as long as we can formulate everything in terms of kernel evaluations, we never explicitly have to work in the high dimensional feature space.

2 Kernels

2.1 An Introductory Example

Suppose we are given empirical data

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}. \tag{1}$$

Here, the domain $\mathcal{X}$ is some nonempty set that the *inputs* x_i are taken from; the $y_i \in \mathcal{Y}$ are called *targets*. Here and below, $i, j = 1, \ldots, n$.

Note that we have not made any assumptions on the domain $\mathcal{X}$ other than it being a set. In order to study the problem of learning, we need additional structure. In learning, we want to be able to *generalize* to unseen data points. In the case of pattern recognition, given some new input $x \in \mathcal{X}$, we want to predict the corresponding $y \in \{\pm 1\}$. Loosely speaking, we want

[*]Universität Darmstadt

[†]Max-Planck-Institut für biologische Kybernetik

[‡]Alex.Smola@nicta.com.au, Statistical Machine Learning Program, National ICT Australia, and Computer Science Laboratory, Australian National University, Canberra, 0200 ACT

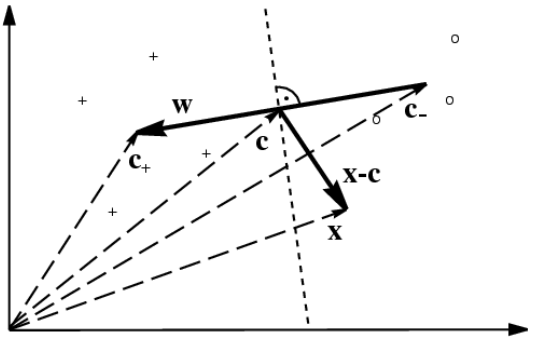

Figure 1: A simple geometric classification algorithm: given two classes of points (depicted by 'o' and '+'), compute their means c_+, c_- and assign a test input x to the one whose mean is closer. This can be done by looking at the dot product between $x - c$ (where $c = (c_+ + c_-)/2$) and $\mathbf{w} := c_+ - c_-$, which changes sign as the enclosed angle passes through $\pi/2$. Note that the corresponding decision boundary is a hyperplane (the dotted line) orthogonal to w (from [Schölkopf and Smola, 2002]).

to choose y such that (x, y) is in some sense *similar* to the training examples. To this end, we need similarity measures in $\mathcal{X}$ and in $\{\pm 1\}$. The latter is easier, as two target values can only be identical or different. For the former, we require a function

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \qquad (x, x') \mapsto k(x, x') \tag{2}$$

satisfying, for all $x, x' \in \mathcal{X}$,

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \,, \tag{3}$$

where Φ maps into some dot product space $\mathcal{H}$, sometimes called the *feature space*. The similarity measure k is usually called a *kernel*, and Φ is called its *feature map*.

The advantage of using such a kernel as a similarity measure is that it allows us to construct algorithms in dot product spaces. For instance, consider the following simple classification algorithm, where $\mathcal{Y} = \{\pm 1\}$. The idea is to compute the means of the two classes in the feature space, $c_+ = \frac{1}{n_+} \sum_{\{i:y_i=+1\}} \Phi(x_i)$, and $c_- = \frac{1}{n_-} \sum_{\{i:y_i=-1\}} \Phi(x_i)$, where n_+ and n_- are the number of examples with positive and negative target values, respectively. We then assign a new point $\Phi(x)$ to the class whose mean is closer to it. This leads to

$$y \quad = \quad \mathrm{sgn}(\langle \Phi(x), c_+ \rangle - \langle \Phi(x), c_- \rangle + b) \tag{4}$$

with $b = \frac{1}{2} \left(\|c_-\|^2 - \|c_+\|^2 \right)$. Substituting the expressions for $c_\pm$ yields

$$y = \mathrm{sgn} \left(\frac{1}{n_+} \sum_{\{i:y_i=+1\}} \langle \Phi(x), \Phi(x_i) \rangle - \frac{1}{n_-} \sum_{\{i:y_i=-1\}} \langle \Phi(x), \Phi(x_i) \rangle + b \right) \tag{5}$$

Rewritten in terms of k, this reads

$$y = \mathrm{sgn} \left(\frac{1}{n_+} \sum_{\{i:y_i=+1\}} k(x, x_i) - \frac{1}{n_-} \sum_{\{i:y_i=-1\}} k(x, x_i) + b \right), \tag{6}$$

where $b = \frac{1}{2}\left(\frac{1}{n_-^2}\sum_{\{(i,j):y_i=y_j=-1\}} k(x_i,x_j) - \frac{1}{n_+^2}\sum_{\{(i,j):y_i=y_j=+1\}} k(x_i,x_j)\right)$.

Let us consider one well-known special case of this type of classifier. Assume that the class means have the same distance to the origin (hence $b = 0$), and that $k(.,x)$ is a density for all $x \in \mathcal{X}$. If the two classes are equally likely and were generated from two probability distributions that are correctly estimated by the Parzen windows estimators

$$p_+(x) := \frac{1}{n_+}\sum_{\{i:y_i=+1\}} k(x,x_i), \qquad p_-(x) := \frac{1}{n_-}\sum_{\{i:y_i=-1\}} k(x,x_i), \tag{7}$$

then (6) is the Bayes decision rule.

The classifier (6) is quite close to the *Support Vector Machine (SVM)* that we will discuss below. It is linear in the feature space (see (4)), while in the input domain, it is represented by a kernel expansion (6). In both cases, the decision boundary is a hyperplane in the feature space; however, the normal vectors are usually rather different.[1]

2.2 Positive Definite Kernels

We have above required that a kernel satisfy (3), i.e., correspond to a dot product in some dot product space. In the present section, we show that the class of kernels that can be written in the form (3) coincides with the class of positive definite kernels. This has far-reaching consequences. There are examples of positive definite kernels which can be evaluated efficiently even though via (3) they correspond to dot products in infinite dimensional dot product spaces. In such cases, substituting $k(x,x')$ for $\langle\Phi(x),\Phi(x')\rangle$, as we have done when going from (5) to (6), is crucial. In the machine learning community, this substitution is called the *kernel trick*.

2.2.1 Prerequisites

Definition 1 (Gram matrix) *Given a kernel k and inputs $x_1,\ldots,x_n \in \mathcal{X}$, the $n \times n$ matrix*

$$K := (k(x_i,x_j))_{ij} \tag{8}$$

is called the Gram matrix *(or* kernel matrix*) of k with respect to $x_1,\ldots,x_n$.*

Definition 2 (Positive definite matrix) *A real $n \times n$ symmetric matrix K_{ij} satisfying*

$$\sum_{i,j} c_i c_j K_{ij} \geq 0 \tag{9}$$

for all $c_i \in \mathbb{R}$ is called positive definite*. If equality in (9) only occurs for $c_1 = \cdots = c_n = 0$ then we shall call the matrix* strictly positive definite*.*

Definition 3 (Positive definite kernel) *Let $\mathcal{X}$ be a nonempty set. A function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ which for all $n \in \mathbb{N}, x_i \in \mathcal{X}$ gives rise to a positive definite Gram matrix is called a* positive definite kernel*. A function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ which for all $n \in \mathbb{N}$ and distinct $x_i \in \mathcal{X}$ gives rise to a strictly positive definite Gram matrix is called a* strictly positive definite kernel*.*

[1] For (4), the normal vector is $w = c_+ - c_-$. As an aside, note that if we normalize the targets such that $\hat{y}_i = y_i/|\{j : y_j = y_i\}|$, in which case the $\hat{y}_i$ sum to zero, then $\|w\|^2 = \langle K, \hat{y}\hat{y}^\top\rangle_F$, where $\langle .,.\rangle_F$ is the Frobenius dot product. If the two classes have equal size, then up to a scaling factor involving $\|K\|_2$ and n, this equals the *kernel-target alignment* defined in [Cristianini et al., 2002].

Occasionally, we shall refer to positive definite kernels simply as a *kernels*. Note that for simplicity we have restricted ourselves to the case of real valued kernels. However, with small changes, the below will also hold for the complex valued case.

Since $\sum_{i,j} c_i c_j \langle \Phi(x_i), \Phi(x_j) \rangle = \left\langle \sum_i c_i \Phi(x_i), \sum_j c_j \Phi(x_j) \right\rangle \geq 0$, kernels of the form (3) are positive definite for any choice of Φ. In particular, if $\mathcal{X}$ is already a dot product space, we may choose Φ to be the identity. Kernels can thus be regarded as generalized dot products. Whilst they are not generally bilinear, they share important properties with dot products, such as the Cauchy-Schwartz inequality:

Proposition 4 *If k is a positive definite kernel, and $x_1, x_2 \in \mathcal{X}$, then*

$$k(x_1, x_2)^2 \leq k(x_1, x_1) \cdot k(x_2, x_2). \tag{10}$$

Proof The 2×2 Gram matrix with entries $K_{ij} = k(x_i, x_j)$ is positive definite. Hence both its eigenvalues are nonnegative, and so is their product, K's determinant, i.e.,

$$0 \leq K_{11}K_{22} - K_{12}K_{21} = K_{11}K_{22} - K_{12}^2. \tag{11}$$

Substituting $k(x_i, x_j)$ for K_{ij}, we get the desired inequality. ■

2.2.2 Construction of the Reproducing Kernel Hilbert Space

We now define a map from $\mathcal{X}$ into the space of functions mapping $\mathcal{X}$ into $\mathbb{R}$, denoted as $\mathbb{R}^{\mathcal{X}}$, via

$$\begin{aligned} \Phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ x &\mapsto k(., x). \end{aligned} \tag{12}$$

Here, $\Phi(x) = k(., x)$ denotes the function that assigns the value $k(x', x)$ to $x' \in \mathcal{X}$.

We next construct a dot product space containing the images of the inputs under Φ. To this end, we first turn it into a vector space by forming linear combinations

$$f(.) = \sum_{i=1}^{n} \alpha_i k(., x_i). \tag{13}$$

Here, $n \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$ are arbitrary.

Next, we define a dot product between f and another function $g(.) = \sum_{j=1}^{n'} \beta_j k(., x'_j)$ (with $n' \in \mathbb{N}$, $\beta_j \in \mathbb{R}$ and $x'_j \in \mathcal{X}$) as

$$\langle f, g \rangle := \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j k(x_i, x'_j). \tag{14}$$

To see that this is well-defined although it contains the expansion coefficients, note that $\langle f, g \rangle = \sum_{j=1}^{n'} \beta_j f(x'_j)$. The latter, however, does not depend on the particular expansion of f. Similarly, for g, note that $\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x_i)$. This also shows that $\langle \cdot, \cdot \rangle$ is bilinear. It is symmetric, as $\langle f, g \rangle = \langle g, f \rangle$. Moreover, it is positive definite, since positive definiteness of k implies that for any function f, written as (13), we have

$$\langle f, f \rangle = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0. \tag{15}$$

Next, note that given functions $f_1, \ldots, f_p$, and coefficients $\gamma_1, \ldots, \gamma_p \in \mathbb{R}$, we have

$$\sum_{i,j=1}^{p} \gamma_i \gamma_j \langle f_i, f_j \rangle = \left\langle \sum_{i=1}^{p} \gamma_i f_i, \sum_{j=1}^{p} \gamma_j f_j \right\rangle \geq 0. \tag{16}$$

Here, the left hand equality follows from the bi-linearity of $\langle \cdot, \cdot \rangle$, and the right hand inequality from (15).

By (16), $\langle \cdot, \cdot \rangle$ is a positive definite kernel, defined on our vector space of functions. For the last step in proving that it even is a dot product, we note that by (14), for all functions (13),

$$\langle k(., x), f \rangle = f(x), \tag{17}$$

and in particular

$$\langle k(., x), k(., x') \rangle = k(x, x'). \tag{18}$$

By virtue of these properties, k is called a *reproducing kernel* [Aronszajn, 1950].

Due to (17) and Proposition 4, we have

$$|f(x)|^2 = |\langle k(., x), f \rangle|^2 \leq k(x, x) \cdot \langle f, f \rangle. \tag{19}$$

By this inequality, $\langle f, f \rangle = 0$ implies $f = 0$, which is the last property that was left to prove in order to establish that $\langle ., . \rangle$ is a dot product.

Skipping some details, we add that one can complete the space of functions (13) in the norm corresponding to the dot product, and thus gets a Hilbert space H, called a *reproducing kernel Hilbert space (RKHS)*.

One can define an RKHS as a Hilbert space $\mathcal{H}$ of functions on a set $\mathcal{X}$ with the property that for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the point evaluations $f \mapsto f(x)$ are continuous linear functionals (in particular, all point values $f(x)$ are well defined, which already distinguishes RKHS's from many L_2 Hilbert spaces). From the point evaluation functional, one can then construct the reproducing kernel using the Riesz representation theorem. The Moore-Aronszajn-Theorem [Aronszajn, 1950] states that for every positive definite kernel on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS and vice versa.

There is an analogue of the kernel trick for distances rather than dot products, i.e., dissimilarities rather than similarities. This leads to the larger class of *conditionally positive definite kernels*. Those kernels are defined just like positive definite ones, with the one difference being that their Gram matrices need to satisfy (9) only subject to

$$\sum_{i=1}^{n} c_i = 0. \tag{20}$$

Interestingly, it turns out that many kernel algorithms, including SVMs and kernel PCA (see Section 3), can be applied also with this larger class of kernels, due to their being translation invariant in feature space [Schölkopf and Smola, 2002, Hein et al., 2005].

We conclude this section with a note on terminology. In the early years of kernel machine learning research, it was not the notion of positive definite kernels that was being used. Instead, researchers considered kernels satisfying the conditions of Mercer's theorem [Mercer, 1909], see e.g. [Vapnik, 1998, Cristianini and Shawe-Taylor, 2000]. However, whilst all such kernels do satisfy (3), the converse is not true. Since (3) is what we are interested in, positive definite kernels are thus the right class of kernels to consider.

2.2.3 Properties of Positive Definite Kernels

We begin with some closure properties of the set of positive definite kernels.

Proposition 5 *Below, $k_1, k_2, \ldots$ are arbitrary positive definite kernels on $\mathcal{X} \times \mathcal{X}$, where $\mathcal{X}$ is a nonempty set.*
(i) The set of positive definite kernels is a closed convex cone, i.e., (a) if $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 k_1 + \alpha_2 k_2$ is positive definite; and (b) If $k(x, x') := \lim_{n\to\infty} k_n(x, x')$ exists for all x, x', then k is positive definite.
(ii) The pointwise product $k_1 k_2$ is positive definite.
(iii) Assume that for $i = 1, 2$, k_i is a positive definite kernel on $\mathcal{X}_i \times \mathcal{X}_i$, where $\mathcal{X}_i$ is a nonempty set. Then the tensor product $k_1 \otimes k_2$ and the direct sum $k_1 \oplus k_2$ are positive definite kernels on $(\mathcal{X}_1 \times \mathcal{X}_2) \times (\mathcal{X}_1 \times \mathcal{X}_2)$.

The proofs can be found in [Berg et al., 1984]. We only give a short proof of (ii), a result due to Schur (see e.g., [Berg and Forst, 1975]). Denote by K, L the positive definite Gram matrices of k_1, k_2 with respect to some data set $x_1, \ldots, x_n$. Being positive definite, we can express L as $SS^\top$ in terms of an $n \times n$ matrix S (e.g., the positive definite square root of L). We thus have

$$L_{ij} = \sum_{m=1}^{n} S_{im} S_{jm}, \tag{21}$$

and therefore, for any $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} c_i c_j K_{ij} L_{ij} = \sum_{i,j=1}^{n} c_i c_j K_{ij} \sum_{m=1}^{n} S_{im} S_{jm} = \sum_{m=1}^{n} \sum_{i,j=1}^{n} (c_i S_{im})(c_j S_{jm}) K_{ij} \geq 0, \tag{22}$$

where the inequality follows from (9).

It is reassuring that sums and products of positive definite kernels are positive definite. We will now argue that, loosely speaking, there are no other operations that preserve positive definiteness. To this end, let C denote the set of all functions $\psi\colon \mathbb{R} \to \mathbb{R}$ that map positive definite kernels to (conditionally) positive definite kernels (readers who are not interested in the case of conditionally positive definite kernels may ignore the term in parentheses):

$$C = \{\psi \mid k \text{ is a positive definite kernel } \Rightarrow \psi(k) \text{ is a (conditionally) positive definite kernel}\}. \tag{23}$$

Moreover, define

$$C' = \{\psi \mid \text{for any Hilbert space } \mathcal{F},\ \psi(\langle x, x' \rangle_{\mathcal{F}}) \text{ is (conditionally) positive definite } \} \tag{24}$$

and

$$C'' = \{\psi \mid \text{for all } n \in \mathbb{N}\colon K \text{ is a positive definite } n\times n \text{ matrix } \Rightarrow \psi(K) \text{ is (conditionally) positive definite } \}, \tag{25}$$

where $\psi(K)$ is the $n \times n$ matrix with elements $\psi(K_{ij})$.

Proposition 6 *$C = C' = C''$*

Proof Let $\psi \in C$. For any Hilbert space $\mathcal{F}$, $\langle x, x' \rangle_{\mathcal{F}}$ is positive definite, thus $\psi(\langle x, x' \rangle_{\mathcal{F}})$ is (conditionally) positive definite. Hence $\psi \in C'$.

Next, assume $\psi \in C'$. Consider, for $n \in \mathbb{N}$, an arbitrary positive definite $n \times n$-matrix K. We can express K as the Gram matrix of some $x_1, \ldots, x_n \in \mathbb{R}^n$, i.e., $K_{ij} = \langle x_i, x_j \rangle_{\mathbb{R}^n}$. Since $\psi \in C'$,

we know that $\psi(\langle x, x' \rangle_{\mathbb{R}^n})$ is a (conditionally) positive definite kernel, hence in particular the matrix $\psi(K)$ is (conditionally) positive definite, thus $\psi \in C''$.

Finally, assume $\psi \in C''$. Let k be a positive definite kernel, $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in \mathcal{X}$. Then

$$\sum_{ij} c_i c_j \psi(k(x_i, x_j)) = \sum_{ij} c_i c_j (\psi(K))_{ij} \geq 0 \tag{26}$$

(for $\sum_i c_i = 0$), where the inequality follows from $\psi \in C''$. Therefore, $\psi(k)$ is (conditionally) positive definite, and hence $\psi \in C$. ■

The following proposition follows from a result of FitzGerald et al. [1995] for (conditionally) positive definite matrices; by Proposition 6, it also applies for (conditionally) positive definite kernels, and for functions of dot products. We state the latter case.

Proposition 7 *Let $\psi : \mathbb{R} \to \mathbb{R}$. Then*

$$\psi(\langle x, x' \rangle_{\mathcal{F}}) \tag{27}$$

is positive definite for any Hilbert space $\mathcal{F}$ if and only if ψ is real entire of the form

$$\psi(t) = \sum_{n=0}^{\infty} a_n t^n \tag{28}$$

with $a_n \geq 0$ for $n \geq 0$.

Moreover, $\psi(\langle x, x' \rangle_{\mathcal{F}})$ is conditionally positive definite for any Hilbert space $\mathcal{F}$ if and only if ψ is real entire of the form (28) with $a_n \geq 0$ for $n \geq 1$.

There are further properties of k that can be read off the coefficients a_n.

- Steinwart [2002a] showed that if all a_n are strictly positive, then the kernel (27) is *universal* on every compact subset S of $\mathbb{R}^d$ in the sense that its RKHS is dense in the space of continuous functions on S in the $\|.\|_\infty$ norm. For support vector machines using universal kernels, he then shows (universal) consistency [Steinwart, 2002b]. Examples of universal kernels are (29) and (30) below.

  In Lemma 13, we will show that the a_0 term does not affect an SVM. Hence we infer that it is actually sufficient for consistency to have $a_n > 0$ for $n \geq 1$.

- Let I_{even} and I_{odd} denote the sets of even and odd incides i with the property that $a_i > 0$, respectively. Pinkus [2004] showed that for any $\mathcal{F}$, the kernel (27) is strictly positive definite for any if and only if $a_0 > 0$ and neither I_{even} nor I_{odd} is finite.[2] Moreover, he states that the necessary and sufficient conditions for universality of the kernel are that in addition both $\sum_{i \in I_{\text{even}}} \frac{1}{i}$ and $\sum_{i \in I_{\text{odd}}} \frac{1}{i}$ diverge.

While the proofs of the above statements are fairly involved, one can make certain aspects plausible using simple arguments. For instance [Pinkus, 2004], suppose k is not strictly positive definite. Then we know that for some $c \neq 0$ we have $\sum_{ij} c_i c_j k(x_i, x_j) = 0$. This means that

[2] One might be tempted to think that given some strictly positive definite kernel k with feature map Φ and feature space $\mathcal{H}$, we could set $\mathcal{F}$ to equal $\mathcal{H}$ and consider $\psi(\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}})$. In this case, it would seem that choosing $a_1 = 1$ and $a_n = 0$ for $n \neq 1$ should give us a strictly positive definite kernel which violates Pinkus' conditions. However, the latter kernel is only strictly positive definite on $\Phi(\mathcal{X})$, which is a weaker statement than it being strictly positive definite on all of $\mathcal{H}$.

$\left\langle \sum_i c_i k(x_i,.), \sum_j c_j k(x_j,.) \right\rangle = 0$, implying that $\sum_i c_i k(x_i,.) = 0$. For any $f = \sum_j \alpha_j k(x_j,.)$, this implies that $\sum_i c_i f(x_i) = \left\langle \sum_i c_i k(x_i,.), \sum_j \alpha_j k(x_j,.) \right\rangle = 0$. This equality thus holds for any function in the RKHS. Therefore, the RKHS cannot lie dense in the space of continuous functions on $\mathcal{X}$, and k is thus not universal.

We thus know that if k is universal, it is necessarily strictly positive definite. The latter, in turn, implies that its RKHS $\mathcal{H}$ is infinite dimensional, since otherwise the maximal rank of the Gram matrix of k with respect to a data set in $\mathcal{H}$ would equal its dimensionality. If we assume that $\mathcal{F}$ is finite dimensional, then infinite dimensionality of $\mathcal{H}$ implies that infinitely many of the a_i in (27) must be strictly positive. If only finitely many of the a_i in Proposition 7 were positive, we could construct the feature space $\mathcal{H}$ of $\psi(\langle x, x' \rangle)$ by taking finitely many tensor products and (scaled) direct sums of copies of $\mathcal{F}$, and $\mathcal{H}$ would be finite.[3]

We conclude the section with an example of a kernel which is positive definite by Proposition 7. To this end, let $\mathcal{X}$ be a dot product space. The power series expansion of $\psi(x) = e^x$ then tells us that

$$k(x, x') = e^{\frac{\langle x, x' \rangle}{\sigma^2}} \tag{29}$$

is positive definite [Haussler, 1999]. If we further multiply k with the positive definite kernel $f(x)f(x')$, where $f(x) = e^{-\frac{\|x\|^2}{2\,\sigma^2}}$ and $\sigma > 0$, this leads to the positive definiteness of the Gaussian kernel

$$k'(x, x') = k(x, x')f(x)f(x') = e^{-\frac{\|x - x'\|^2}{2\,\sigma^2}}. \tag{30}$$

2.2.4 Properties of Positive Definite Functions

We now let $\mathcal{X} = \mathbb{R}^d$ and consider positive definite kernels of the form

$$k(x, x') = f(x - x'), \tag{31}$$

in which case f is called a *positive definite function.*

The following characterization is due to Bochner [1933], see also [Rudin, 1962]. We state it in the form given by Wendland [2005].

Theorem 8 *A continuous function f on $\mathbb{R}^d$ is positive definite if and only there exists a finite nonnegative Borel measure μ on $\mathbb{R}^d$ such that*

$$f(x) = \int_{\mathbb{R}^d} e^{-i\langle x, \omega \rangle}\, d\mu(\omega). \tag{32}$$

Whilst normally formulated for complex valued functions, the theorem also holds true for real functions. Note, however, that if we start with an arbitrary nonnegative Borel measure, its Fourier transform may not be real. Real valued positive definite functions are distinguished by the fact that the corresponding measures μ are symmetric.

We may normalize f such that $f(0) = 1$ (hence by Proposition 4 $|f(x)| \leq 1$), in which case μ is a probability measure and f is its characteristic function. For instance, if μ is a normal distribution of the form $(2\pi/\sigma^2)^{-d/2} e^{-\frac{\sigma^2 \|\omega\|^2}{2}} d\omega$, then the corresponding positive definite function is the Gaussian $e^{-\frac{\|x\|^2}{2\sigma^2}}$, cf. (30).

Bochner's theorem allows us to interpret the similarity measure $k(x, x') = f(x - x')$ in the frequency domain. The choice of the measure μ determines which frequency components occur

[3] Recall that the dot product of $\mathcal{F} \otimes \mathcal{F}$ is the square of the original dot product, and the one of $\mathcal{F} \oplus \mathcal{F}'$, where $\mathcal{F}'$ is another dot product space, is the sum of the original dot products.

in the kernel. Since the solutions of kernel algorithms will turn out to be finite kernel expansions, the measure μ will thus determine which frequencies occur in the estimates, i.e., it will determine their regularization properties — more on that in Section 2.3.3 below.

A proof of the theorem can be found for instance in [Rudin, 1962]. One part of the theorem is easy to prove: if f takes the form (32), then

$$\sum_{ij} a_i a_j f(x_i - x_j) = \sum_{ij} a_i a_j \int_{\mathbb{R}^d} e^{-i\langle x_i - x_j, \omega\rangle} \, d\mu(\omega) = \int_{\mathbb{R}^d} \left| \sum_i a_i e^{-i\langle x_i, \omega\rangle} \right|^2 \, d\mu(\omega) \geq 0. \quad (33)$$

Bochner's theorem generalizes earlier work of Mathias, and has itself been generalized in various ways, e.g. by Schoenberg [1938]. An important generalization considers Abelian semigroups (e.g., [Berg et al., 1984]). In that case, the theorem provides an integral representation of positive definite functions in terms of the semigroup's semicharacters. Further generalizations were given by Krein, for the cases of positive definite kernels and functions with a limited number of negative squares (see [Stewart, 1976] for further details and references).

As in the previous section, there are conditions that ensure that the positive definiteness becomes strict.

Proposition 9 *[Wendland, 2005] A positive definite function f is strictly positive definite if the carrier of the measure in the representation (32) contains an open subset.*

This implies that the Gaussian kernel is strictly positive definite.

An important special case of positive definite functions, which includes the Gaussian, are *radial basis functions*. These are functions that can be written as $f(x) = g(\|x\|_2)$ for some function $g : [0, \infty[\to \mathbb{R}$. They have the property of being invariant under the Euclidean group. If we would like to have the additional property of compact support, which is computationally attractive in a number of large scale applications, the following result becomes relevant.

Proposition 10 *[Wendland, 2005] Assume that $g : [0, \infty[\to \mathbb{R}$ is continuous and compactly supported. Then $f(x) = g(\|x\|_2)$ cannot be positive definite on every $\mathbb{R}^d$.*

•[4]

2.2.5 Examples of Kernels

We have above already seen several instances of positive definite kernels, and now intend to complete our selection with a few more examples. In particular, we discuss polynomial kernels, convolution kernels, ANOVA expansions, and kernels on documents.

Polynomial Kernels From Proposition 5 it is clear that homogeneous polynomial kernels $k(x, x') = \langle x, x' \rangle^p$ are positive definite for $p \in \mathbb{N}$ and $x, x' \in \mathbb{R}^d$. By direct calculation we can derive the corresponding feature map [Poggio, 1975]

$$\langle x, x' \rangle^p = \left(\sum_{j=1}^{d} [x]_j [x']_j \right)^p = \sum_{j \in [d]^p} [x]_{j_1} \cdot \cdots \cdot [x]_{j_p} \cdot [x']_{j_1} \cdot \cdots \cdot [x']_{j_p} = \langle C_p(x), C_p(x') \rangle \,, \quad (34)$$

[4] **Bernhard:** universality of pdf functions should follow from standard results of Fourier transf., if μ has full support?

where C_p maps $x \in \mathbb{R}^d$ to the vector $C_p(x)$ whose entries are all possible p-th degree ordered products of the entries of x (note that we use $[d]$ as a shorthand for $\{1, \ldots, d\}$). The polynomial kernel of degree p thus computes a dot product in the space spanned by all monomials of degree p in the input coordinates. Other useful kernels include the inhomogeneous polynomial,

$$k(x, x') = (\langle x, x' \rangle + c)^p \text{ where } p \in \mathbb{N} \text{ and } c \geq 0, \tag{35}$$

which computes all monomials up to degree p.

Spline Kernels It is possible to obtain spline functions as a result of kernel expansions [Smola, 1996, Vapnik et al., 1997] simply by noting that convolution of an even number of indicator functions yields a positive kernel function. Denote by I_X the indicator (or characteristic) function on the set X, and denote by $\otimes$ the convolution operation, $(f \otimes g)(x) := \int_{\mathbb{R}^d} f(x')g(x' - x)dx'$. Then the B-spline kernels are given by

$$k(x, x') = B_{2p+1}(x - x') \text{ where } p \in \mathbb{N} \text{ with } B_{i+1} := B_i \otimes B_0. \tag{36}$$

Here B_0 is the characteristic function on the unit ball[5] in $\mathbb{R}^d$. From the definition of (36) it is obvious that for odd m we may write B_m as inner product between functions $B_{m/2}$. Moreover, note that for even m, B_m is not a kernel.

Convolutions and Structures Let us now move to kernels defined on structured objects [Haussler, 1999, Watkins, 2000]. Suppose the object $x \in \mathcal{X}$ is composed of $x_p \in \mathcal{X}_p$, where $p \in [P]$ (note that the sets $\mathcal{X}_p$ need not be equal). For instance, consider the string $x = ATG$, and $P = 2$. It is composed of the parts $x_1 = AT$ and $x_2 = G$, or alternatively, of $x_1 = A$ and $x_2 = TG$. Mathematically speaking, the set of "allowed" decompositions can be thought of as a *relation* $R(x_1, \ldots, x_P, x)$, to be read as "$x_1, \ldots, x_P$ constitute the composite object x."

Haussler [1999] investigated how to define a kernel between composite objects by building on similarity measures that assess their respective *parts*; in other words, kernels k_p defined on $\mathcal{X}_p \times \mathcal{X}_p$. Define the *R-convolution* of $k_1, \ldots, k_P$ as

$$[k_1 \star \cdots \star k_P](x, x') := \sum_{\bar{x} \in R(x), \bar{x}' \in R(x')} \prod_{p=1}^{P} k_p(\bar{x}_p, \bar{x}'_p), \tag{37}$$

where the sum runs over all possible ways $R(x)$ and $R(x')$ in which we can decompose x into $\bar{x}_1, \ldots, \bar{x}_P$ and x' analogously.[6] If there is only a finite number of ways, the relation R is called finite. In this case, it can be shown that the R-convolution is a valid kernel [Haussler, 1999].

ANOVA Kernels Specific examples of convolution kernels are Gaussians and ANOVA kernels [Wahba, 1990, Burges and Vapnik, 1995, Vapnik, 1998, Stitson et al., 1999]. To construct an ANOVA kernel, we consider $\mathcal{X} = S^N$ for some set S, and kernels $k^{(i)}$ on $S \times S$, where $i = 1, \ldots, N$. For $P = 1, \ldots, N$, the *ANOVA kernel of order* P is defined as

$$k_P(x, x') := \sum_{1 \leq i_1 < \cdots < i_P \leq N} \prod_{p=1}^{P} k^{(i_p)}(x_{i_p}, x'_{i_p}). \tag{38}$$

[5] Note that in $\mathbb{R}$ one typically uses $I_{[-\frac{1}{2}, \frac{1}{2}]}$.

[6] We use the convention that an empty sum equals zero, hence if either x or x' cannot be decomposed, then $(k_1 \star \cdots \star k_P)(x, x') = 0$.

Note that if $P = N$, the sum consists only of the term for which $(i_1, \ldots, i_P) = (1, \ldots, N)$, and k equals the tensor product $k^{(1)} \otimes \cdots \otimes k^{(N)}$. At the other extreme, if $P = 1$, then the products collapse to one factor each, and k equals the direct sum $k^{(1)} \oplus \cdots \oplus k^{(N)}$. For intermediate values of P, we get kernels that lie in between tensor products and direct sums.

ANOVA kernels typically use some moderate value of P, which specifies the order of the interactions between attributes x_{i_p} that we are interested in. The sum then runs over the numerous terms that take into account interactions of order P; fortunately, the computational cost can be reduced to $O(Pd)$ cost by utilizing recurrent procedures for the kernel evaluation [Burges and Vapnik, 1995, Stitson et al., 1999]. ANOVA kernels have been shown to work rather well in multi-dimensional SV regression problems [Stitson et al., 1999].

Brownian Bridge and Related Kernels The Brownian bridge kernel $\min(x, x')$ defined on $\mathbb{R}$ likewise is a positive kernel. Note that one-dimensional linear spline kernels with an infinite number of nodes [Vapnik et al., 1997] for $x, x' \in \mathbb{R}_0^+$ are given by

$$k(x, x') = \frac{\min(x, x')^3}{3} + \frac{\min(x, x')^2 |x - x'|}{2} + 1 + xx'. \tag{39}$$

These kernels can be used as basis functions $k^{(n)}$ in ANOVA expansions. Note that it is advisable to use a $k^{(n)}$ which never or rarely takes the value zero, since a single zero term would eliminate the product in (38).

Bag of Words One way in which SVMs have been used for text categorization [Joachims, 1998] is the *bag-of-words* representation. This maps a given text to a sparse vector, where each component corresponds to a word, and a component is set to one (or some other number) whenever the related word occurs in the text. Using an efficient sparse representation, the dot product between two such vectors can be computed quickly. Furthermore, this dot product is by construction a valid kernel, referred to as a *sparse vector kernel*. One of its shortcomings, however, is that it does not take into account the word ordering of a document. Other sparse vector kernels are also conceivable, such as one that maps a text to the set of pairs of words that are in the same sentence [Joachims, 1998, Watkins, 2000], or those which look only at pairs of words within a certain vicinity with respect to each other [Sim, 2001].

n-grams and Suffix Trees A more sophisticated way of dealing with string data was proposed [Watkins, 2000, Haussler, 1999]. The basic idea is as described above for general structured objects (37): Compare the strings by means of the substrings they contain. The more substrings two strings have in common, the more similar they are. The substrings need not always be contiguous; that said, the further apart the first and last element of a substring are, the less weight should be given to the similarity. Depending on the specific choice of a similarity measure it is possible to define more or less efficient kernels which compute the dot product in the feature space spanned by *all* substrings of documents.

Consider a finite alphabet Σ, the set of all strings of length n, Σ^n, and the set of all finite strings, $\Sigma^* := \cup_{n=0}^{\infty} \Sigma^n$. The length of a string $s \in \Sigma^*$ is denoted by $|s|$, and its elements by $s(1) \ldots s(|s|)$; the concatenation of s and $t \in \Sigma^*$ is written st. Denote by

$$k(x, x') = \sum_s \#(x, s) \#(x', s) c_s$$

a string kernel computed from exact matches. Here $\#(x, s)$ is the number of occurrences of s in x and $c_s \geq 0$.

Vishwanathan and Smola [2004a] prove that for arbitrary c_s the value of the kernel $k(x, x')$ can be computed in $O(|x| + |x'|)$ time and memory. This is done by means of suffix trees, which allow one to find a compact representation of all substrings of x in only $O(|x|)$ time and space. Even better, also $f(x) = \langle w, \Phi(x) \rangle$ can be computed in $O(|x|)$ time if preprocessing linear in the size of the support vectors is carried out.

For inexact matches of a limited degree, typically up to $\epsilon = 3$, and strings of bounded length a similar data structure can be built by explicitly generating a dictionary of strings and their neighborhood in terms of a Hamming distance [Leslie et al., 2002b,a]. These kernels are defined by replacing $\#(x, s)$ by a mismatch function $\#(x, s, \epsilon)$ which reports the number of approximate occurrences of s in x. By trading off computational complexity with storage (hence the restriction to small numbers of mismatches) essentially linear-time algorithms can be designed. Whether a general purpose algorithm exists which allows for efficient comparisons of strings with mismatches in linear time is still an open question. Tools from approximate string matching [Navarro and Raffinot, 1999, Cole and Hariharan, 2000, Gusfield, 1997] promise to be of help in this context.

Mismatch Kernels In the general case it is only possible to find algorithms whose complexity is linear in the lengths of the documents being compared, and the length of the substrings, i.e. $O(|x| \cdot |x'|)$ or worse. We now describe such a kernel with a specific choice of weights, following Lodhi et al. [2000].

Let us now form subsequences u of strings. Given an index sequence $\mathbf{i} := (i_1, \ldots, i_{|u|})$ with $1 \leq i_1 < \cdots < i_{|u|} \leq |s|$, we define $u := s(\mathbf{i}) := s(i_1) \ldots s(i_{|u|})$. We call $l(\mathbf{i}) := i_{|u|} - i_1 + 1$ *the length of the subsequence in* s. Note that if $\mathbf{i}$ is not contiguous, then $l(\mathbf{i}) > |u|$.

The feature space built from strings of length n is defined to be $\mathcal{H}_n := \mathbb{R}^{(\Sigma^n)}$. This notation means that the space has one dimension (or coordinate) for each element of Σ^n, labelled by that element (equivalently, we can think of it as the space of all real-valued functions on Σ^n). We can thus describe the feature map coordinate-wise for each $u \in \Sigma^n$ via

$$[\Phi_n(s)]_u := \sum_{\mathbf{i}: s(\mathbf{i})=u} \lambda^{l(\mathbf{i})}. \tag{40}$$

Here, $0 < \lambda \leq 1$ is a decay parameter: The larger the length of the subsequence in s, the smaller the respective contribution to $[\Phi_n(s)]_u$. The sum runs over all subsequences of s which equal u.

For instance, consider a dimension of $\mathcal{H}_3$ spanned (that is, labelled) by the string `asd`. In this case, we have $[\Phi_3(\texttt{N}\underline{\texttt{as}}\underline{\texttt{d}}\texttt{aq})]_{\texttt{asd}} = \lambda^3$, while $[\Phi_3(\texttt{l}\underline{\texttt{ass d}}\texttt{as})]_{\texttt{asd}} = 2\lambda^5$.[7] The kernel induced by the map Φ_n takes the form

$$k_n(s, t) = \sum_{u \in \Sigma^n} [\Phi_n(s)]_u [\Phi_n(t)]_u = \sum_{u \in \Sigma^n} \sum_{(\mathbf{i},\mathbf{j}): s(\mathbf{i})=t(\mathbf{j})=u} \lambda^{l(\mathbf{i})} \lambda^{l(\mathbf{j})}. \tag{41}$$

The string kernel k_n can be computed using a dynamic programming scheme, see [Watkins, 2000, Lodhi et al., 2000, Durbin et al., 1998].

Locality Improved Kernels They are adjusted for the structure of the data. Recall the Gaussian RBF and polynomial kernels. When applied to an image, it makes no difference whether one uses as x the image or a version of x where all locations of the pixels have been permuted. This indicates that function space on $\mathcal{X}$ induced by k does not take advantage of the *locality* properties of the data.

[7] In the first string, `asd` is a contiguous substring. In the second string, it appears twice as a non-contiguous substring of length 5 in `lass das`, the two occurrences are `l`$\underline{\texttt{ass}}$ $\underline{\texttt{d}}$`as` and `l`$\underline{\texttt{ass}}$ $\underline{\texttt{d}}$`as`.

By taking advantage of the local structure, estimates can be improved. On biological sequences [Zien et al., 2000] one may assign more weight to the entries of the sequence close to the location where estimates should occur. In other words, one replaces $\langle x, x' \rangle$ by $x^\top \Omega x$, where $\Omega \succeq 0$ is a diagonal matrix with largest terms at the location which needs to be classified.

For images, local interactions between image patches need to be considered. One way is to use the *pyramidal* kernel [Schölkopf, 1997, DeCoste and Schölkopf, 2002], inspired by the pyramidal cells of the brain: It takes inner products between corresponding image patches, then raises the latter to some power p_1, and finally raises their sum to another power p_2. This means that mainly short-range interactions are considered and that the long-range interactions are taken with respect to short-range groups.

Tree Kernels We now discuss similarity measures on more structured objects. For trees Collins and Duffy [2001] propose a decomposition method which maps a tree x into its set of subtrees. The kernel between two trees x, x' is then computed by taking a weighted sum of all terms between both trees. In particular, Collins and Duffy [2001] show a quadratic time algorithm, i.e. $O(|x| \cdot |x'|)$ to compute this expression, where $|x|$ is the number of nodes of the tree. When restricting the sum to all proper rooted subtrees it is possible to reduce the computational cost to linear time, i.e. $O(|x| + |x'|)$ by means of a tree to string conversion [Vishwanathan and Smola, 2004a].

Tree Kernels Graphs pose a dual challenge: one may both design a kernel *on* vertices of them and also a kernel *between* them. In the former case, the graph itself becomes the object defining the metric between the vertices. See [Gärtner, 2003, Kashima et al., 2004, 2003] for details on the latter. In the following we discuss kernels *on* graphs.

Denote by $W \in \mathbb{R}^{n \times n}$ the adjacency matrix of a graph with $W_{ij} > 0$ if an edge between i, j exists. Moreover, assume for simplicity that the graph is undirected, that is $W^\top = W$ (see Zhou et al. [2005] for extensions to directed graphs). Denote by $L = D - W$ the graph Laplacian and by $\tilde{L} = \mathbf{1} - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ the normalized graph Laplacian. Here D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$ denoting the degree of vertex i.

It is well known [Fiedler, 1973] that the second largest eigenvector of L approximately decomposes the graph into two parts according to their sign. The other large eigenvectors partition the graph into correspondingly smaller portions. L arises from the fact that for a function f defined on the vertices of the graph $\sum_{i,j} (f(i) - f(j))^2 = 2 f^\top L f$.

Finally, Smola and Kondor [2003] show that under mild conditions and up to rescaling L is the only quadratic permutation invariant form which can be obtained as a linear function of W.

Hence it is reasonable to consider kernel matrices K obtained from L (and $\tilde{L}$). [Smola and Kondor, 2003] suggest kernels $K = r(L)$ or $K = r(\tilde{L})$ and they show that they have desirable smoothness properties. Here $r : [0, \infty) \to [0, \infty)$ is a monotonically decreasing function. Popular choices include

$$r(\xi) = \exp(-\lambda \xi) \qquad \text{diffusion kernel} \tag{42}$$

$$r(\xi) = (\xi + \lambda)^{-1} \qquad \text{regularized graph Laplacian} \tag{43}$$

$$r(\xi) = (\lambda - \xi)^p \qquad p\text{-step random walk} \tag{44}$$

where $\lambda > 0$ is suitably chosen. Eq. 42 was proposed by Kondor and Lafferty [2002]. In Section 2.3.3 we will discuss the connection between regularization operators and kernels in $\mathbb{R}^n$. Without going into details, the function $r(\xi)$ describes the smoothness properties on the graph and L plays the role of the Laplace operator.

Kernels on Sets and Subspaces Whenever each observation x_i consists of a *set* of instances we may use a range of methods to capture the specific properties of these sets:

- Gärtner et al. [2002] take the average of the elements of the set in feature space, that is, $\phi(x_i) = \frac{1}{n}\sum_j \phi(x_{ij})$. This yields good performance in the area of multi-instance learning.
- Jebara and Kondor [2003] extend the idea by dealing with distributions $p_i(x)$ such that $\phi(x_i) = \mathbf{E}[\phi(x)]$ where $x \sim p_i(x)$. This is used with great success in the classification of images with missing pixels.
- Wolf and Shashua [2003] take a different approach by studying the *subspace* spanned by the observations. More specifically, they consider subspace angles [Martin, 2000, Cock and Moor, 2002] between spaces and design a kernel based on this. In a nutshell, if U, U' denote the orthogonal matrices spanning the subspaces of x and x' respectively, then $k(x, x') = \det U^\top U'$.
- Vishwanathan and Smola [2004b] extend this to arbitrary compound matrices (i.e. matrices composed of subdeterminants of matrices) and dynamical systems. Their result exploits the Binet Cauchy theorem which states that compound matrices are a representation of the matrix group, i.e. $C_q(AB) = C_q(A)C_q(B)$. This leads to kernels of the form $k(x, x') = \operatorname{tr} C_q(AB)$. Note that for $q = \dim A$ we recover the determinant kernel of Wolf and Shashua [2003].

Fisher Kernels In their quest to to make density estimates directly accessible to kernel methods Jaakkola and Haussler [1999b,a] designed kernels which work directly on probability density estimates $p(x|\theta)$. Denote by

$$U_\theta(x) := \partial_\theta - \log p(x|\theta) \tag{45}$$

$$I := \mathbf{E}_x\left[U_\theta(x)U_\theta^\top(x)\right] \tag{46}$$

the Fisher scores and the Fisher information matrix respectively. Note that for maximum likelihood estimators $\mathbf{E}_x[U_\theta(x)] = 0$ and therefore I is the covariance of $U_\theta(x)$. The Fisher kernel is defined as

$$k(x, x') := U_\theta^\top(x) I^{-1} U_\theta(x') \text{ or } k(x, x') := U_\theta^\top(x) U_\theta(x') \tag{47}$$

depending on whether whether we study the normalized or the unnormalized kernel respectively. It is a versatile tool to re-engineer existing density estimators for the purpose of discriminative estimation.

In addition to that, it has several attractive theoretical properties: Oliver et al. [2000] show that estimation using the normalized Fisher kernel corresponds to estimation subject to a regularization on the $L_2(p(\cdot|\theta))$ norm.

Moreover, in the context of exponential families (see Section 4.1 for a more detailed discussion) where $p(x|\theta) = \exp(\langle\phi(x), \theta\rangle - g(\theta))$ we have

$$k(x, x') = [\phi(x) - \partial_\theta g(\theta)]\,[\phi(x') - \partial_\theta g(\theta)]\,. \tag{48}$$

for the unnormalized Fisher kernel. This means that up to centering by $\partial_\theta g(\theta)$ the Fisher kernel is identical to the kernel arising from the inner product of the sufficient statistics $\phi(x)$. This is not a coincidence. In fact, in our analysis of nonparametric exponential families we will encounter this fact several times. See Section 4 for further details. Moreover, note that the centering is immaterial, as can be seen in Lemma 13.

While we presented an overview over recent kernel design in this section the overview is by no means complete. The reader is referred to textbooks by Cristianini and Shawe-Taylor [2000], Shawe-Taylor and Cristianini [2004], Schölkopf and Smola [2003], Joachims [2002], Herbrich [2002] for further details on how kernels may be constructed.

2.3 Kernel Function Classes

2.3.1 The Representer Theorem

From kernels, we now move to functions that can be expressed in terms of kernel expansions. The representer theorem [Kimeldorf and Wahba, 1971, Cox and O'Sullivan, 1990] shows that solutions of a large class of optimization problems can be expressed as kernel expansions over the sample points. We present a slightly more general version of the theorem with a simple proof [Schölkopf et al., 2001]. As above, $\mathcal{H}$ is the RKHS associated to the kernel k.

Theorem 11 (Representer Theorem) *Denote by $\Omega : [0, \infty) \to \mathbb{R}$ a strictly monotonic increasing function, by $\mathcal{X}$ a set, and by $c : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$ an arbitrary loss function. Then each minimizer $f \in \mathcal{H}$ of the regularized risk functional*

$$c\left((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))\right) + \Omega\left(\|f\|_{\mathcal{H}}^2\right) \tag{49}$$

admits a representation of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x). \tag{50}$$

Proof We decompose any $f \in \mathcal{H}$ into a part contained in the span of the kernel functions $k(x_1, \cdot), \cdots, k(x_n, \cdot)$, and one in the orthogonal complement;

$$f(x) = f_{\|}(x) + f_{\perp}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) + f_{\perp}(x). \tag{51}$$

Here $\alpha_i \in \mathbb{R}$ and $f_{\perp} \in \mathcal{H}$ with $\langle f_{\perp}, k(x_i, \cdot)\rangle_{\mathcal{H}} = 0$ for all $i \in [n] := \{1, \ldots, n\}$. By (17) we may write $f(x_j)$ (for all $j \in [n]$) as

$$f(x_j) = \langle f(\cdot), k(x_j, .)\rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x_j) + \langle f_{\perp}(\cdot), k(x_j, .)\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i k(x_i, x_j). \tag{52}$$

Second, for all $f_{\perp}$,

$$\Omega(\|f\|_{\mathcal{H}}^2) = \Omega\left(\left\|\sum_{i}^{n} \alpha_i k(x_i, \cdot)\right\|_{\mathcal{H}}^2 + \|f_{\perp}\|_{\mathcal{H}}^2\right) \geq \Omega\left(\left\|\sum_{i}^{n} \alpha_i k(x_i, \cdot)\right\|_{\mathcal{H}}^2\right). \tag{53}$$

Thus for any fixed $\alpha_i \in \mathbb{R}$ the risk functional (49) is minimized for $f_{\perp} = 0$. Since this also has to hold for the solution, the theorem holds. ■

Monotonicity of Ω does not prevent the regularized risk functional (49) from having multiple local minima. To ensure a global minimum, we would need to require convexity. If we discard the strictness of the monotonicity, then it no longer follows that each minimizer of the regularized risk admits an expansion (50); it still follows, however, that there is always another solution that is as good, and that *does* admit the expansion.

The significance of the Representer Theorem is that although we might be trying to solve an optimization problem in an infinite-dimensional space $\mathcal{H}$, containing linear combinations of kernels centered on *arbitrary* points of $\mathcal{X}$, it states that the solution lies in the span of n particular kernels — those centered on the training points. We will encounter (50) again further below, where it is called the *Support Vector expansion*. For suitable choices of loss functions, many of the α_i often equal 0.

2.3.2 Reduced Set Methods

Nevertheless, it can often be the case that the expansion (50) contains a large number of terms. This can be problematic in practical applications, since the time required to evaluate (50) is proportional to the number of terms. To deal with this issue, we thus need to express or at least approximate (50) using a smaller number of terms. Exact expressions in a reduced number of basis points are often not possible; e.g., if the kernel is strictly positive definite and the data points $x_1, \ldots, x_n$ are distinct, then no exact reduction is possible. For the purpose of approximation, we need to decide on a discrepancy measure that we want to minimize. The most natural such measure is the norm of the RKHS. From a practical point of view, this norm has the advantage that it can be computed using evaluations of the kernel function which opens the possibility for relatively efficient approximation algorithms. Clearly, if we were given a set of points $z_1, \ldots, z_m$ such that the function (cf. (50))

$$f = \sum_{i=1}^{n} \alpha_i k(x_i, .). \tag{54}$$

can be expressed with small RKHS-error in the subspace spanned by $k(z_1, .), \ldots, k(z_m, .)$, then we can compute an m-term approximation of f by projection onto that subspace.

The problem thus reduces to the choice of the expansion points $z_1, \ldots, z_m$. In the literature, two classes of methods have been pursued. In the first, it is attempted to choose the points as a subset of some larger set $x_1, \ldots, x_n$, which is usually the training set [Schölkopf, 1997, Frieß and Harrison, 1998]. In the second, we compute the difference in the RKHS between (54) and the *reduced set expansion* [Burges, 1996]

$$g = \sum_{p=1}^{m} \beta_p k(z_p, .), \tag{55}$$

leading to

$$\|f - g\|^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) + \sum_{p,q=1}^{m} \beta_p \beta_q k(z_p, z_q) - 2 \sum_{i=1}^{n} \sum_{p=1}^{m} \alpha_i \beta_p k(x_i, z_p). \tag{56}$$

To obtain the $z_1, \ldots, z_m$ and $\beta_1, \ldots, \beta_m$, we can minimize this objective function, which for most kernels will be a nonconvex optimization problem. Specialized methods for specific kernels exist, as well as iterative methods taking into account the fact that given the $z_1, \ldots, z_m$, the coefficients $\beta_1, \ldots, \beta_m$ can be computed in closed form. See [Schölkopf and Smola, 2002] for further details.

Once we have constructed or selected a set of expansion points, we are working in a finite dimensional subspace of $\mathcal{H}$ spanned by these points.

It is known that each closed subspace of an RKHS is itself an RKHS. Denote its kernel by l, then the projection P onto the subspace is given by (e.g., [Meschkowski, 1962])

$$(Pf)(x') = \langle f, l(., x') \rangle . \tag{57}$$

If the subspace is spanned by a finite set of linearly independent kernels $k(., z_1), \ldots, k(., z_m)$, the kernel of the subspace takes the form

$$k_m(x, x') = (k(x, z_1), \ldots, k(x, z_m))\, K^{-1}\, (k(x', z_1), \ldots, k(x', z_m))^\top , \tag{58}$$

where K is the Gram matrix of k with respect to $z_1, \ldots, z_m$. This can be seen by noting that (i) the $k_m(., x)$ span an m-dimensional subspace, and (ii) we have $k_m(z_i, z_j) = k(z_i, z_j)$ for

$i, j = 1, \ldots, m$. The kernel of the subspace's orthogonal complement, on the other hand, is $k - k_m$.

The projection onto the m-dimensional subspace can be written

$$P = \sum_{i,j=1}^{m} \left(K^{-1}\right)_{ij} k(., z_i) k(., z_j)^\top, \tag{59}$$

where $k(., z_j)^\top$ denotes the linear form mapping $k(., x)$ to $k(z_j, x)$ for $x \in \mathcal{X}$.

If we prefer to work with coordinates and the Euclidean dot product, we can use the *kernel PCA map* [Schölkopf and Smola, 2002]

$$\Phi_m : \mathcal{X} \to \mathbb{R}^m, \; x \mapsto K^{-1/2} \left(k(x, z_1), \ldots, k(x, z_m)\right)^\top, \tag{60}$$

which directly projects into an orthonormal basis of the subspace.[8]

A number of articles exist discussing how to best choose the expansion points z_i for the purpose of processing a given dataset, e.g., [Smola and Schölkopf, 2000, Williams and Seeger, 2000].

2.3.3 Regularization Properties

The regularizer $\|f\|^2_{\mathcal{H}}$ used in Theorem 11 stems from the dot product $\langle f, f \rangle_k$ in the RKHS $\mathcal{H}$ associated with a positive definite kernel. Since it is somewhat hard to interpret which aspects of f are influencing the value of this dot product, it is instructive to consider the cases where there exists a positive operator Υ mapping the RKHS $\mathcal{H}$ into a function space endowed with the dot product

$$\langle f, g \rangle = \int f(x) \overline{g(x)} \, dx, \tag{61}$$

with the property that for all $f, g \in \mathcal{H}$

$$\langle \Upsilon f, \Upsilon g \rangle = \langle f, g \rangle_k . \tag{62}$$

We shall now sketch this, with a particular focus on the case where the dot product can be written conveniently in the Fourier domain. Our exposition will be informal (see also [Poggio and Girosi, 1990, Girosi et al., 1993, Smola et al., 1998]), and we will implicitly assume that all integrals are over $\mathbb{R}^d$ and exist, and that the operators are well defined. Rather than (62), we consider the condition

$$\langle \Upsilon k(x, .), \Upsilon k(x', .) \rangle = \langle k(x, .), k(x', .) \rangle_k , \tag{63}$$

which by linearity extends to (62) (cf. Section 2.2.2).

Let us rewrite the first dot product as $\left\langle \Upsilon^2 k(x, .), k(x', .) \right\rangle$ (recall that Υ is positive and thus symmetric). We then see that if $k(x, .)$ is a *Greens function* of Υ^2, we have

$$\left\langle \Upsilon^2 k(x, .), k(x', .) \right\rangle = \langle \delta_x, k(x', .) \rangle = k(x, x'), \tag{64}$$

which by the reproducing property (18) amounts to the desired equality (63). For conditionally positive definite kernels, a similar correspondence can be established, with a regularization operator whose null space is spanned by a set of functions which are not regularized (in the case (20), which is sometimes called *conditionally positive definite of order 1*, these are the constant functions).

[8] In case K does not have full rank, projections onto the span $\{k(\cdot, z_1), \ldots, k(\cdot, z_m)\}$ are achieved by using the pseudoinverse of K instead.

We now consider the particular case where the kernel can be written $k(x, x') = f(x - x')$ with a continuous strictly positive definite function $f \in L_1(\mathbb{R}^d)$ (cf. Section 2.2.4). A variation of Bochner's theorem (Theorem 8), stated in Wendland [2005], then tells us that the measure corresponding to f has a nonvanishing density υ with respect to the Lebesgue measure, i.e., that k can be written as

$$k(x, x') = \int e^{-i\langle x-x',\omega\rangle} \upsilon(\omega) d\omega = \int e^{-i\langle x,\omega\rangle} \overline{e^{-i\langle x',\omega\rangle}} \upsilon(\omega) d\omega. \tag{65}$$

We would like to rewrite this as $\langle \Upsilon k(x, .), \Upsilon k(x', .)\rangle$ for some linear operator Υ. It turns out that a multiplication operator in the Fourier domain will do the job. To this end, recall the d-dimensional Fourier transform, given by

$$F[f](\omega) := (2\pi)^{-\frac{d}{2}} \int f(x) e^{-i\langle x,\omega\rangle} dx, \tag{66}$$

with the inverse

$$F^{-1}[f](x) = (2\pi)^{-\frac{d}{2}} \int f(\omega) e^{i\langle x,\omega\rangle} d\omega. \tag{67}$$

Next, compute the Fourier transform of k as

$$\begin{aligned} F[k(x, .)](\omega) &= (2\pi)^{-\frac{d}{2}} \int \int \left(\upsilon(\omega') e^{-i\langle x,\omega'\rangle} \right) e^{i\langle x',\omega'\rangle} d\omega' e^{-i\langle x',\omega\rangle} dx' && (68) \\ &= (2\pi)^{\frac{d}{2}} \upsilon(\omega) e^{-i\langle x,\omega\rangle}. && (69) \end{aligned}$$

Hence we can rewrite (65) as

$$k(x, x') = (2\pi)^{-d} \int \frac{F[k(x, .)](\omega) \overline{F[k(x', .)](\omega)}}{\upsilon(\omega)} d\omega. \tag{70}$$

If our regularization operator maps

$$\Upsilon : f \mapsto (2\pi)^{-\frac{d}{2}} \upsilon^{-\frac{1}{2}} F[f], \tag{71}$$

we thus have

$$k(x, x') = \int (\Upsilon k(x, .))(\omega) \overline{(\Upsilon k(x', .))(\omega)} d\omega, \tag{72}$$

i.e., our desired identity (63) holds true.

As required in (62), we can thus interpret the dot product $\langle f, g\rangle_k$ in the RKHS as a dot product $\int (\Upsilon f)(\omega) \overline{(\Upsilon g)(\omega)} d\omega$. This allows us to understand regularization properties of k in terms of its (scaled) Fourier transform $\upsilon(\omega)$. *Small* values of $\upsilon(\omega)$ amplify the corresponding frequencies in (71). Penalizing $\langle f, f\rangle_k$ thus amounts to a *strong* attenuation of the corresponding frequencies. Hence small values of $\upsilon(\omega)$ for large $\|\omega\|$ are desirable, since high frequency components of $F[f]$ correspond to rapid changes in f. It follows that $\upsilon(\omega)$ describes the filter properties of the corresponding regularization operator Υ. In view of our comments following Theorem 8, we can translate this insight into probabilistic terms: if the probability measure $\frac{\upsilon(\omega)d\omega}{\int \upsilon(\omega)d\omega}$ describes the desired filter properties, then the natural translation invariant kernel to use is the characteristic function of the measure.

2.3.4 Remarks and Notes

The notion of kernels as dot products in Hilbert spaces was brought to the field of machine learning by Aizerman et al. [1964], Boser et al. [1992], Vapnik [1998], Schölkopf et al. [1998]. Aizerman et al. [1964] used kernels as a tool in a convergence proof, allowing them to apply the Perceptron convergence theorem to their class of potential function algorithms. To the best of our knowledge, Boser et al. [1992] were the first to use kernels to construct a nonlinear estimation algorithm, the hard margin predecessor of the Support Vector Machine, from its linear counterpart, the *generalized portrait* [Vapnik and Lerner, 1963, Vapnik, 1982]. Whilst all these uses were limited to kernels defined on vectorial data, Schölkopf [1997] observed that this restriction is unnecessary, and nontrivial kernels on other data types were proposed by Haussler [1999], Watkins [2000]. Schölkopf et al. [1998] applied the kernel trick to generalize principal component analysis and pointed out the (in retrospect obvious) fact that any algorithm which only uses the data via dot products can be generalized using kernels.

In addition to the above uses of positive definite kernels in machine learning, there has been a parallel, and partly earlier development in the field of statistics, where such kernels have been used for instance for time series analysis [Parzen, 1970] as well as regression estimation and the solution of inverse problems [Wahba, 1990].

In probability theory, positive definite kernels have also been studied in depth since they arise as covariance kernels of stochastic processes, see e.g. [Loève, 1978]. This connection is heavily being used in a subset of the machine learning community interested in prediction with Gaussian Processes [Rasmussen and Williams, 2006].

In functional analysis, the problem of Hilbert space representations of kernels has been studied in great detail; a good reference is [Berg et al., 1984]; indeed, a large part of the material in the present section is based on that work. Interestingly, it seems that for a fairly long time, there have been two separate strands of development [Stewart, 1976]. One of them was the study of positive definite functions, which started later but seems to have been unaware of the fact that it considered a special case of positive definite kernels. The latter was initiated by Hilbert [1904], Mercer [1909], and pursued for instance by Schoenberg [1938]. Hilbert calls a kernel k *definit* if it satisfies

$$\int_a^b \int_a^b k(x, x') f(x) f(x') dx dx' > 0 \tag{73}$$

for all nonzero continuous functions f, and shows that all eigenvalues of the corresponding integral operator $f \mapsto \int_a^b k(x, .) f(x) dx$ are then positive. If k satisfies the condition (73) subject to the constraint that

$$\int_a^b f(x) g(x) dx = 0, \tag{74}$$

for some fixed function g, Hilbert calls it *relativ definit*. For that case, he shows that k has at most one negative eigenvalue. Note that if f is chosen to be constant, then this notion is closely related to the one of conditionally positive definite kernels, cf. (20).

For further historical remarks, cf. the review of Stewart [1976] or the book of Berg et al. [1984].

3 Convex Programming Methods for Estimation

As we saw, kernels can be used both for the purpose of describing nonlinear functions subject to smoothness constraints and for the purpose of computing inner products in some feature space efficiently. In this section we focus on the latter and how it allows us to design methods of estimation based on the geometry of the problems at hand.

3.1 Support Vector Estimation

Given a sample $Z := \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$ we now aim at finding an affine function $f(x) = \langle w, x \rangle + b$ such that $\operatorname{sgn} f(x) = y$ for all $(x, y) \in Z$ or if this is not possible, such that $\mathbf{E}_{\text{emp}}[\{yf(x) < 0\}]$ is minimized. This problem is hard for two reasons:

- Minimization of the empirical risk with respect to (w, b) is NP-hard [Minsky and Papert, 1969]. In fact, Ben-David et al. [2003] show that even approximately minimizing the empirical risk is NP-hard, not only for linear function classes but also for spheres and other simple geometrical objects. This means that even if the statistical challenges could be solved, we still would be saddled with a formidable algorithmic problem.
- The indicator function $\{yf(x) < 0\}$ is discontinuous and even small changes in f may lead to large changes in both empirical and expected risk. Properties of such functions can be captured by the VC-dimension [Vapnik and Chervonenkis, 1971], that is, the maximum number of observations which can be labeled in an arbitrary fashion by functions of the class. Necessary and sufficient conditions for estimation can be stated in these terms Vapnik and Chervonenkis [1991]. However, much tighter bounds can be obtained by using the scale of the class, too [Alon et al., 1993, Bartlett et al., 1996, Williamson et al., 2001]. In fact, there exist function classes parameterized by a scalar which have infinite VC-dimension [Vapnik, 1995].

Given the difficulty arising from minimizing the empirical risk of misclassification, we now discuss algorithms which minimize an upper bound on the empirical risk, while providing good computational properties and consistency of the estimators. The statistical analysis is relegated to Section 3.2.

3.1.1 Support Vector Classification

Assume that Z is separable, i.e. there exists a linear function $f(x)$ such that $\operatorname{sgn} yf(x) = 1$ on Z. In this case, the task of finding a large margin separating hyperplane can be viewed as one of solving [Vapnik and Lerner, 1963]

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\|w\|^2 \ \text{ subject to } \ y_i\left(\langle w, x\rangle + b\right) \geq 1. \tag{75}$$

Note that $\|w\|^{-1} f(x_i)$ is the distance of the point x_i to the hyperplane $H(w, b) := \{x | \langle w, x\rangle + b = 0\}$. The condition $y_i f(x_i) \geq 1$ implies that the margin of separation is at least $2\|w\|^{-1}$. The bound becomes exact if equality is attained for some $y_i = 1$ and $y_j = -1$. Consequently minimizing $\|w\|$ subject to the constraints maximizes the margin of separation. Eq. (75) is a quadratic program which can be solved efficiently [Luenberger, 1984, Fletcher, 1989, Boyd and Vandenberghe, 2004].

Mangasarian [1965] devised a similar optimization scheme using $\|w\|_1$ instead of $\|w\|_2$ in the objective function of (75). The result is a *linear* program. In general, one may show [Smola et al., 2000] that minimizing the ℓ_p norm of w leads to the maximizing of the margin of separation in

the ℓ_q norm where $\frac{1}{p} + \frac{1}{q} = 1$. The ℓ_1 norm leads to sparse approximation schemes (see also Chen et al. [1999]), whereas the ℓ_2 norm can be extended to Hilbert spaces and kernels.

To deal with nonseparable problems, i.e. cases when (75) is infeasible, we need to relax the constraints of the optimization problem. Bennett and Mangasarian [1992] and Cortes and Vapnik [1995] impose a linear penalty on the violation of the large-margin constraints to obtain:

$$\underset{w,b,\xi}{\text{minimize}} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \text{ subject to } y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0. \tag{76}$$

Eq. (76) is a quadratic program which is always feasible (e.g. $w, b = 0$ and $\xi_i = 1$ satisfy the constraints). $C > 0$ is a regularization constant trading off the violation of the constraints vs. maximizing the overall margin.

Whenever the dimensionality of $\mathcal{X}$ exceeds n direct optimization of (76) is computationally inefficient. This is particularly true if we map from $\mathcal{X}$ into an RKHS. To address these problems one may solve the problem in dual space as follows. The Lagrange function of (76) is given by

$$L(w, b, \xi, \alpha, \eta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i\left(\langle w, x_i\rangle + b\right)\right) - \sum_{i=1}^{n}\eta_i\xi_i \tag{77}$$

where $\alpha_i, \eta_i \geq 0$ for all $i \in [n]$. To compute the dual of L we need to identify the first order conditions in w, b. They are given by

$$\partial_w L = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \text{ and } \partial_b L = -\sum_{i=1}^{n}\alpha_i y_i = 0 \text{ and } \partial_{\xi_i} L = C - \alpha_i + \eta_i = 0 \tag{78}$$

This translates into $w = \sum_{i=1}^{n}\alpha_i y_i x_i$, the linear constraint $\sum_{i=1}^{n}\alpha_i y_i = 0$, and the box-constraint $\alpha_i \in [0, C]$ arising from $\eta_i \geq 0$. Substituting (78) into L yields the Wolfe [1961] dual

$$\underset{\alpha}{\text{minimize}} \ \frac{1}{2}\alpha^\top Q\alpha - \alpha^\top 1 \text{ subject to } \alpha^\top y = 0 \text{ and } \alpha_i \in [0, C]. \tag{79}$$

$Q \in \mathbb{R}^{n\times n}$ is the matrix of inner products $Q_{ij} := y_i y_j \langle x_i, x_j\rangle$. Clearly this can be extended to feature maps and kernels easily via $K_{ij} := y_i y_j \langle \Phi(x_i), \Phi(x_j)\rangle = y_i y_j k(x_i, x_j)$. Note that w lies in the span of the x_i. This is an instance of the Representer Theorem (Th. 11). The KKT conditions [Karush, 1939, Kuhn and Tucker, 1951, Boser et al., 1992, Cortes and Vapnik, 1995] require that at optimality $\alpha_i(y_i f(x_i) - 1) = 0$. This means that only those x_i may appear in the expansion (78) for which $y_i f(x_i) \leq 1$, as otherwise $\alpha_i = 0$. The x_i are commonly referred to as Support Vectors.

Note that $\sum_{i=1}^{n}\xi_i$ is an upper bound on the empirical risk, as $y_i f(x_i) \leq 0$ implies $\xi_i \geq 1$ (see also Lemma 12). The number of misclassified points x_i itself depends on the configuration of the data and the value of C. The result of [Ben-David et al., 2003] suggests that finding a even an approximate minimum classification error solution is difficult. That said, it is possible to modify (76) such that a desired target number of observations violates $y_i f(x_i) \geq \rho$ for some $\rho \in \mathbb{R}$ by making the threshold itself a variable of the optimization problem [Schölkopf et al., 2000]. This leads to the following optimization problem (ν-SV classification):

$$\underset{w,b,\xi}{\text{minimize}} \ \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\xi_i - n\nu\rho \text{ subject to } y_i\left(\langle w, x_i\rangle + b\right) \geq \rho - \xi_i \text{ and } \xi_i \geq 0. \tag{80}$$

The dual of (80) is essentially identical to (79) with the exception of an additional constraint:

$$\underset{\alpha}{\text{minimize}} \ \frac{1}{2}\alpha^\top Q\alpha \text{ subject to } \alpha^\top y = 0 \text{ and } \alpha^\top 1 = n\nu \text{ and } \alpha_i \in [0,1]. \tag{81}$$

One can show that for every C there exists a ν such that the solution of (81) is a multiple of the solution of (79). Schölkopf et al. [2000] prove that solving (81) for which $\rho > 0$ satisfies:

1. ν is an upper bound on the fraction of margin errors.
2. ν is a lower bound on the fraction of SVs.

Moreover, under mild conditions with probability 1, asymptotically, ν equals both the fraction of SVs and the fraction of errors.

This statement implies that whenever the data are sufficiently well separable (that is $\rho > 0$), ν-SV Classification finds a solution with a fraction of at most ν margin errors. Also note that for $\nu = 1$, all $\alpha_i = 1$, that is, f becomes an affine copy of the Parzen windows classifier (6).

3.1.2 Estimating the Support of a Density

We now extend the notion of linear separation to that of estimating the support of a density [Schölkopf et al., 2001, Tax and Duin, 1999]. Denote by $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ the sample drawn iid from $\mathrm{P}(x)$. Let $\mathcal{C}$ be a class of measurable subsets of $\mathcal{X}$ and let λ be a real-valued function defined on $\mathcal{C}$. The *quantile function* [Einmal and Mason, 1992] with respect to $(\mathrm{P}, \lambda, \mathcal{C})$ is defined as

$$U(\mu) = \inf\{\lambda(C) | \mathrm{P}(C) \geq \mu, C \in \mathcal{C}\} \ \text{ where } \mu \in (0,1]. \tag{82}$$

We denote by $C_\lambda(\mu)$ and $C^m_\lambda(\mu)$ the (not necessarily unique) $C \in \mathcal{C}$ that attain the infimum (when it is achievable) on $\mathrm{P}(x)$ and on the empirical measure given by X respectively. A common choice of λ is Lebesgue measure, in which case $C_\lambda(\mu)$ is the minimum volume set $C \in \mathcal{C}$ that contains at least a fraction μ of the probability mass.

Support estimation requires us to find some $C^m_\lambda(\mu)$ such that $|\mathrm{P}\left(C^m_\lambda(\mu)\right) - \mu|$ is small. This is where the complexity trade-off enters: On the one hand, we want to use a rich class $\mathcal{C}$ to capture all possible distributions, on the other hand large classes lead to large deviations between μ and $\mathrm{P}\left(C^m_\lambda(\mu)\right)$. Therefore, we have to consider classes of sets which are suitably restricted. This can be achieved using an SVM regularizer.

In the case where $\mu < 1$, it seems the first work was reported in [Sager, 1979, Hartigan, 1987], in which $\mathcal{X} = \mathbb{R}^2$, with $\mathcal{C}$ being the class of closed convex sets in $\mathcal{X}$. Nolan [1991] considered higher dimensions, with $\mathcal{C}$ being the class of ellipsoids. Tsybakov [1997] studied an estimator based on piece-wise polynomial approximation of $C_\lambda(\mu)$ and showed it attains the asymptotically minimax rate for certain classes of densities. Polonik [1997] studied the estimation of $C_\lambda(\mu)$ by $C^m_\lambda(\mu)$. He derived asymptotic rates of convergence in terms of various measures of richness of $\mathcal{C}$. More information on minimum volume estimators can be found in that work, and in [Schölkopf et al., 2001].

SV support estimation works by using SV support estimation relates to previous work as follows: set $\lambda(C_w) = \|w\|^2$, where $C_w = \{x | f_w(x) \geq \rho\}$, and (w, ρ) are respectively a weight vector and an offset with $f_w(x) = \langle w, x\rangle$. Stated as a convex optimization problem we want to separate the data from the origin with maximum margin via:

$$\underset{w,\xi,\rho}{\text{minimize}} \ \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\xi_i - n\nu\rho \text{ subject to } \langle w, x_i\rangle \geq \rho - \xi_i \text{ and } \xi_i \geq 0. \tag{83}$$

Here, $\nu \in (0,1]$ plays the same role as in (80), controlling the number of observations x_i for which $f(x_i) \leq \rho$. Since nonzero slack variables ξ_i are penalized in the objective function, if w and ρ solve this problem, then the decision function $f(x)$ will attain or exceed ρ for at least $1-\nu$ instances x_i contained in X while the regularization term $\|w\|$ will still be small. The dual of (83) yields:

$$\underset{\alpha}{\text{minimize}} \ \frac{1}{2}\alpha^\top K\alpha \text{ subject to } \alpha^\top 1 = \nu n \text{ and } \alpha_i \in [0,1] \tag{84}$$

To compare (84) to a Parzen windows estimator assume that k is such that it can be normalized as a density in input space, such as a Gaussian. Using $\nu = 1$ in (84) the constraints automatically imply $\alpha_i = 1$. Thus f reduces to a Parzen windows estimate of the underlying density. For $\nu < 1$, the equality constraint (84) still ensures that f is a thresholded density, now depending only on a *subset* of X — those which are important for the decision $f(x) \leq \rho$ to be taken.

3.1.3 Regression Estimation

SV regression was first proposed in [Vapnik, 1995, Vapnik et al., 1997, Drucker et al., 1997] using the so-called ϵ-insensitive loss function. It is a direct extension of the soft-margin idea to regression: instead of requiring that $yf(x)$ exceeds some margin value, we now require that the values $y - f(x)$ are bounded by a margin on both sides. That is, we impose the constraints

$$y_i - f(x_i) \leq \epsilon_i - \xi_i \text{ and } f(x_i) - y_i \leq \epsilon_i - \xi_i^*. \tag{85}$$

where $\xi_i, \xi_i^* \geq 0$. If $|y_i - f(x_i)| \leq \epsilon$ no penalty occurs. The objective function is given by the sum of the slack variables ξ_i, ξ_i^* penalized by some $C > 0$ and a measure for the slope of the function $f(x) = \langle w, x\rangle + b$, that is $\frac{1}{2}\|w\|^2$.

Before computing the dual of such an optimization problem let us consider a somewhat more general situation where we use a range of different convex penalties for the deviation between y_i and $f(x_i)$. One may check that minimizing $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i + \xi_i^*$ subject to (85) is equivalent to solving

$$\underset{w,b,\xi}{\text{minimize}} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \psi(y_i - f(x_i)) \text{ where } \psi(\xi) = \max(0, |\xi| - \epsilon). \tag{86}$$

Choosing different loss functions ψ leads to a rather rich class of estimators:

- $\psi(\xi) = \frac{1}{2}\xi^2$ yields penalized least mean squares (LMS) regression [Hoerl and Kennard, 1970, Tikhonov, 1963, Morozov, 1984, Wahba, 1990]. The corresponding optimization problem can be minimized by solving a linear system.
- For $\psi(\xi) = |\xi|$ we obtain the penalized least absolute deviations (LAD) estimator [Bloomfield and Steiger, 1983]. That is, we obtain a quadratic program to estimate the conditional median.
- A combination of LMS and LAD loss yields a penalized version of Huber's robust regression [Huber, 1981, Smola and Schölkopf, 1998]. In this case we have $\psi(\xi) = \frac{1}{2\sigma}\xi^2$ for $|\xi| \leq \sigma$ and $\psi(\xi) = |\xi| - \frac{\sigma}{2}$ for $|\xi| \geq \sigma$.
- Note that also quantile regression [Koenker, 2005] can be modified to work with kernels [Schölkopf et al., 2000, Le et al., 2005] by using as loss function the "pinball" loss, that is $\psi(\xi) = (1-\tau)\psi$ if $\psi < 0$ and $\psi(\xi) = \tau\psi$ if $\psi > 0$.

All the optimization problems arising from the above five cases are convex quadratic programs. By and large their dual resembles that of (85), namely

$$\underset{\alpha,\alpha^*}{\text{minimize}} \frac{1}{2}(\alpha-\alpha^*)^\top K(\alpha-\alpha^*) + \epsilon^\top(\alpha+\alpha^*) - y^\top(\alpha-\alpha^*) \tag{87a}$$

$$\text{subject to } (\alpha-\alpha^*)^\top 1 = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C]. \tag{87b}$$

Here $K_{ij} = \langle x_i, x_j\rangle$ for linear models and $K_{ij} = k(x_i, x_j)$ if we map $x \to \Phi(x)$. The ν-trick, as described in (80) [Schölkopf et al., 2000], can be extended to regression, allowing one to choose the margin of approximation automatically. In this case (87a) drops the terms in ϵ. In its place, we add a linear constraint $(\alpha-\alpha^*)^\top 1 = \nu n$. Likewise, LAD is obtained from (87) by dropping the terms in ϵ without additional constraints. Robust regression leaves (87) unchanged, however, in the definition of K we have an additional term of σ^{-1} on the main diagonal. Further details can be found in [Schölkopf and Smola, 2002]. For Quantile Regression we drop ϵ and we obtain different constants $C(1-\tau)$ and $C\tau$ for the constraints on α^* and α. We will discuss uniform convergence properties of the empirical risk estimates with respect to various $\psi(\xi)$ in Section 3.2.

3.1.4 Multicategory Classification, Ranking and Ordinal Regression

Many estimation problems cannot be described by assuming that $\mathcal{Y} = \{\pm 1\}$. In this case it is advantageous to go beyond a simple function $f(x)$ depending on x only. Instead, we can encode a larger degree of information by estimating a function $f(x, y)$ directly and subsequently obtaining the estimate $\hat{y}(x) := \operatorname{argmax}_{y\in\mathcal{Y}} f(x, y)$. In other words, we study problems where y is obtained as the solution of an optimization problem over $f(x, y)$ and we wish to find f such that y matches y_i as well as possible.

Unlike binary classification, the loss, however, may be more than just a simple $0-1$ loss. In the following we denote by $\Delta(y, y')$ the loss incurred by estimating y' instead of y. Without loss of generality we require that $\Delta(y, y) = 0$ and that $\Delta(y, y') \geq 0$ for all $y, y' \in \mathcal{Y}$. Key in our reasoning is the following lemma:

Lemma 12 *Let $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ and assume that $\Delta(y, y') \geq 0$ with $\Delta(y, y) = 0$. Moreover let $\xi \geq 0$ such that $f(x, y) - f(x, y') \geq \Delta(y, y') - \xi$ for all $y' \in \mathcal{Y}$. In this case $\xi \geq \Delta(y, \operatorname{argmax}_{y'\in\mathcal{Y}} f(x, y'))$.*

Proof Denote by $y^* := \operatorname{argmax}_{y\in\mathcal{Y}} f(x, y)$. By assumption we have $\xi \geq \Delta(y, y^*) + f(x, y^*) - f(x, y)$. Since $f(x, y^*) \geq f(x, y')$ for all $y' \in \mathcal{Y}$ the inequality holds. ■

The construction of the estimator was suggested in [Taskar et al., 2003, Tsochantaridis et al., 2004], and a special instance of the above lemma is given by Joachims [2005]. While the bound appears quite innocuous, it allows us to describe a much richer class of estimation problems as a convex program.

To deal with the added complexity we assume that f is given by $f(x, y) = \langle\Phi(x, y), w\rangle$. Given the possibly nontrivial connection between x and y the use of $\Phi(x, y)$ cannot be avoided. Corresponding kernel functions are given by $k(x, y, x', y') = \langle\Phi(x, y), \Phi(x', y')\rangle$. We have the following optimization problem [Tsochantaridis et al., 2004]:

$$\underset{w,\xi}{\text{minimize}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i \text{ subject to } \langle w, \phi(x_i, y_i) - \phi(x_i, y)\rangle \geq \Delta(y_i, y) - \xi_i \text{ for all } y \in \mathcal{Y} \tag{88}$$

This is a convex optimization problem which can be solved efficiently if the constraints can be evaluated without high computational cost. One typically employs column-generation methods

[Hettich and Kortanek, 1993, Rätsch, 2001, Bennett et al., 2000, Tsochantaridis et al., 2004, Fletcher, 1989] which identify one violated constraint at a time to find an approximate minimum of the optimization problem.

To describe the flexibility of the framework set out by (88) we give several examples of its application. Further details, in particular when dealing with sequences and other objects of exponential cardinality in $\mathcal{Y}$ are relegated to Section 4.1.5

- Binary classification can be recovered by setting $\Phi(x, y) = y\Phi(x)$, in which case the constraint of (88) reduces to $2y_i \langle \Phi(x_i), w \rangle \geq 1 - \xi_i$. Ignoring constant offsets and a scaling factor of 2, this is exactly the standard SVM optimization problem.
- Multicategory classification problems [Crammer and Singer, 2000, Collins, 2002, Allwein et al., 2000, Rätsch et al., 2003] can be encoded via $\mathcal{Y} = [N]$, where N is the number of classes and $\Delta(y, y') = 1 - \delta_{y,y'}$. In other words, the loss is 1 whenever we predict the wrong class and 0 for correct classification. Corresponding kernels are typically chosen to be $\delta_{y,y'} k(x, x')$.
- We can deal with joint labeling problems by setting $\mathcal{Y} = \{\pm 1\}^n$. In other words, the error measure does not depend on a single observation but on an entire set of labels. Joachims [2005] shows that the so-called F_1 score [van Rijsbergen, 1979] used in document retrieval and the area under the ROC curve [Bamber, 1975, Gribskov and Robinson, 1996] fall into this category of problems. Moreover, Joachims [2005] derives an $O(n^2)$ method for evaluating the inequality constraint over $\mathcal{Y}$.
- Multilabel estimation problems deal with the situation where we want to find the best subset of labels $\mathcal{Y} \subseteq 2^{[N]}$ which correspond to some observation x. The problem is described in [Elisseeff and Weston, 2001], where the authors devise a ranking scheme such that $f(x, i) > f(x, j)$ if label $i \in y$ and $j \notin y$. It is a special case of a general ranking approach described next.

Note that (88) is invariant under translations $\phi(x, y) \leftarrow \phi(x, y) + \phi_0$ where ϕ_0 is constant, as $\phi(x_i, y_i) - \phi(x_i, y)$ remains unchanged. In practice this means that transformations $k(x, y, x', y') \leftarrow k(x, y, x', y') + \langle \phi_0, \phi(x, y) \rangle + \langle \phi_0, \phi(x', y') \rangle + \|\phi_0\|^2$ do not affect the outcome of the estimation process. Since ϕ_0 was arbitrary, we have the following lemma:

Lemma 13 *Let $\mathcal{H}$ be an RKHS on $\mathcal{X} \times \mathcal{Y}$ with kernel k. Moreover let $g \in \mathcal{H}$. Then the function $k(x, y, x', y') + f(x, y) + f(x', y') + \|g\|^2_{\mathcal{H}}$ is a kernel and it yields the same estimates as k.*

We need a slight extension to deal with general ranking problems. Denote by $\mathcal{Y} = \mathrm{Graph}[N]$ the set of all directed graphs on N vertices which do not contain loops of less than three nodes. Here an edge $(i, j) \in y$ indicates that i is preferred to j with respect to the observation x. It is the goal to find some function $f : \mathcal{X} \times [N] \to \mathbb{R}$ which imposes a total order on $[N]$ (for a given x) by virtue of the function values $f(x, i)$ such that the total order and y are in good agreement.

More specifically, Dekel et al. [2003], Crammer [2005], Crammer and Singer [2005] propose a decomposition algorithm $\mathcal{A}$ for the graphs y such that the estimation error is given by the number of subgraphs of y which are in disagreement with the total order imposed by f. As an example, multiclass classification can be viewed as a graph y where the correct label i is at the root of a directed graph and all incorrect labels are its children. Multilabel classification can be seen as a bipartite graph where the correct labels only contain outgoing arcs and the incorrect labels only incoming ones.

This setting leads to a form similar to (88) except for the fact that we now have constraints

over each subgraph $G \in \mathcal{A}(y)$. We solve

$$\underset{w,\xi}{\text{minimize}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} |\mathcal{A}(y_i)|^{-1} \sum_{G\in\mathcal{A}(y_i)} \xi_{iG} \tag{89}$$
$$\text{subject to } \langle w, \Phi(x_i,u) - \Phi(x_i,v)\rangle \geq 1 - \xi_{iG} \text{ and } \xi_{iG} \geq 0 \text{ for all } (u,v) \in G \in \mathcal{A}(y_i).$$

That is, we test for all $(u,v) \in G$ whether the ranking imposed by the subgraph $G \in y_i$ is satisfied.

Finally, ordinal regression problems which perform ranking not over labels y but rather over observations x were studied by Herbrich et al. [2000] and Chapelle and Harchaoui [2005] in the context of ordinal regression and conjoint analysis respectively. In ordinal regression x is preferred to x' if $f(x) > f(x')$ and hence one minimizes an optimization problem akin to (88), with constraint $\langle w, \Phi(x_i) - \Phi(x_j)\rangle \geq 1 - \xi_{ij}$. In conjoint analysis the same operation is carried out for $\Phi(x,u)$, where u is the user under consideration. Similar models were also studied by Basilico and Hofmann [2004]. Further models will be discussed in Section 4, in particular situations where $\mathcal{Y}$ is of exponential size. These models allow one to deal with sequences and more sophisticated structures.

3.1.5 Applications of SVM Algorithms

When SVMs were first presented, they initially met with scepticism in the statistical community. Part of the reason was that as described, SVMs construct their decision rules in potentially very high-dimensional feature spaces associated with kernels. Although there was a fair amount of theoretical work addressing this issue (see Section 3.2 below), it was probably to a larger extent the empirical success of SVMs that paved its way to become a standard method of the statistical toolbox. The first successes of SVMs on practical problem were in handwritten digit recognition, which was the main benchmark task considered in the Adaptive Systems Department at AT&T Bell Labs where SVMs were developed (see e.g. LeCun et al. [1998]). Using methods to incorporate transformation invariances, SVMs were shown to beat the world record on the MNIST benchmark set, at the time the gold standard in the field [DeCoste and Schölkopf, 2002]. There has been a significant number of further computer vision applications of SVMs since then, including tasks such as object recognition [Blanz et al., 1996, Chapelle et al., 1999, Holub et al., 2005, Everingham et al., 2005] and detection [Romdhani et al., 2004]. Nevertheless, it is probably fair to say that two other fields have been more influential in spreading the use of SVMs: bioinformatics and natural language processing. Both of them have generated a spectrum of challenging high-dimensional problems on which SVMs excel, such as microarray processing tasks [Brown et al., 2000] and text categorization [Dumais, 1998]. For further references, see [Schölkopf et al., 2004, Joachims, 2002].

Many successful applications have been implemented using SV classifiers; however, also the other variants of SVMs have led to very good results, including SV regression Müller et al. [1997], SV novelty detection Hayton et al. [2001], SVMs for ranking Herbrich et al. [2000] and problems with interdependent labels McCallum et al. [2005].

3.2 Margins and Uniform Convergence Bounds

So far we motivated the algorithms by means reasons of practicality and the fact that $0-1$ loss functions yield hard-to-control estimators. We now follow up on the analysis by providing uniform convergence bounds for large margin classifiers. We focus on the case of scalar valued functions applied to classification for two reasons: The derivation is well established and it can

be presented in a concise fashion. Secondly, the derivation of corresponding bounds for the vectorial case is by and large still an open problem. Preliminary results exist, such as the bounds by Collins [2002] for the case of perceptrons, Taskar et al. [2003] who derive capacity bounds in terms of covering numbers by an explicit covering construction, and Bartlett and Mendelson [2002], who give Gaussian average bounds for vectorial functions. We believe that the scaling behavior of these bounds in the number of classes $|\mathcal{Y}|$ is currently not optimal, when applied to the problems of type (88).

Our analysis is based on the following ideas: firstly the $0-1$ loss is upper bounded by some function $\psi(yf(x))$ which can be minimized, such as the soft margin function $\max(0, 1-yf(x))$ of the previous section. Secondly we prove that the empirical average of the ψ-loss is concentrated close to its expectation. This will be achieved by means of Rademacher averages. Thirdly we show that under rather general conditions the minimization of the ψ-loss is consistent with the minimization of the expected risk. Finally, we combine these bounds to obtain rates of convergence which only depend on the Rademacher average and the approximation properties of the function class under consideration.

3.2.1 Margins and Empirical Risk

Unless stated otherwise $\mathbf{E}[\cdot]$ denotes the expectation with respect to all random variables of the argument. Subscripts, such as $\mathbf{E}_X[\cdot]$, indicate that the expectation is taken over X. We will omit them wherever obvious. Finally we will refer to $\mathbf{E}_{\text{emp}}[\cdot]$ as the empirical average with respect to an n-sample.

While the sign of $yf(x)$ can be used to assess to accuracy of a binary classifier we saw that for algorithmic reasons one rather optimizes a (smooth function of) $yf(x)$ directly. In the following we assume that the binary loss $\chi(\xi) = \frac{1}{2}(1 - \operatorname{sgn}\xi)$ is majorized by some function $\psi(\xi) \geq \chi(\xi)$, e.g. via the construction of Lemma 12. Consequently $\mathbf{E}[\chi(yf(x))] \leq \mathbf{E}[\psi(yf(x))]$ and likewise $\mathbf{E}_{\text{emp}}[\chi(yf(x))] \leq \mathbf{E}_{\text{emp}}[\psi(yf(x))]$. The hope is (as will be shown in Section 3.2.3) that minimizing the upper bound leads to consistent estimators.

There is a long-standing tradition of minimizing $yf(x)$ rather than the number of misclassifications. $yf(x)$ is known as "margin" (based on the geometrical reasoning) in the context of SVMs [Vapnik and Lerner, 1963, Mangasarian, 1965], as "stability" in the context of Neural Networks [Krauth and Mézard, 1987, Ruján, 1993], and as the "edge" in the context of arcing [Breiman, 1999]. One may show [Makovoz, 1996, Barron, 1993, Herbrich and Williamson, 2002] that functions f in an RKHS achieving a large margin can be approximated by another function f' achieving almost the same empirical error using a much smaller number of kernel functions.

Note that by default, uniform convergence bounds are expressed in terms of minimization of the empirical risk average with respect to a *fixed* function class $\mathcal{F}$, e.g. [Vapnik and Chervonenkis, 1971]. This is very much unlike what is done in practice: in SVM (88) the sum of empirical risk and a regularizer is minimized. However, one may check that minimizing $\mathbf{E}_{\text{emp}}[\psi(yf(x))]$ subject to $\|w\|^2 \leq W$ is equivalent to minimizing $\mathbf{E}_{\text{emp}}[\psi(yf(x))] + \lambda\|w\|^2$ for suitably chosen values of λ. The equivalence is immediate by using Lagrange multipliers. For numerical reasons, however, the second formulation is much more convenient [Tikhonov, 1963, Morozov, 1984], as it acts as a regularizer. Finally, for the design of adaptive estimators, so-called luckiness results exist, which provide risk bounds in a data-dependent fashion [Shawe-Taylor et al., 1998, Herbrich and Williamson, 2002].

3.2.2 Uniform Convergence and Rademacher Averages

The next step is to bound the deviation $\mathbf{E}_{\text{emp}}[\psi(yf(x))] - \mathbf{E}[\psi(yf(x))]$ by means of Rademacher averages. For details see [Bousquet et al., 2004, Mendelson, 2003, Bartlett et al., 2002, Koltchin-

skii, 2001]. Denote by $g : \mathcal{X}^n \to \mathbb{R}$ a function of n variables and let $c > 0$ such that $|g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c$ for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$ and for all $i \in [n]$, then [McDiarmid, 1989]

$$\mathrm{P}\left\{\mathbf{E}\left[g(x_1, \ldots, x_n)\right] - g(x_1, \ldots, x_n) > \epsilon\right\} \leq \exp\left(-2\epsilon^2/nc^2\right). \tag{90}$$

Assume that $f(x) \in [0, B]$ for all $f \in \mathcal{F}$ and let $g(x_1, \ldots, x_n) := \sup_{f \in \mathcal{F}} \left|\mathbf{E}_{\mathrm{emp}}\left[f(x)\right] - \mathbf{E}\left[f(x)\right]\right|$. Then it follows that $c \leq \frac{B}{n}$. Solving (90) for g we obtain that with probability at least $1 - \delta$

$$\sup_{f \in \mathcal{F}} \mathbf{E}\left[f(x)\right] - \mathbf{E}_{\mathrm{emp}}\left[f(x)\right] \leq \mathbf{E}\left[\sup_{f \in \mathcal{F}} \mathbf{E}\left[f(x)\right] - \mathbf{E}_{\mathrm{emp}}\left[f(x)\right]\right] + B\sqrt{-\frac{\log \delta}{2n}} \tag{91}$$

This means that with high probability the largest deviation between the sample average and its expectation is concentrated around its mean and within an $O(n^{-\frac{1}{2}})$ term. The expectation can be bounded by a classical symmetrization argument [Vapnik and Chervonenkis, 1971] as follows:

$$\begin{aligned}
&\mathbf{E}_X\left[\sup_{f \in \mathcal{F}} \mathbf{E}[f(x')] - \mathbf{E}_{\mathrm{emp}}[f(x)]\right] && \leq \mathbf{E}_{X,X'}\left[\sup_{f \in \mathcal{F}} \mathbf{E}_{\mathrm{emp}}[f(x')] - \mathbf{E}_{\mathrm{emp}}[f(x)]\right] \\
=&\mathbf{E}_{X,X',\sigma}\left[\sup_{f \in \mathcal{F}} \mathbf{E}_{\mathrm{emp}}[\sigma f(x')] - \mathbf{E}_{\mathrm{emp}}[\sigma f(x)]\right] && \leq 2\mathbf{E}_{X,\sigma}\left[\sup_{f \in \mathcal{F}} \mathbf{E}_{\mathrm{emp}}[\sigma f(x)]\right] =: 2R_n\left[\mathcal{F}\right].
\end{aligned}$$

The first inequality follows from the convexity of the argument of the expectation, the second equality follows from the fact that x_i and x_i' are drawn iid from the same distribution, hence we may swap terms. Here σ_i are independent ± 1-valued zero-mean Rademacher random variables. $R_n\left[\mathcal{F}\right]$ is referred as the Rademacher average [Mendelson, 2001, Bartlett and Mendelson, 2002, Koltchinskii, 2001] of $\mathcal{F}$ wrt. sample size n.

For linear function classes $R_n\left[\mathcal{F}\right]$ takes on a particularly nice form. We begin with $\mathcal{F} := \{f | f(x) = \langle x, w\rangle \text{ and } \|w\| \leq 1\}$. It follows that $\sup_{\|w\| \leq 1} \sum_{i=1}^n \sigma_i \langle w, x_i\rangle = \|\sum_{i=1}^n \sigma_i x_i\|$. Hence

$$nR_n\left[\mathcal{F}\right] = \mathbf{E}_{X,\sigma}\Big\|\sum_{i=1}^n \sigma_i x_i\Big\| \leq \mathbf{E}_X\left[\mathbf{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i x_i\right\|^2\right]\right]^{\frac{1}{2}} = \mathbf{E}_X\left[\sum_{i=1}^n \|x_i\|^2\right]^{\frac{1}{2}} \leq \sqrt{n\mathbf{E}\left[\|x\|^2\right]}. \tag{92}$$

Here the first inequality is a consequence of Jensen's inequality, the second equality follows from the fact that σ_i are iid zero-mean random variables, and the last step again is a result of Jensen's inequality. Corresponding tight lower bounds by a factor of $1/\sqrt{2}$ exist and they are a result of the Khintchine-Kahane inequality [Kahane, 1968].

Note that (92) allows us to bound $R_n\left[\mathcal{F}\right] \leq n^{-\frac{1}{2}} r$ where r is the average length of the sample. An extension to kernel functions is straightforward: by design of the inner product we have $r = \sqrt{\mathbf{E}_x\left[k(x, x)\right]}$. Note that this bound is *independent* of the dimensionality of the data but rather only depends on the expected length of the data. Moreover r is the trace of the integral operator with kernel $k(x, x')$ and probability measure on $\mathcal{X}$.

Since we are computing $\mathbf{E}_{\mathrm{emp}}\left[\psi(yf(x))\right]$ we are interested in the Rademacher complexity of $\psi \circ \mathcal{F}$. Bartlett and Mendelson [2002] show that $R_n\left[\psi \circ \mathcal{F}\right] \leq LR_n\left[\mathcal{F}\right]$ for any Lipschitz continuous function ψ with Lipschitz constant L and with $\psi(0) = 0$. Secondly, for $\{yb \text{ where } |b| \leq B\}$ the Rademacher average can be bounded by $B\sqrt{2\log 2/n}$, as follows from [Bousquet et al., 2004, eq. (4)]. This takes care of the offset b. For sums of function classes $\mathcal{F}$ and $\mathcal{G}$ we have $R_n\left[\mathcal{F} + \mathcal{G}\right] \leq R_n\left[\mathcal{F}\right] + R_n\left[\mathcal{G}\right]$. This means that for linear functions with $\|w\| \leq W$, $|b| \leq B$, and ψ Lipschitz continuous with constant L we have $R_n \leq \frac{L}{\sqrt{n}}(Wr + B\sqrt{2\log 2})$.

3.2.3 Upper Bounds and Convex Functions

We briefly discuss consistency of minimization of the surrogate loss function $\psi : \mathbb{R} \to [0, \infty)$ about which assume that it is convex and that $\psi \geq \chi$ [Jordan et al., 2003, Zhang, 2004]. Examples of such functions are the soft-margin loss $\max(0, 1 - \gamma\xi)$, which we discussed in Section 3.1, the Boosting Loss $e^{-\xi}$, which is commonly used in AdaBoost [Schapire et al., 1998, Rätsch et al., 2001], and the logistic loss $\log_2\left(1 + e^{-\xi}\right)$ (see Section 4).

Denote by f_χ^* the minimizer of the expected risk and let f_ψ^* be the minimizer of $\mathbf{E}\left[\psi(yf(x)\right]$ with respect to f. Then under rather general conditions on ψ [Zhang, 2004] for all f the following inequality holds:

$$\mathbf{E}\left[\chi(yf(x))\right] - \mathbf{E}\left[\chi(yf_\chi^*(x))\right] \leq c\left(\mathbf{E}\left[\psi(yf(x))\right] - \mathbf{E}\left[\psi(yf_\psi^*(x))\right]\right)^s. \tag{93}$$

In particular we have $c = 4$ and $s = 1$ for soft margin loss, whereas for Boosting and logistic regression $c = \sqrt{8}$ and $s = \frac{1}{2}$. Note that (93) implies that the minimizer of the ψ loss is consistent, i.e. $\mathbf{E}\left[\chi(yf_\psi(x))\right] = \mathbf{E}\left[\chi(yf_\chi(x))\right]$.

3.2.4 Rates of Convergence

We now have all tools at our disposition to obtain rates of convergence to the minimizer of the expected risk which depend only on the complexity of the function class and its approximation properties in terms of the ψ-loss. Denote by $f_{\psi,\mathcal{F}}^*$ the minimizer of $\mathbf{E}\left[\psi(yf(x))\right]$ restricted to $\mathcal{F}$, let $f_{\psi,\mathcal{F}}^n$ be the minimizer of the empirical ψ-risk, and let $\delta(\mathcal{F}, \psi) := \mathbf{E}\left[yf_{\psi,\mathcal{F}}^*(x)\right] - \mathbf{E}\left[yf_\psi^*(x)\right]$ be the approximation error due to the restriction of f to $\mathcal{F}$. Then a simple telescope sum yields

$$\begin{aligned}
\mathbf{E}\left[\chi(yf_{\psi,\mathcal{F}}^n)\right] \leq & \mathbf{E}\left[\chi(yf_\chi^*)\right] + 4\left[\mathbf{E}\left[\psi(yf_{\psi,\mathcal{F}}^n)\right] - \mathbf{E}_{\text{emp}}\left[\psi(yf_{\psi,\mathcal{F}}^n)\right]\right] \\
& + 4\left[\mathbf{E}_{\text{emp}}\left[\psi(yf_{\psi,\mathcal{F}}^*)\right] - \mathbf{E}\left[\psi(yf_{\psi,\mathcal{F}}^*)\right]\right] + \delta(\mathcal{F}, \psi) \\
\leq & \mathbf{E}\left[\chi(yf_\chi^*)\right] + \delta(\mathcal{F}, \psi) + 4\frac{RW\gamma}{\sqrt{n}}\left[\sqrt{-2\log\delta} + r/R + \sqrt{8\log 2}\right]
\end{aligned} \tag{94}$$

Here γ is the effective margin of the soft-margin loss $\max(0, 1 - \gamma yf(x))$, W is an upper bound on $\|w\|$, $R \geq \|x\|$, r is the average radius, as defined in the previous section, and we assumed that b is bounded by the largest value of $\langle w, x\rangle$. A similar reasoning for logistic and exponential loss is given in [Bousquet et al., 2004].

Note that we get an $O(1/\sqrt{n})$ rate of convergence regardless of the dimensionality of x. Moreover note that the rate is dominated by $RW\gamma$, that is, the classical radius-margin bound [Vapnik, 1995]. Here R is the radius of an enclosing sphere for the data and $1/(W\gamma)$ is an upper bound on the radius of the data — the soft-margin loss becomes active only for $yf(x) \leq \gamma$.

3.2.5 Localization and Noise Conditions

In many cases it is possible to obtain better rates of convergence than $O(1/\sqrt{n})$ by exploiting information about the magnitude of the error of misclassification and about the variance of f on $\mathcal{X}$. Such bounds use Bernstein-type inequalities and they lead to localized Rademacher averages [Bartlett et al., 2002, Mendelson, 2003, Bousquet et al., 2004].

Basically the slow $O(1/\sqrt{n})$ rates arise whenever the region around the Bayes optimal decision boundary is large. In this case, determining this region produces the slow rate, whereas the well-determined region could be estimated at an $O(1/n)$ rate.

Tsybakov's noise condition [Tsybakov, 2003] requires that there exist $\beta, \gamma \geq 0$ such that

$$\mathrm{P}\left\{\left|\mathrm{P}\left\{y=1|x\right\}-\frac{1}{2}\right| \leq t\right\} \leq \beta t^{\gamma} \text{ for all } t \geq 0. \tag{95}$$

Note that for $\gamma = \infty$ the condition implies that there exists some s such that $\left|\mathrm{P}\left\{y=1|x\right\}-\frac{1}{2}\right| \geq s > 0$ almost surely. This is also known as Massart's noise condition.

The key benefit of (95) is that it implies a relationship between variance and expected value of classification loss. More specifically for $\alpha = \frac{\gamma}{1+\gamma}$ and $g : \mathcal{X} \to \mathcal{Y}$ we have

$$\mathbf{E}\left[\left[\{g(x) \neq y\} - \{g^*(x) \neq y\}\right]^2\right] \leq c\left[\mathbf{E}\left[\{g(x) \neq y\} - \{g^*(x) \neq y\}\right]\right]^{\alpha}. \tag{96}$$

Here $g^*(x) := \operatorname{argmax}_y \mathrm{P}(y|x)$ denotes the Bayes optimal classifier. This is sufficient to obtain faster rates for finite sets of classifiers. For more complex function classes localization is used. See, e.g. [Bousquet et al., 2004, Bartlett et al., 2002] for more details.

3.3 Distribution of Data in Feature Space

This section discusses various methods of data analysis by modeling the distribution of data in feature space. That is, we study the behavior of $\Phi(x)$ by means of rather simple linear methods, which has implications for nonlinear methods on the space of the data itself, that is $\mathcal{X}$. In particular, we will discuss the extension of PCA to Hilbert spaces, which allows for image denoising, clustering, and nonlinear dimensionality reduction, the study of covariance operators for the measure of independence, and the modeling of complex dependencies between sets of random variables via kernel dependency estimation and canonical correlation analysis.

3.3.1 Kernel Principal Component Analysis

Principal Component Analysis (PCA) is a powerful technique for extracting structure from possibly high-dimensional data sets. It is readily performed by solving an eigenvalue problem, or by using iterative algorithms which estimate principal components. For reviews of the existing literature, see [Jolliffe, 1986, Diamantaras and Kung, 1996]; some of the classical papers are [Pearson, 1901, Hotelling, 1933, Karhunen, 1946]. PCA is an orthogonal transformation of the coordinate system in which we describe our data. The new coordinate system is obtained by projection onto the so-called principal axes of the data. A small number of principal components is often sufficient to account for most of the structure in the data, e.g. for the purpose of principal component regression [Draper and Smith, 1981].

The basic idea is strikingly simple: denote by $X = \{x_1, \ldots, x_n\}$ an n-sample drawn from $\mathrm{P}(x)$. Then the covariance operator C is given by $C = \mathbf{E}\left[(x - \mathbf{E}\left[x\right])(x - \mathbf{E}\left[x\right])^{\top}\right]$. PCA aims at estimating leading eigenvectors of C via the empirical estimate $C_{\mathrm{emp}} = \mathbf{E}_{\mathrm{emp}}\left[(x - \mathbf{E}_{\mathrm{emp}}\left[x\right])(x - \mathbf{E}_{\mathrm{emp}}\left[x\right])^{\top}\right]$. If $\mathcal{X}$ is d-dimensional, then the eigenvectors can be computed in $O(d^3)$ time [Press et al., 1994].

The problem can also be posed in feature space [Schölkopf et al., 1998] by replacing x by $\Phi(x)$. In this case, however, it is impossible to compute the eigenvectors directly. However, note that the image of C_{emp} lies in the span of $\{\Phi(x_1), \ldots, \Phi(x_n)\}$. Hence it is sufficient to diagonalize C_{emp} in that subspace. In other words, we replace the *outer* product C_{emp} by an *inner* product matrix, leaving the eigenvalues unchanged, which can be computed efficiently. Using $w = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$ it follows that α needs to satisfy $PKP\alpha = \lambda\alpha$, where P is the projection operator with $P_{ij} = \delta_{ij} - n^{-2}$ and K is the kernel matrix on X.

Note that the problem can also be recovered as one of maximizing some Contrast$[f, X]$ subject to $f \in \mathcal{F}$. This means that the projections onto the leading eigenvectors correspond to the most

reliable features in the sense of Section 3.2.2. Such an optimization problem also allows us to unify various feature extraction methods as follows:

- For Contrast$[f, X] = \mathrm{Var}_{\mathrm{emp}}[f, X]$ and $\mathcal{F} = \{\langle w, x\rangle \text{ subject to } \|w\| \leq 1\}$ we recover PCA.
- Changing $\mathcal{F}$ to $\mathcal{F} = \{\langle w, \Phi(x)\rangle \text{ subject to } \|w\| \leq 1\}$ we recover Kernel-PCA.
- For Contrast$[f, X]$ = Curtosis$[f, X]$ and $\mathcal{F} = \{\langle w, x\rangle \text{ subject to } \|w\| \leq 1\}$ we have Projection Pursuit [Huber, 1985, Friedman and Tukey, 1974]. Other contrasts lead to further variants, e.g. the Epanechikov kernel, entropic contrasts, etc. [Cook et al., 1993, Friedman, 1987, Jones and Sibson, 1987].
- If $\mathcal{F}$ is a convex combination of basis functions and the contrast function is convex in w one obtains computationally efficient algorithms, as the solution of the optimization problem can be found at one of the vertices [Rockafellar, 1970, Schölkopf and Smola, 2002].

Subsequent projections are obtained, e.g. by seeking directions orthogonal to f or other computationally attractive variants thereof.

Kernel-PCA has been applied to numerous problems, from dimensionality reduction for estimation [Blanz et al., 1996] by capturing only the leading principal components, invariant feature extraction [Mika et al., 2003] by simultaneous diagonalization of the data and a noise covariance matrix, to image denoising [Mika et al., 1999] and super-resolution Kim et al. [2005]. The basic idea in the latter case is to obtain a set of principal directions in feature space $w_1, \ldots, w_l$, obtained from noise-free data, and to project the image $\Phi(x)$ of a noisy observation x onto the space spanned by $w_1, \ldots, w_l$. This yields a "denoised" solution $\tilde{\Phi}(x)$ in feature space. Finally, to obtain the pre-image of this denoised solution one minimizes $\left\|\Phi(x') - \tilde{\Phi}(x)\right\|$. The fact that projections onto the leading principal components turn out to be good starting points for pre-image iterations is further exploited in kernel dependency estimation (Section 3.3.4).

3.3.2 Canonical Correlation Analysis

Given two samples X, Y canonical correlation analysis [Hotelling, 1936] aims at finding directions of projection u, v such that the correlation coefficient between X and Y is maximized. That is, (u, v) are given by

$$\operatorname*{argmax}_{u,v} \mathrm{Var}_{\mathrm{emp}}\left[\langle u, x\rangle\right]^{-1} \mathrm{Var}_{\mathrm{emp}}\left[\langle v, y\rangle\right]^{-1} \mathbf{E}_{\mathrm{emp}}\left[\langle u, x - \mathbf{E}_{\mathrm{emp}}[x]\rangle \langle v, y - \mathbf{E}_{\mathrm{emp}}[y]\rangle\right]. \tag{97}$$

One can show that this problem can be solved by finding the eigensystem of $C_x^{-\frac{1}{2}} C_{xy} C_y^{-\frac{1}{2}}$, where C_x, C_y are the covariance matrices of X and Y and C_{xy} is the covariance matrix between X and Y, respectively. Extensions to the case of several random variables are discussed in [Kettenring, 1971].

Melzer et al. [2001] and Bach and Jordan [2002] extended CCA to kernels by means of replacing linear projections $\langle u, x\rangle$ by projections in feature space $\langle u, \Phi(x)\rangle$. More specifically, Bach and Jordan [2002] used the so-derived contrast to obtain a measure of independence and they applied it to Independent Component Analysis with great success. Unfortunately, the formulation only works because of an additional regularization term (otherwise the optimization problem becomes distribution independent). The reason for this is that the function class $\langle \Phi(x), u\rangle$ can be quite rich and consequently it is generally possible to find some directions u, v for which the projections are perfectly correlated. Such disastrous behavior is prevented in [Bach and Jordan, 2002] by means of an additional regularization term.

The problem is that bounding the variance of each projections is not a good way to control the capacity of the function class. Instead, we may modify the correlation coefficient by normalizing

by the norms of u and v. Hence we maximize

$$\|u\|^{-1} \|v\|^{-1} \mathbf{E}_{\text{emp}} \left[\langle u, x - \mathbf{E}_{\text{emp}}[x] \rangle \langle v, y - \mathbf{E}_{\text{emp}}[y] \rangle \right], \tag{98}$$

which can be solved by diagonalization of the covariance matrix C_{xy}. One may check that in feature space, this amounts to finding the eigenvectors of $(PK_xP)^{\frac{1}{2}}(PK_yP)^{\frac{1}{2}}$ [Gretton et al., 2005b].

3.3.3 Measures of Independence

Note that the correlation coefficient allows one to define measures of independence. As shown by Rényi [1959], independence between random variables is equivalent to the condition that the covariance $\text{Cov}[f(x), g(y)] = 0$ for all C^1 functions f, g bounded by L_∞ norm 1 on $\mathcal{X}$ and $\mathcal{Y}$. In [Das and Sen, 1994, Dauxois and Nkiet, 1998, Bach and Jordan, 2002, Gretton et al., 2005b,a] a constrained empirical estimate of the above criterion is used. That is, one studies

$$\Lambda(X, Y, \mathcal{F}, \mathcal{G}) := \sup_{f,g} \text{Cov}_{\text{emp}}[f(x), g(y)] \text{ subject to } f \in \mathcal{F} \text{ and } g \in \mathcal{G}. \tag{99}$$

This statistic is often extended to use the entire series $\Lambda_1, \ldots, \Lambda_d$ of maximal correlations where each of the function pairs (f_i, g_i) are orthogonal to the previous set of terms. More specifically Dauxois and Nkiet [1998] restrict $\mathcal{F}, \mathcal{G}$ to finite-dimensional linear function classes subject to their L_2 norm bounded by 1, Bach and Jordan [2002] use functions in the RKHS for which some sum of the ℓ_2^n and the RKHS norm on the sample is bounded. Martin [2000], Cock and Moor [2002], Vishwanathan and Smola [2004b] show that degrees of maximum correlation can also be viewed as an inner product between subspaces spanned by he observations.

Finally, Gretton et al. [2005b,a] use functions with bounded RKHS norm only. The authors show that for universal kernels on $\mathcal{X}$ and $\mathcal{Y}$ the function $\Lambda(X, Y, \mathcal{F}, \mathcal{G}) = 0$ if and only if x and y are independent. Moreover, Gretton et al. [2005a] show that $\text{tr}\, PK_xPK_yP$ has the same theoretical properties and it can be computed much more easily. In fact, by using incomplete factorizations it is possible to compute an approximation to the above quantity in $O(n)$ time. Here K_x and K_y are the kernel matrices on X and Y respectively.

The above criteria can be used to derive algorithms for Independent Component Analysis, which are some of the best methods available [Bach and Jordan, 2002, Gretton et al., 2005a]. In this case one seeks rotations of the data $S = VX$ where $V \in \text{SO}(d)$ such that S is coordinate-wise independent, i.e. $\text{P}(S) = \prod_{i=1}^d \text{P}(s_i)$.

3.3.4 Kernel Dependency Estimation

A large part of the previous discussion revolved around estimating dependencies between sample $\mathcal{X}$ and $\mathcal{Y}$ for rather structured spaces $\mathcal{Y}$, in particular (88). In general, however, such dependencies can be hard to compute. Weston et al. [2003] proposed an algorithm which allows one to extend standard regularized LMS regression models, as described in Section 3.1.3, to cases where $\mathcal{Y}$ has complex structure.

It works by recasting the estimation problem as a *linear* estimation problem for the map $f : \Phi(x) \to \Phi(y)$ and then as a nonlinear pre-image estimation problem for finding $\hat{y} := \text{argmin} \|f(x)\Phi(y)\|$ as the point in $\mathcal{Y}$ closest to $f(x)$.

Since $\Phi(y)$ is a possibly infinite dimensional space, which may lead to problems of capacity of the estimator, Weston et al. [2003] restrict f to the leading principal components of $\Phi(y)$ and subsequently they perform regularized LMS regression onto each of the projections $\langle v_j, \Phi(y) \rangle$ separately, yielding functions f_j. This yields an estimate of f via $f(x) = \sum_j v_j f_j(x)$. The authors report good performance of the estimator on strings and other structured data.

4 Statistical Models and RKHS

As we have argued so far, the Reproducing Kernel Hilbert Space approach offers many significant advantages in machine learning: (i) powerful and flexible hypothesis classes or feature spaces can be defined, (ii) many results and algorithms for linear models in Euclidean spaces can be generalized to RKHS, (iii) learning theory assures that effective learning in RKHS is possible, for instance, by means of regularization.

In this chapter, we will show how kernel methods can be utilzed in the context of *statistical models*. There are several reasons to pursue such an avenue. First of all, in conditional modeling, it is often insufficent to compute a prediction without assessing confidence and reliability. Second, when dealing with multiple or structured responses, it is important to model *dependencies between responses* in addition to the dependence on a set of covariates. Third, incomplete data, be it due to missing variables, incomplete training targets, or a model structure involving latent variables, needs to be dealt with in a principled manner. All of these issues can be addressed by using the RKHS approach to define statistical models and by combining kernels with statistical approaches such as exponential models, generalized linear models, and Markov networks.

4.1 Exponential RKHS Models

4.1.1 Exponential Models

Exponential models or *exponential families* are among the most important class of parametric models studied in statistics. Given a canonical vector of statistics Φ and a σ-finite measure ν over the sample space $\mathcal{X}$, an exponential model can be defined via its probability density with respect to ν (cf. **?**),

$$p(x;\theta) = \exp\left[\langle\theta, \Phi(x)\rangle - g(\theta)\right], \quad \text{where} \quad g(\theta) := \ln \int_{\mathcal{X}} e^{\langle\theta,\Phi(x)\rangle} \nu(dx)\,. \tag{100}$$

The m-dimensional parameter vector $\theta \in \Theta$ with $\Theta := \{\theta \in \mathbb{R}^m : g(\theta) < \infty\}$ are also called *canonical parameter*. In general there are multiple exponential representations of the same model via canonical parameters that are affinely related to one another (Murray and Rice [1993]). A representation with minimal dimension m is called a minimal representation in which case m is the *order* of the exponential model. One of the most important properties of exponential families is that they have sufficient statistics of fixed dimensionality (**??**), i.e. the joint density for i.i.d. random variables $X_1, X_2, \ldots, X_n$ is also exponential, the corresponding canonical statistics simply being $\sum_{i=1}^n \Phi(X_i)$. It is well-known that much of the structure of exponential models can be derived from the log partition function $g(\theta)$, in particular (cf. Lauritzen [1996])

$$\nabla_\theta g(\theta) = \mu(\theta) := \mathbf{E}_\theta\left[\Phi(X)\right], \qquad \partial^2_\theta g(\theta) = \mathbf{V}_\theta\left[\Phi(X)\right], \tag{101}$$

where μ is known as the mean-value map. Being a covariance matrix, the Hessian of g is positive semi-definite and consequently g is convex.

Maximum likelihood estimation (MLE) in exponential families leads to a particularly elegant form for the MLE equations: the expected and the observed canonical statistics agree at the MLE $\hat{\theta}$. This means, given an i.i.d. sample $\mathcal{S} = (x_i)_{i\in[n]}$, $\mathbf{E}_{\hat{\theta}}\left[\Phi(X)\right] = \mu(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \Phi(x_i) =: \mathbf{E}_{\mathcal{S}}\left[\Phi(X)\right]$.

4.1.2 Exponential RKHS Models

One can extend the parameteric exponential model in (100) by defining a statistical model via an RKHS $\mathcal{H}$ with generating kernel k. Linear function $\langle\theta, \Phi(\cdot)\rangle$ over $\mathcal{X}$ are replaced with functions

$f \in \mathcal{H}$, which yields an exponential RKHS model

$$p(x; f) = \exp[f(x) - g(f)], \quad f \in \mathcal{H} := \left\{ f : f(\cdot) = \sum_{x \in \mathcal{S}} \alpha_x k(\cdot, x),\ \mathcal{S} \subseteq \mathcal{X},\ |\mathcal{S}| < \infty \right\}. \tag{102}$$

A justification for using exponential RKHS families with rich canonical statistics as a generic way to define non-parametric models stems from the fact that if the chosen kernel k is powerful enough, the associated exponential families become universal density estimators. This can be made precise using the concept of universal kernels (Steinwart [2002a], cf. Section 2).

Proposition 14 (Dense Densities) *Let $\mathcal{X}$ be a measurable set with a fixed σ-finite measure ν and denote by $\mathcal{P}$ a family of densities on $\mathcal{X}$ with respect to ν such that $p \in \mathcal{P}$ is bounded from above and continuous. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a universal kernel for $\mathcal{H}$. Then the exponential RKHS family of densities generated by k according to Eq.* (102) *are dense in $\mathcal{P}$ in the L_∞ sense.*
Proof *Let $D := \|p\|_\infty$ and chose $\eta = \eta(\epsilon, D) > 0$ appropriately (see below). By assumption there exists $f \in \mathcal{H}$ such that $\|f - \ln p\|_\infty < \eta$. Exploiting the strict convexity of the exp-function one gets that $e^{-\eta} p < e^f < e^\eta p$, which directly implies $e^{-\eta} < e^{g(f)} = \int_{\mathcal{X}} e^{f(x)} \nu(dx) < e^\eta$. Moreover $\|e^f - p\|_\infty < D(e^\eta - 1)$, and hence*

$$\|p(f) - p\|_\infty \leq \|e^f - p\|_\infty + \|e^{f - g(f)} - e^f\|_\infty < D(e^\eta - 1)(1 + e^\eta) < 2De^\eta(e^\eta - 1)\,,$$

where we utilized the upper bound

$$\|e^{f-g(f)} - e^f\|_\infty \leq |e^{-g(f)} - 1| \cdot \|e^f\|_\infty < (e^\eta - 1)\|e^f\|_\infty < De^\eta(e^\eta - 1).$$

Equaling with ϵ and solving for η results in $\eta(\epsilon, D) = \ln\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{2\epsilon}{D}}\right) > 0$. ■

4.1.3 Conditional Exponential Models

For the rest of the paper, we will focus on the case of predictive or conditional modeling with a – potentially compound or structured – response variable Y and predictor variables X. Taking up the concept of joint kernels introduced in the previous section, we will investigate conditional models that are defined by functions $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ from some RKHS $\mathcal{H}$ over $\mathcal{X} \times \mathcal{Y}$ with kernel k as follows

$$p(y|x; f) = \exp\left[f(x, y) - g(x, f)\right], \quad \text{where} \quad g(x, f) := \ln \int_{\mathcal{Y}} e^{f(x,y)} \mu(dy)\,. \tag{103}$$

Notice that in the finite-dimensional case we have a feature map $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$ from which parametric models are obtained via $\mathcal{H} := \{f : \exists w, f(x, y) = f(x, y; w) := \langle w, \Psi(x, y)\rangle\}$ and each f can be identified with its parameter w. Let us discuss some concrete examples to illustrate the rather general model equation (103).

- Let Y be univariate and define $\Psi(x, y) = y\Phi(x)$. Then simply $f(x, y; w) = \langle w, \Psi(x, y)\rangle = y\tilde{f}(x; w)$, $\tilde{f}(x; w) := \langle w, \Phi(x)\rangle$ and the model equation in (103) reduces to

  $$p(y|x; w) = \exp\left[y \langle w, \Phi(x)\rangle - g(x, w)\right]. \tag{104}$$

  This is a *generalized linear model* (GLM) (Nelder and Wedderburn [1972], McCullagh and Nelder [1983]) with a canonical link, i.e. the canonical parameters depend linearly on the covariates. For different response scales we get several well-known models such as, for instance, logistic regression where $y \in \{-1, 1\}$.

- In the non-parameteric extension of generalized linear models following Green and Yandell [1985], O'Sullivan et al. [1986] the parametric assumption on the linear predictor $\tilde{f}(x;w) = \langle w, \Phi(x)\rangle$ in GLMs is relaxed by requiring that $\tilde{f}$ comes from some sufficiently smooth class of functions, namely a RKHS defined over $\mathcal{X}$. In combination with a parametric part, this can also be uses to define semi-parametric models. Popular choices of reproducing kernels include the ANOVA kernel investigated in Wahba et al. [1995]. This is a special case of defining joint kernels from an existing kernel k over inputs via $k((x,y),(x',y')) := yy'k(x,x')$.
- Joint kernels provide a powerful framework for prediction problems with structured outputs. An illuminating example is statistical natural language parsing with lexicalized probabilistic context free grammars (cf. Magerman [1996], Charniak [1997]). Here x will be an English sentence and y a parse tree for x. The productions of the grammar are known, but the conditional probability $p(y|x)$ needs to be estimated based on training data of parsed/annotated sentences. In the simplest case, the extracted statistics Ψ may encode the frequencies of the use of different productions in a sentence with known parse tree. More sophisticated feature encodings are discussed in Taskar et al. [2004], Zettlemoyer and Collins [2005]. The latter conditional modeling approaches provide alternatives to state-of-the art approaches for estimating joint models $p(x,y)$ with maximum likelihood or maximum entropy (Charniak [2000]).

4.1.4 Risk Functions for Model Fitting

There are different inference principles to determine the optimal function $f \in \mathcal{H}$ for the conditional exponential model in (103). One standard approach to parametric model fitting is to maximize the conditional log-likelihood – or equivalently – minimize a logarithmic loss, a strategy pursued in the Conditional Random Field (CRF) approach of Lafferty et al. [2001]. Here we consider the more general case of minimizing a functional that includes (a monotone function of) the Hilbert space norm $\|f\|_{\mathcal{H}}$ as a stabilizer (cf. Wahba [1990]). This reduces to penalized log-likelihood estimation in the finite dimensional case,

$$C^{\text{ll}}(f;\mathcal{S}) := -\frac{1}{n}\sum_{i=1}^{n} \ln p(y_i|x_i;f), \quad \hat{f}^{\text{ll}} := \operatorname*{argmin}_{f\in\mathcal{H}} \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 + C^{\text{ll}}(f;\mathcal{S})\,. \tag{105}$$

- For the parametric case, Lafferty et al. [2001] have employed variants of improved iterative scaling (Della Pietra et al. [1997], Darroch and Ratcliff [1972]) to optimize Eq.(105) whereas Sha and Pereira [2003] have investigated preconditioned conjugate gradient descent and limited memory quasi-Newton methods (Nocedal and Wright [1999]). In addition, the generalized perceptron algorithm of Collins [2002] has been suggested as an efficient approximation scheme.
- In order to optimize Eq.(105) one usually needs to compute expectations of the canoncial statistics $\mathbf{E}_f[\Psi(Y,x)]$ at sample points $x = x_i$. This requires the availability of efficient inference algorithms, a topic that will be discussed more systematically in Section 4.2.

As we have seen in the case of classification and regression, likelihood-based criteria are by no means the only justifiable choice and large margin methods offer an interesting alternative. To that extend, we will present a general formulation of large margin methods for response variables Y over finite sample spaces $\mathcal{Y}$ that is based on Altun et al. [2003] and Taskar et al. [2003]. Define

$$r(x,y;f) := f(x,y) - \max_{y'\neq y} f(x,y') = \min_{y'\neq y} \log\frac{p(y|x;f)}{p(y'|x;f)} \quad \text{and} \quad r(\mathcal{S};f) := \min_{i=1}^{n} r(x_i,y_i;f)\,. \tag{106}$$

Here $r(\mathcal{S}; f)$ generalizes the notion of separation margin used in SVMs. Since the log-odds ratio is sensitive to rescaling of f, i.e. $r(x, y; \beta f) = \beta r(x, y; f)$, we need to constrain $\|f\|_{\mathcal{H}}$ to make the problem well-defined. We thus replace f by $\phi^{-1} f$ for some fixed dispersion parameter $\phi > 0$ and define the maximum margin problem $\hat{f}^{\text{mm}}(\mathcal{S}) := \phi^{-1} \operatorname{argmax}_{\|f\|_{\mathcal{H}}=1} r(\mathcal{S}; f/\phi)$. Using the same line of arguments as was used in Section 3, this can be re-formulated as a constrained optimization problem

$$\hat{f}^{\text{mm}}(\mathcal{S}) \propto \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{2}\|f\|_{\mathcal{H}}^2, \quad \text{s.t. } r(x_i, y_i; f) \geq 1, \forall i \in [n]\,, \tag{107}$$

provided the latter is feasible, i.e. if there exists $f \in \mathcal{H}$ such that $r(\mathcal{S}; f) > 0$. To make the connection to SVMs consider the case of binary classification case with $\Psi(x, y) = y\Phi(x)$, $f(x, y; w) = \langle w, y\Phi(x)\rangle$ where $r(x, y; f) = \langle w, y\Phi(x)\rangle - \langle w, -y\Phi(x)\rangle = 2y \langle w, \Phi(x)\rangle = 2\rho(x, y; w)$. The latter is twice the standard margin for binary classification in SVMs.

A soft margin version can be defined based on the Hinge loss as follows

$$C^{\text{hl}}(f; \mathcal{S}) := \frac{1}{n} \sum_{i=1}^{n} \min\{1 - r(x_i, y_i; f), 0\}\,, \quad \hat{f}^{\text{sm}}(\mathcal{S}) := \operatorname*{argmin}_{f \in \mathcal{H}} \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 + C^{\text{hl}}(f, \mathcal{S})\,. \tag{108}$$

- An equivalent formulation using slack variables ξ_i as discussed in Section 3 can be obtained by introducing soft-margin constraints $r(x_i, y_i; f) \geq 1 - \xi_i$, $\xi_i \geq 0$ and by defining $C^{\text{hl}} = \frac{1}{n}\xi_i$. Each non-linear constraint can be further expanded into $|\mathcal{Y}|$ linear constraints $f(x_i, y_i) - f(x_i, y) \geq 1 - \xi_i$ for all $y \neq y_i$.
- Prediction problems with structured outputs often involve task-specific loss function $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ discussed in Section 3.1.4. As suggested in Taskar et al. [2003] cost sensitive large margin methods can be obtained by defining re-scaled margin constraints $f(x_i, y_i) - f(x_i, y) \geq \triangle(y_i, y) - \xi_i$.
- Another sensible option in the parametric case is to minimize an exponential risk function of the following type

$$\hat{f}^{\text{exp}}(\mathcal{S}) := \operatorname*{argmin}_{w} \frac{1}{n} \sum_{i=1}^{n} \sum_{y \neq y_i} \exp\left[f(x_i, y_i; w) - f(x_i, y; w)\right]\,. \tag{109}$$

  This is related to the exponential loss used in the AdaBoost method of Freund and Schapire [1996] (cf. Friedman et al. [1998], Collins et al. [2000]). Since we are mainly interested in kernel-based methods here, we refrain from further elaborating on this connection.

4.1.5 Generalized Representer Theorem and Dual Soft-Margin Formulation

It is crucial to understand how the representer theorem applies in the setting of arbitrary discrete output spaces, since a finite representation for the optimal $\hat{f} \in \{\hat{f}^{\text{ll}}, \hat{f}^{\text{sm}}\}$ is the basis for constructive model fitting. Notice that the regularized log-loss as well as the soft margin functional introduced above depend not only on the values of f on the sample $\mathcal{S}$, but rather on the evaluation of f on the augmented sample $\tilde{\mathcal{S}} := \{(x_i, y) : i \in [n],\, y \in \mathcal{Y}\}$. This is the case, because for each x_i, output values $y \neq y_i$ not observed with x_i show up in the log-partition function $g(x_i, f)$ in (103) as well as in the log-odds ratios in (106). This adds an additional complication compared to binary classification.

Corollary 15 *Denote by $\mathcal{H}$ an RKHS on $\mathcal{X} \times \mathcal{Y}$ with kernel k and let $\mathcal{S} = ((x_i, y_i))_{i \in [n]}$. Furthermore let $C(f; \mathcal{S})$ be a functional depending on f only via its values on the augmented sample*

$\tilde{\mathcal{S}}$. *Let Ω be a strictly monotonically increasing function. Then the solution of the optimization problem $\hat{f}(\mathcal{S}) := \operatorname{argmin}_{f\in\mathcal{H}} C(f;\tilde{\mathcal{S}}) + \Omega(\|f\|_{\mathcal{H}})$ can be written as:*

$$\hat{f}(\cdot) = \sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} \beta_{iy} k(\cdot,(x_i,y)) \tag{110}$$

Proof *Follows directly from Theorem 11.* ■

Let us focus on the soft margin maximizer $\hat{f}^{\text{sm}}$. Instead of solving (108) directly, we first derive the dual program, following essentially the derivation in Section 3.

Proposition 16 (Tsochantaridis et al. [2004]) *The minimizer $\hat{f}^{\text{sm}}(\mathcal{S})$ can be written as in Corollary 15 where the expansion coefficients can be computed from the solution of the following convex quadratic program*

$$\alpha^* = \operatorname*{argmin}_{\alpha} \left\{ \frac{1}{2}\sum_{i,j=1}^{n}\sum_{y\neq y_i}\sum_{y'\neq y_j} \alpha_{iy}\alpha_{jy'}K_{iy,jy'} - \sum_{i=1}^{n}\sum_{y\neq y_i}\alpha_{iy} \right\} \tag{111a}$$

$$s.t. \quad \lambda n \sum_{y\neq y_i}\alpha_{iy} \leq 1,\ \forall i\in[n]; \quad \alpha_{iy}\geq 0,\ \forall i\in[n], y\in\mathcal{Y}\,, \tag{111b}$$

where $K_{iy,jy'} := k((x_i,y_i),(x_j,y_j)) + k((x_i,y),(x_j,y')) - k((x_i,y_i),(x_j,y')) - k((x_i,y),(x_j,y_j))$.
Proof *The primal program for soft-margin maximization can be expressed as*

$$\min_{f\in\mathcal{H},\xi} \mathcal{R}(f,\xi,\mathcal{S}) := \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i, \quad s.t.\ f(x_i,y_i) - f(x_i,y) \geq 1-\xi_i,\ \forall y\neq y_i; \quad \xi_i\geq 0,\ \forall i\in[n]\,.$$

Introducing dual variables α_{iy} for the margin constraints and ζ_i for the non-negativity constraints on ξ_i yields the Lagrange function (the objective has been divided by λ)

$$\mathcal{L}(f,\alpha) := \frac{1}{2}\langle f,f\rangle_{\mathcal{H}} + \frac{1}{\lambda n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\sum_{y\neq y_i}\alpha_{iy}\left[f(x_i,y_i) - f(x_i,y) - 1 + \xi_i\right] - \sum_{i=1}^{n}\xi_i\zeta_i\,.$$

Solving for f results in

$$\nabla_f \mathcal{L}(f,\alpha) = 0 \iff f(\cdot) = \sum_{i=1}^{n}\sum_{y\neq y_i}\alpha_{iy}\left[k(\cdot,(x_i,y_i)) - k(\cdot,(x_i,y))\right]$$

since for $f\in\mathcal{H}$, $\nabla_f f(x,y) = k(\cdot,(x,y))$. Solving for ξ_i implies $\lambda n\sum_{y\neq y_i}\alpha_{iy}\leq 1$ and plugging the solution back into $\mathcal{L}$ yields the (negative) objective function in the claim. Finally note that the representation in (50) can be obtained by identifying

$$\beta_{iy} := \begin{cases} -\alpha_{iy} & \text{if } y\neq y_i \\ \sum_{y\neq y_i}\alpha_{iy} & \text{if } y = y_i \end{cases}$$

■

- The multiclass SVM formulation of Crammer and Singer [2001] can be recovered as a special case for kernels that are diagonal with respect to the outputs, i.e. $k((x,y),(x',y')) = \delta_{y,y'}k(x,x')$. Notice that in this case the quadratic term in Eq. (111a) simplifies to

$$\sum_{i,j} k(x_i,x_j) \sum_y \alpha_{iy}\alpha_{jy} \left[1 + \delta_{y_i,y}\delta_{y_j,y} - \delta_{y_i,y} - \delta_{y_j,y}\right].$$

- The pairs (x_i, y) for which $\alpha_{iy} > 0$ are the *support pairs*, generalizing the notion of support vectors. As in binary SVMs their number can be much smaller than the total number of constraints. Notice also that in the final expansion contributions $k(\cdot,(x_i,y_i))$ will get non-negative weights whereas $k(\cdot,(x_i,y))$ for $y \neq y_i$ will get non-positive weights. Due to the equality constraints they will fulfill a balancing equation $\beta_{iy_i} - \sum_{y\neq y_i} \beta_{iy} = 0$.

4.1.6 Sparse Approximation

Proposition 16 shows that sparseness in the representation of $\hat{f}^{\text{sm}}$ is linked to the fact that only few α_{iy} in the solution to the dual problem in Eq. (111b) are non-zero. Note that each of these Lagrange multipliers is linked to the corresponding soft margin constraint $f(x_i,y_i) - f(x_i,y) \geq 1 - \xi_i$. Hence, sparseness is achieved, if only few constraints are active at the solution. While this may or may not be the case for a given sample, one can still exploit this observation to define a nested sequence of relaxations, where margin constraint are incrementally added. This corresponds to a constraint generation or cutting plane algorithm (cf. Bertsimas and Tsitsiklis [1997]) for the primal or – equivalently – a variable selection or column generation method for the dual program and has been investigated in Tsochantaridis et al. [2005]. In particular, one may iterate through the training examples according to some (fair) visitation schedule and greedily select constraints that are most violated at the current solution f, i.e. for the i-th instance one computes

$$\hat{y}_i = \operatorname*{argmax}_{y\neq y_i} f(x_i,y) = \operatorname*{argmax}_{y\neq y_i} p(y|x_i;f)\,, \tag{112}$$

and strengthens the current relaxation by including $\alpha_{i\hat{y}_i}$ in the optimization of the dual if $f(x_i,y_i) - f(x_i,\hat{y}_i) < 1 - \xi_i - \epsilon$. It is important to understand how many strengthening steps are necessary to achieve a reasonable close approximation to the original problem. The following theorem provides an answer to this question:

Theorem 17 (Tsochantaridis et al. [2005]) *Let $\bar{R} = \max_{i,y} K_{iy,iy}$ and chose $\epsilon > 0$. The described procedure, which optimizes Eq.* (108) *by greedily selecting constraints, will find an approximate solution where all constraints are fulfilled within a precision of ϵ, i.e. $r(x_i,y_i;f) \geq 1-\xi_i-\epsilon$ afer at most*

$$\max\left\{\frac{2n}{\epsilon}, \frac{8\bar{R}^2}{\lambda n \epsilon^2}\right\}$$

constraint selections according to Eq. (112).

Corollary 18 *Denote by $(\hat{f},\hat{\xi})$ an outer approximation of the problem in Proposition 16, minimizing $\mathcal{R}(f,\xi,\mathcal{S})$ while violating no constraint by more than ϵ (cf. Theorem 17). Then*

$$\mathcal{R}(\hat{f},\hat{\xi},\mathcal{S}) \leq \mathcal{R}(\hat{f}^{sm},\xi^*,\mathcal{S}) \leq \mathcal{R}(\hat{f},\hat{\xi},\mathcal{S}) + \epsilon$$

where $(\hat{f}^{sm},\xi^*)$ is the optimal solution.

Proof *The first inequality follows from the fact that $(\hat{f}, \hat{\xi})$ is an outer approximation, the second from the observation that by setting $\tilde{\xi} = \hat{\xi} + \epsilon$ one gets an admissible solution $(\hat{f}, \tilde{\xi})$ such that* $\mathcal{R}(\hat{f}, \tilde{\xi}, \mathcal{S}) = \frac{\lambda}{2}\|\hat{f}\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^n(\hat{\xi}_i + \epsilon) = \mathcal{R}(\hat{f}, \hat{\xi}, \mathcal{S}) + \epsilon$. ■

- Combined with an efficient QP solver, the above theorem guarantees a runtime polynomial in $n, \epsilon^{-1}, \bar{R}, \lambda^{-1}$. This is basically irrespective of special properties of the data set utilized – the only dependeny on the sample points x_i is through the radius $\bar{R}$.
- The remaining key problem is how to compute Eq. (112) efficiently. The answer depends on the specific form of the joint kernel k and/or the feature map Ψ. In many cases, efficient dynamic programming techniques exists, whereas in other cases one has to resort to approximations or use other methods to pre-select a set of candidate distractors for a training pair (x_i, y) (cf. Collins [2000]). It can be shown that the result in Theorem 17 holds as long as we are able to find ϵ-violating constraints, i.e. $\hat{y}_i$ such that $f(x_i, y_i) - f(x_i, \hat{y}_i) < 1 - \xi_i - \epsilon$.

4.1.7 Generalized Gaussian Processes Classification

The model equation Eq. (103) and the minimization of the regularized log-loss can be interpreted as a generalization of Gaussian process classification Williams [1998], Altun et al. [2004b] by assuming that $(f(x,\cdot))_{x\in\mathcal{X}}$ is a vector-valued zero mean Gaussian process; note that the covariance function C is defined over pairs $\mathcal{X}\times\mathcal{Y}$. For a given sample $\mathcal{S} = \{(x_i, y_i) : i \in [n]\}$ let $\tilde{\mathcal{S}} := \{(x_i, y) : i \in [n],\, y \in \mathcal{Y}\}$ and define a multi-index vector $F(\mathcal{S}) := (f(x_i, y))_{i,y}$ as the restriction of the stochastic process f to the augmented sample $\tilde{\mathcal{S}}$. Denote the kernel matrix by $K = (K_{iy,jy'})$, where $K_{iy,jy'} := C((x_i, y), (x_j, y'))$, and indices $i, j \in [n]$ and $y, y' \in \mathcal{Y}$, so that in summary: $F(\mathcal{S}) \sim \mathcal{N}(0, K)$. This induces a predictive model via Bayesian model integration according to

$$p(y|x;\mathcal{S}) = \int p(y|F(x,\cdot))p(F|\mathcal{S})dF\,. \tag{113}$$

For an i.i.d. sample the log-posterior for F can be written as

$$\ln p(F|\mathcal{S}) = -\frac{1}{2}F^T K^{-1} F + \sum_{i=1}^n [f(x_i, y_i) - g(x_i, F)] + const., \quad \hat{F}^{\mathrm{mp}}(\mathcal{S}) := \operatorname*{argmax}_f \ln p(F|\mathcal{S}) \tag{114}$$

Evoking the representer theorem we know that $\hat{F}^{\mathrm{mp}}(\mathcal{S})_{iy} = \sum_{j=1}^n \sum_{y'\in\mathcal{Y}} \alpha_{iy} K_{iy,jy'}$ which we plug into Eq. (114) to arrive at

$$\min_\alpha \alpha^T \mathbf{K}\alpha - \sum_{i=1}^n \left(\alpha^T \mathbf{K} e_{iy_i} + \log \sum_{y\in\mathcal{Y}} \exp\left[\alpha^T \mathbf{K} e_{iy}\right] \right)\,. \tag{115}$$

Notice that for $f = \sum_{i,y} \alpha_{iy} k(\cdot, (x_i, y))$ the first term is equivalent to the squared RKHS norm of $f \in \mathcal{H}$ since $\langle f, f\rangle_{\mathcal{H}} = \sum_{i,j}\sum_{y,y'} \alpha_{iy}\alpha_{jy'} \langle k(\cdot,(x_i,y)), k(\cdot,(x_j,y'))\rangle$ which reduces to $k((x_i, y), (x_j, y'))$ due to the reproducing property. Again, the key issue in solving Eq. (115) is how to achieve spareness in the expansion for $\hat{F}^{\mathrm{mp}}$.

Semiparametric Latent Factor Models by Teh et al. !! Mixing multiple GPs.

4.2 Markov Networks and Kernels

In Section 4.1 no assumptions about the specific structure of the joint kernel defining the model in Eq. (103) has been made. In the following, we will focus on a more specific setting with multiple outputs, where dependencies are modeled by a conditional independence graph. This approach is motivated by the fact that independently predicting individual responses by building marginal response models will often be suboptimal and explicitly modelling these interactions can be of crucial importance.

4.2.1 Markov Networks and Factorization Theorem

Denote predictor variables by X, response variables by Y and define $Z := (X, Y)$ with associated sample space $\mathcal{Z}$. We use Markov networks as the modeling formalism for representing dependencies between and among response variables and covariates.

Definition 19 *A* **conditional independence graph** *(or Markov network) is an undirected graph $\mathcal{G} = (Z, E)$ such that for any pair of variables $(Z_i, Z_j) \notin E$ if and only if $Z_i \perp\!\!\!\perp Z_j | Z_{*-\{i,j\}}$.*

The above definition is based on the pairwise Markov property, but by virtue of the separation theorem (cf. e.g. **?**) this implies the global Markov property for distributions with full support. The global Markov property says that for disjoint subsets $U, V, W \subseteq Z$ where W separates U from V in $\mathcal{G}$ one has that $U \perp\!\!\!\perp V | W$. Even more important in the context of this paper is the factorization result due to Hammersley and Clifford [1971] and Besag [1974].

Theorem 20 *Given a random vector Z with independence graph $\mathcal{G}$. Any density function for Z with full support factorizes over $\mathcal{C}(\mathcal{G})$, the set of maximal cliques of $\mathcal{G}$:*

$$p(z) = \exp\left[\sum_{c \in \mathcal{C}(G)} f_c(z_c)\right] \tag{116}$$

where f_c are clique compatibility functions that depend on z only through the restriction z_c.

The significance of this result is that in order to estimate the distribution for Z, one only needs to estimate the simpler functions f_c.

4.2.2 Kernel Decomposition over Markov Networks

It is of interest to analyse the structure of kernels k that generate Hilbert spaces $\mathcal{H}$ of functions that are consistent with a graph.

Definition 21 *A function $f : \mathcal{Z} \to \mathbb{R}$ is* **compatible** *with a conditional independence graph $\mathcal{G}$, if f decomposes additively as $f(z) = \sum_{c \in \mathcal{C}(\mathcal{G})} f_c(z_c)$. A Hilbert space $\mathcal{H}$ over $\mathcal{Z}$ is compatible with $\mathcal{G}$, if every function $f \in \mathcal{H}$ is compatible. Such f and $\mathcal{H}$ are also called $\mathcal{G}$-compatible.*

Proposition 22 *Let $\mathcal{H}$ be a $\mathcal{G}$-compatible RKHS with kernel k. There are functions $k_{cd} : \mathcal{Z}_c \times \mathcal{Z}_d \to \mathbb{R}$ such that the kernel decomposes as*

$$k(u, z) = \sum_{c,d \in \mathcal{C}} k_{cd}(u_c, z_d)\,.$$

Proof *Let $z \in \mathcal{Z}$, then $k(\cdot, z) = k(z, \cdot) \in \mathcal{H}$ and by assumption for all $u, z \in \mathcal{Z}$ there exist functions $k_d(\cdot; z)$ such that*

$$k(u, z) = \sum_d k_d(u_d; z) = \sum_c k_c(z_c; u)\,.$$

The left hand sum involves functions restricted to cliques of the first argument, whereas the right hand sum involves functions restricted to cliques of the second argument. Then a simple lemma implies that there have to be function $k_{cd}(u_c, z_d)$ as claimed. ■

Lemma 23 *Let $\mathcal{X}$ be a set of n-tupels and $f_i, g_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for $i \in [n]$ functions such that $f_i(x, y) = f_i(x_i, y)$ and $g_i(x, y) = g_i(x, y_i)$. If $\sum_i f_i(x_i, y) = \sum_j g_j(x, y_j)$ for all x, y, then there exist functions h_{ij} such that $\sum_i f_i(x_i, y) = \sum_{i,j} h_{ij}(x_i, y_j)$.*

- Proposition 22 is useful for the design of kernels, since it states that only kernels allowing an additive decomposition into local functions k_{cd} are compatible with a given Markov network $\mathcal{G}$. Lafferty et al. [2004] have pursued a similar approach by considering kernels for RKHS with functions defined over $\mathcal{Z}_{\mathcal{C}} := \{(c, z_c) : c \in c, z_c \in \mathcal{Z}_c\}$.
- An illuminating example of how to design kernels via the decomposition in Proposition 22 is the case of it conditional Markov chains, for which models based on joint kernels have been proposed in Lafferty et al. [2001], Collins [2002], Altun and Hofmann [2003], Taskar et al. [2003]. Given an input sequences $X = (X_t)_{t\in[T]}$, the goal is to predict a sequence of labels or class variables $Y = (Y_t)_{t\in[T]}$, $Y_t \in \mathcal{Y}$. Dependencies between class variables are modeled in terms of a Markov chain, whereas outputs Y_t are assumed to depend (directly) on the observation X_{t-1}, X_t, X_{t+1}. Notice that this goes beyond the standard Hidden Markov Model structure (cf. Rabiner [1989]): the clique set is given by $\mathcal{C} := \{c_t := (x_t, y_t, y_{t+1}), c'_t := (x_{t+1}, y_t, y_{t+1}) : t \in [T-1]\}$. We assume an input kernel k is given and introduce indicator vectors (or dummy variates) $I(Y_{\{t,t+1\}}) := (I_{\omega,\omega'}(Y_{\{t,t+1\}}))_{\omega,\omega'\in\mathcal{Y}}$. Now we can define the local kernel functions as

$$k_{cd}(z_c, z'_d) := \left\langle I(y_{\{s,s+1\}}), I(y'_{\{t,t+1\}}) \right\rangle \begin{cases} k(x_s, x_t) & \text{if } c = c_s \text{ and } d = c_t \\ k(x_{s+1}, x_t) & \text{if } c = c'_s \text{ and } d = c_t \\ k(x_s, x_{t+1}) & \text{if } c = c_s \text{ and } d = c_{t+1} \\ k(x_{s+1}, x_{t+1}) & \text{if } c = c'_s \text{ and } d = c'_t \end{cases} \tag{117}$$

  Notice that the inner product between indicator vectors is zero, unless the variable pairs are in the same configuration.
  Conditional Markov chain models have found widespread applications, in natural language processing (e.g. for part of speech tagging and shallow parsing, cf. Sha and Pereira [2003]), in information retrieval (e.g. for information extraction, cf. McCallum et al. [2005]), or in computational biology (e.g. for gene prediction Culotta et al. [2005]).

4.2.3 Clique-based Sparse Approximation

Proposition 22 immediately leads to an alternative version of the representer theorem as observed by Lafferty et al. [2004] and Altun et al. [2004b].

Corollary 24 *If $\mathcal{H}$ is compatible with some graph $\mathcal{G}$ then in the same setting as in Corollary 15, the optimizer $\hat{f}$ can be written as*

$$\hat{f}(u) = \sum_{i=1}^{n} \sum_{c \in \mathcal{C}} \sum_{y_c \in \mathcal{Y}_c} \beta^i_{c,z_c} \sum_{d \in \mathcal{C}} k_{cd}((x_{ic}, y_c), u_d) \tag{118}$$

here x_{ic} are the variables of of x_i belonging to clique c and $\mathcal{Y}_c$ is the subspace of $\mathcal{Z}_c$ that contains response variables.

Proof *According to Proposition 22 the kernel k of $\mathcal{H}$ can be written as $k(z,u) = \sum_{c,d} k_{cd}(z_c, u_d)$ plugging this into the expansion of Corollary 15 yields*

$$\hat{f}(u) = \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \beta_{iy} \sum_{c,d \in \mathcal{C}} k_{cd}((x_{ic}, y_c), u_d) = \sum_{i=1}^{n} \sum_{c,d \in \mathcal{C}} \sum_{y_c \in \mathcal{Y}_c} k_{cd}((x_{ic}, y_c), u_d) \sum_{y'} I_{y_c}(y'_c) \beta_{iy'}.$$

Setting $\beta^i_{c,y_c} := \sum_{y'} I_{y_c}(y'_c)\beta_{iy'}$ completes the proof. ■

- Notice that the number of parameters in the representation Eq. (118) scales with $n \cdot \sum_{c \in \mathcal{C}} |\mathcal{Y}_c|$ as opposed to $n \cdot |\mathcal{Y}|$ in Eq. (50). For reasonably small cliques, this will be a significantly more compact representation. Notice also that the evaluation of a sub-kernel k_{cd} will typically be more efficient than evaluating k.
- In spite of this improvement, the number of terms in the expansion in Eq. (118) is often too large to deal with in practice. In this case, one can pursue a reduced set approach, which selects a subset of variables to be included in a sparsified expansion. This has been proposed in Taskar et al. [2003] for the soft margin maximization problem as well as in Lafferty et al. [2004] for conditional random fields and in Altun et al. [2004a] for Gaussian processes. For instance, in Lafferty et al. [2004] parameters $\beta^i_{cy_c}$ that maximize the functional gradient of the regularized log-loss are greedily included in the reduced set. In Taskar et al. [2003] a similar selection criterion is utilized with respect to margin violations, leading to an SMO-like optimization algorithm (Platt [1999]).

4.2.4 Exact Inference

In dealing with structured or interdependent response variables, computing marginal probabilities of interest or computing the most probable response may be non-trivial. Therefore we provide a short overview of *inference* techniques for Markov networks. In general, MAP inference in Markov networks is NP-hard (cf. Cowell et al. [1999]), however, there are several special classes of networks with small tree width for which efficient inference algorithms exist.

The standard algorithm for inference in Markov network is the junction tree algorithm (cf. Jensen et al. [1990], Dawid [1992]). The algorithm works on a data structure that is called the *junction tree*, which is a special type of tree defined over the cliques of the graph fulfilling the *running intersection* property.

Definition 25 *A* **junction tree** *for some graph $\mathcal{G}$ is a tree $\mathcal{T} = (\mathcal{C}(\mathcal{G}), E)$ over the cliques of the graph, with the following property: for all pairs of cliques $c, c' \in \mathcal{C}$ and every clique $d \in \mathcal{C}$ on the (unique) path from c to c' ones has that $c \cap c' \subseteq d$.*

The standard procedure to obtain a junction tree first triangulates $\mathcal{G}$ and then applies a maximum spanning tree algorithm to the weighted clique graph $(\mathcal{C}(\mathcal{G}), W)$ with edge weights $W(c,d) = |c \cap d|$ (cf. Shibata [1988], Jensen and Jensen [1994]).

Once this data structure is complied, inference can be performed by local message passing between neigboring cliques in the junction tree. We will restrict ourselves to inference in the conditional model of Eq. (103) and assume $\mathcal{H}$ is consistent with a triangulated independence graph $\mathcal{G}$. All cliques that do not depend on the response variables can be removed prior to the triangulation.

Let us investigate the problem of computing marginal probabilities for y_t over some target clique $t \in \mathcal{C}$. We define t to be the root of the junction tree $\mathcal{T}$ and order cliques $c_1, \ldots, c_{k-1}, c_k = t$ consistent with the partial order defined by $\mathcal{T}$, i.e. starting from the leaves to the root. Denote by $s(c) := c \cap c'$ separator sets where c' is the ancester of c in the tree $\mathcal{T}$ rooted at t. Moreover define $r(c) := c - s(c)$ and as a shorthand $r(j) := r(c_j)$. Now

$$p(y_t|x, f) \propto e^{f_t} \sum_{y \sim y_t} \prod_{c \neq r} e^{f_c} = e^{f_t} \sum_{y_{r(k-1)}} \cdots \sum_{y_{r(2)}} \sum_{y_{r(1)}} \prod_{j=1}^{k-1} e^{f_{c(j)}} \tag{119}$$

where we have dropped the arguments of compatibility functions f_c and all sums run over configurations consistent with the (fixed) configuration y_t, $y \sim y_t$. The key observation is that the nested sum can be evaluated efficiently by local computations. To that extent one conveniently defines messages propagated inward from leaves towards the root as follows

$$\mu_{c \to c'}(y_{s(c)}) := \sum_{y_{r(c)}} e^{f_c} \prod_{d : d' = c} \mu_{d \to c}(y_{s(d)}) \,, \tag{120}$$

where the product of received messages from successor nodes is empty at leave nodes. It can be shown that by propagating messages outward from the root, all clique marginals can be computed simultaneously (basically, all nodes are considered as roots for $\mathcal{T}$).

- In the case of the conditional or hidden Markov chain, the junction tree algorithm is equivalent to the well-known forward-backward algorithm Baum [1972]. It can be utilized to efficiently compute the expected (conditional) canoncial statistics $\mathbf{E}[\Psi(Y, x)]$ needed for the optimization methods discussed above (cf. Lafferty et al. [2001]).
- By substituting the sum in the definition of the messages in Eq. (120) with a max operation, one can compute the most probable output

$$y_t^*(x, f) = \operatorname*{argmax}_{y_t} \left\{ e^{f_t} \prod_{c : t = c'} \mu'_{c \to t}(y_{s(c)}) \right\}, \; \mu'_{c \to c'}(y_{s(c)}) := \max_{y_{r(c)}} \left\{ e^{f_c} \prod_{d : d' = c} \mu_{d \to c}(y_{s(d)}) \right\} .$$

  This generalizes Viterbi decoding (cf. Forney Jr. [1973]). With some additional bookkeeping, one can also compute the two most probable outputs, which is necessary in order to evaluate Eq. (112), if $p(\hat{y}_i|x_i) < p(y_i|x_i)$. For the Hidden Markov Model case, this has been developed in Schwarz and Chow [1990].

4.2.5 Variational inference

The junction tree algorithm is tractable as long as the state spaces for every clique remain sufficiently small. However, many interesting models do not fall into this class, in which case one typically resorts to approximate inference methods. A systematic way to go about doing this is to formulate inference problems as a constrained optimization problem and to develop and analyse various approximate optimization methods, thereby leveraging sophisticated methodology from the area of operations research and mathematical programming.

We come back to the beginning of this section and elaborate on certain properties of exponential families. We have seen that within an exponential family, there is a mapping from the canonical parameters to the mean parameters $\mu(\theta) := \mathbf{E}_\theta(\Phi(X))$. This mapping is one-to-one, if the family is minimal. Define the Fenchel-Legendre conjugate (cf. Rockafellar [1970]) of the log-partition function

$$g^*(\theta) = \sup_{\theta \in \theta} \{\langle \mu, \theta \rangle - g(\theta)\}, \tag{121}$$

which for exponential families g^* is basically equal to the negative Boltzmann-Shannon entropy, namely:

Theorem 26 (Wainwright and Jordan [2003a]) *For any $\mu \in \mathcal{M}(\Phi)$, where $\mathcal{M}(\Phi) := \{\mu \in \mathbb{R}^m : \exists p \text{ s.t. } \int \Phi(x) p(x) \nu(dx) = \mu\}$, let $\theta(\mu)$ such that $\mu = \mu(\theta)$, then*

$$g^*(\mu) = \begin{cases} H(\theta(\mu)),\ H(\theta) := -\int_{\mathcal{X}} p(x;\theta) \log p(x;\theta) \nu(dx), & \text{if } \mu \in ri\mathcal{M} \\ +\infty & \text{if } \mu \notin cl\mathcal{M} \\ \lim_{n\to\infty} H(\theta(\mu^n)) & \text{otherwise} \end{cases} \tag{122}$$

where the last case is over a sequences $\{\mu^n\} \subset ri\mathcal{M}$ such that $\lim \mu^n = \mu$. Moreover,

$$g(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - g^*(\mu)\}. \tag{123}$$

Note that for every $\theta \in \theta$ the supremum is attained at $\mu = \mu(\theta)$. Computing the log partition function as well as the mean parameter mapping is thus reduced to solving the optimization problem in Eq. (123). The marginal polytope $\mathcal{M}$ for finite sample space can be characterized as follows.

Lemma 27 (Wainwright and Jordan [2003a]) *For finite sample spaces $\mathcal{X}$ and arbitrary Φ there exists an integer J as well as vectors a_j and numbers b_j for $j \in [J]$ such that*

$$\mathcal{M}(\Phi) = \{\mu \in \mathbb{R}^d : \langle a_j, \mu \rangle \leq b_j, \forall j \in [J]\} \tag{124}$$

Proof *$\mathcal{M}(\Phi)$ can be written as the convex hull of a finite set of vectors, $\mathcal{M}(\Phi) = \operatorname{conv} \Phi(\mathcal{X})$. By the Minokwski-Weil theorem this implies the claim.* ■

In the general case, the optimization problem in Eq. (123) is challenging, since (i) the marginal polytope may have a finite, but very large number of facets, and (ii) the negative entropy $H(\theta(\mu))$ is often not given explicitly as a function of μ and may not even be efficiently computable. Different variational techniques offer different strategies for approximating the exact inference problem.

- The mean field approach is based on the idea of an inner approximation, namely to restrict the admissible domain to a subset $\mathcal{M}_0(\Phi) \subseteq \mathcal{M}(\Phi)$ and maximize H over $\mathcal{M}_0$. Then every solution μ_0 over $\mathcal{M}_0$ will apparently yield a lower bound on the true supremum since $\sup_{\mu \in \mathcal{M}_0} \{\langle \mu, \theta \rangle\} - g^*(\mu) \leq g(\theta)$. A generic way of defining $\mathcal{M}_0$ for statistics Φ arising from graphical models is to pick a tractable substructure $\mathcal{H}$ of the independence graph, such as a tree or – in the most extreme case – the empty graph and then to set

$$\mathcal{M}_0 := \{\mu \in \mathcal{M}(\Phi) : \exists \theta \in \mathcal{E}(\mathcal{H}), \text{ s.t. } \mu = \mathbf{E}_\theta[\Phi(X)]\} \tag{125}$$

where $\mathcal{E}(\mathcal{H}) \subseteq \theta$ denotes the set of canoncial parameters that are consistent with the selected subgraph $\mathcal{H}$.

- The belief propagation algorithm of Pearl [1988], exact for graphs without cycles, has been applied as "loopy" belief propagation to graphs with cycles. Although not guaranteed to converge, one can often obtain reasonable approximations to the MAP or maximial margin probabilities in practice (Murphy et al. [1999]). In this context, the improved Bethe free energy variational approach proposed by Yedidia et al. [2003] can be interpreteted as approximating the entropy by an approximate entropy. For m-dimensional finite sampling spaces $\mathcal{X}$ define

$$\tilde{H}_j(\mu) := \sum_{x_j} \mu_j(x_j) \log \mu_j(x_j), \quad \tilde{I}_{jk}(\mu) := \sum_{x_j, x_k} \mu_{jk}(x_j, x_k) \log \frac{\mu_{jk}(x_j, x_k)}{\mu_j(x_j)\mu_k(x_k)}. \tag{126}$$

  Then approximate the entropy by $\tilde{H}(\mu) := \sum_{j=1}^m H_j(\mu_j) - \sum_{j,k} I_{j,k}(\mu_j, \mu_k)$ which is the correct entropy of a graph without cycles (Wainwright and Jordan [2003a]). This is combined with an outer approximation of the marginal polytope which is also motivated by trees, namely

$$\mathcal{M}_+ := \left\{ \mu \geq 0 : \sum_{x_j} \mu_j(x_j) = 1, \ \sum_{x_k} \mu_{jk}(x_j, x_k) = \mu_j(x_j) \right\} \tag{127}$$

  which formulates necessary conditions on μ.
- More sophisticated approximation scheme, for instance, combining a log-determinant relaxation with a semi-definite constraint on the approximating covariance matrix can be found in Wainwright and Jordan [2003b].

Acknowledgments

The main body of this paper was written at the Mathematisches Forschungsinstitut Oberwolfach. The authors would like to express thanks for the support during that time. National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council. This work was supported by grants of the ARC and by the Pascal Network of Excellence.

xxx ackn. for reading the manuscript

References

M. A. Aizerman, É. M. Braverman, and L. I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. In P. Langley, editor, *Proceedings of the International Conference on Machine Learning*, pages 9–16, San Francisco, California, 2000. Morgan Kaufmann Publishers.

Noga Alon, Shai Ben-David, N. Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Proc. of the 34rd Annual Symposium on Foundations of Computer Science*, pages 292–301. IEEE Computer Society Press, Los Alamitos, CA, 1993.

Y. Altun and T. Hofmann. Large margin methods for label sequence learning. In *8th European Conference on Speech Communication and Technology*, 2003.

Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In *International Conference on Machine Learning*, 2003.

Y. Altun, T. Hofmann, and A. J. Smola. Exponential families for conditional random fields. In *Uncertainty in Artificial Intelligence UAI*, pages 2–9, 2004a.

Y. Altun, T. Hofmann, and A. J. Smola. Gaussian process classification for segmenting and annotating sequences. In *International Conference on Machine Learning ICML*, 2004b.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

Peter Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Localized rademacher averages. In *Proceedings of the 15th conference on Computational Learning Theory COLT'02*, pages 44–58, 2002.

J. Basilico and T. Hofmann. Unifying collaborative and content-based filtering. In *International Conference on Machine Learning ICML*, 2004.

L. E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3, 1972.

S. Ben-David, N. Eiron, and P.M. Long. On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.*, 66(3):496–514, 2003.

K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

K. P. Bennett, A. Demiriz, and J. Shawe-Taylor. A column generation algorithm for boosting. In P. Langley, editor, *Proceedings of the International Conference on Machine Learning*, San Francisco, 2000. Morgan Kaufmann Publishers.

C. Berg and G. Forst. *Potential Theory on Locally Compact Abelian Groups*. Springer-Verlag, Berlin, 1975.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, New York, 1984.

D. Bertsimas and J. Tsitsiklis. *Introduction to linear programming*. Athena Scientific, 1997.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, 36(B):192–326, 1974.

V. Blanz, B. Schölkopf, H. Bülthoff, C. Burges, V. Vapnik, and T. Vetter. Comparison of view-based object recognition algorithms using realistic 3D models. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Artificial Neural Networks ICANN'96*, volume 1112 of *Lecture Notes in Computer Science*, pages 251–256, Berlin, 1996. Springer-Verlag.

P. Bloomfield and W.L. Steiger. *Least absolute deviations: theory, applications and algorithms.* Birkhäuser, Boston, 1983.

S. Bochner. Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse. *Mathematische Annalen*, 108:378–410, 1933.

B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the Annual Conference on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 2004. submitted.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

L. Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11(7):1493–1518, 1999.

M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, J. r. Ares M, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–267, Jan 2000.

C. J. C. Burges. Simplified support vector decision rules. In L. Saitta, editor, *Proceedings of the International Conference on Machine Learning*, pages 71–77, San Mateo, CA, 1996. Morgan Kaufmann Publishers.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

C. J. C. Burges and V. Vapnik. A new method for constructing artificial neural networks. Interim technical report, ONR contract N00014 - 94-c-0186, AT&T Bell Laboratories, 1995.

O. Chapelle and Z. Harchaoui. A machine learning approach to conjoint analysis. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, Cambridge, MA, 2005. MIT Press.

O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5), 1999.

E. Charniak. A maximum-entropy-inspired parser. In *North American Association of Computational Linguistics (NAACL 1)*, pages 132–139, 2000.

E. Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pages 598–603, 1997.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *Siam Journal of Scientific Computing*, 20(1):33–61, 1999.

K. De Cock and B. De Moor. Subspace angles between ARMA models. *Systems and Control Letter*, 46:265–270, 2002.

R. Cole and R. Hariharan. Faster suffix tree construction with missing suffix links. In *Proceedings of the Thirty Second Annual Symposium on the Theory of Computing*, Portland, OR, USA, May 2000. ACM.

M. Collins. Discriminative reranking for natural language parsing. In *Proc. 17th International Conf. on Machine Learning*, pages 175–182. Morgan Kaufmann, San Francisco, CA, 2000.

M. Collins. Discriminative training methods for hidden markov models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

M. Collins and N. Duffy. Convolution kernels for natural language. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2001. MIT Press.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. In *Proc. 13th Annu. Conference on Comput. Learning Theory*, pages 158–169. Morgan Kaufmann, San Francisco, 2000.

D. Cook, A. Buja, and J. Cabrera. Projection pursuit indices based on orthonormal function expansions. *Journal of Computational and Graphical Statistics*, 2:225–250, 1993.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

R. Cowell, A. David, S. Lauritzen, and D. Spiegelhalter. *Probabilistic Networks and Expert Sytems*. Springer, New York, 1999.

D. Cox and F. O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*, 18:1676–1695, 1990.

K. Crammer. *Online Learning for Complex Categorical Problems*. PhD thesis, Hebrew University of Jerusalem, 2005.

K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In N. Cesa-Bianchi and S. Goldman, editors, *Proceedings of the Annual Conference on Computational Learning Theory*, pages 35–46, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2001.

K. Crammer and Y. Singer. Loss bounds for online category ranking. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2005.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 367–373, Cambridge, MA, 2002. MIT Press.

A. Culotta, D. Kulp, and A. McCallum. Gene prediction with conditional random fields. Technical Report UM-CS-2005-028, University of Massachusetts, Amherst, 2005.

J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43, 1972.

D. Das and P. Sen. Restricted canonical correlations. *Linear Algebra and its Applications*, 210: 29–47, 1994.

J. Dauxois and G. M. Nkiet. Nonlinear canonical analysis and independence tests. *Annals of Statistics*, 26(4):1254–1278, 1998.

A. P. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2:25–36, 1992.

D. DeCoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46: 161–190, 2002. Also: Technical Report JPL-MLTR-00-1, Jet Propulsion Laboratory, Pasadena, CA, 2000.

O. Dekel, C.D. Manning, and Y. Singer. Log-linear models for label ranking. In *Advances in Neural Information Processing Systems 15*, 2003.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:380–393, 1997.

K. I. Diamantaras and S. Y. Kung. *Principal Component Neural Networks*. John Wiley and Sons, New York, 1996.

N.R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley and Sons, New York, NY, 1981.

H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161, Cambridge, MA, 1997. MIT Press.

S. Dumais. Using SVMs for text categorization. *IEEE Intelligent Systems*, 13(4), 1998. In: M. A. Hearst, B. Schölkopf, S. Dumais, E. Osuna, and J. Platt: Trends and Controversies - -Support Vector Machines.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

J. H. J. Einmal and D. M. Mason. Generalized quantile processes. *Annals of Statistics*, 20(2): 1062–1078, 1992.

A. Elisseeff and J. Weston. A kernel method for multi-labeled classification. In *Advances in Neural Information Processing Systems 14*. MIT press, 2001.

M. Everingham, L. Van Gool, C. Williams, and A. Zisserman. Pascal visual object classes challenge results. Available from www.pascal-network.org, 2005.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.

C. H. FitzGerald, C. A. Micchelli, and A. Pinkus. Functions that preserve families of positive semidefinite matrices. *Linear Algebra and its Applications*, 221:83–102, 1995.

R. Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, New York, 1989.

G. D. Forney Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61:268–278, 1973.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learing*, pages 148–146. Morgan Kaufmann Publishers, 1996.

J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266, 1987.

J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23(9):881–890, 1974.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Stanford University, Dept. of Statistics, 1998.

T. T. Frieß and R. F. Harrison. Linear programming support vector machiens for pattern classification and regression estimation and the set reduction algorithm. TR RR-706, University of Sheffield, Sheffield, UK, 1998.

T. Gärtner. A survey of kernels for structured data. *SIGKDD Explorations*, 5(1):49–58, July 2003.

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proceedings of the 19th International Conference on Machine Learning ICML*, 2002.

F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

P. Green and B. Yandell. Semi-parametric generalized linear models. In *Proceedings 2nd International GLIM Conference*, number 32 in Lecture Notes in Statistics, pages 44–55. Springer Verlag, New York, 1985.

A. Gretton, O. Bousquet, A.J. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In S. Jain and W.-S. Lee, editors, *Proceedings Algorithmic Learning Theory*, 2005a.

A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. Logothetis. Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, 2005b.

M. Gribskov and N. L. Robinson. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers and Chemistry*, 20(1):25–33, 1996.

D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, June 1997. ISBN 0 - 521 - 58519 - 8.

J. M. Hammersley and P. E. Clifford. Markov fields on finite graphs and lattices. unpublished manuscript, 1971.

J. A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82:267–270, 1987.

D. Haussler. Convolutional kernels on discrete structures. Technical Report UCSC-CRL-99 - 10, Computer Science Department, UC Santa Cruz, 1999.

P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis. Support vector novelty detection applied to jet engine vibration spectra. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 946–952, Cambridge, MA, 2001. MIT Press.

M. Hein, O. Bousquet, and B. Schölkopf. Maximal margin classification for metric spaces. *Journal of Computer and System Sciences*, 2005. in press.

R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, 2002.

R. Herbrich and R. C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2002.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132, Cambridge, MA, 2000. MIT Press.

R. Hettich and K. O. Kortanek. Semi-infinite programming: theory, methods, and applications. *SIAM Rev.*, 35(3):380–429, 1993.

D. Hilbert. Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen. *Nachrichten der Akademie der Wissenschaften in Göttingen, Mathematisch-Physikalische Klasse*, pages 49–91, 1904.

A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

A. D. Holub, M. Welling, and P. Perona. Combining generative models and fisher kernels for object recognition. In *Proceedings of the International Conference on Computer Vision*, 2005.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 and 498 – 520, 1933.

H. Hotelling. Relations between two seets of variates. *Biometrika*, 28:321–377, 1936.

P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

P. J. Huber. Projection pursuit. *Annals of Statistics*, 13(2):435–475, 1985.

T. S. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*, 1999a.

T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1999b.

Tony Jebara and Imre Risi Kondor. Bhattacharyya and expected likelihood kernels. In Bernhard Schölkopf and Manfred K Warmuth, editors, *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, number 2777 in LNAI, pages 57–71, 2003.

F. V. Jensen and F. Jensen. Optimal junction trees. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 360–366. Morgan Kaufmann Publishers, San Francisco, 1994.

F. V. Jensen, S. L . Lauritzen, and K. G. Olesen. Bayesian updates in causal probabilistic networks by local computation. *Computational Statistics Quaterly*, 4:269–282, 1990.

T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. The Kluwer International Series In Engineering And Computer Science. Kluwer Academic Publishers, Boston, May 2002. ISBN 0 - 7923 - 7679-X.

T. Joachims. A support vector method for multivariate performance measures. In *Proc. ICML*, 2005.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142, Berlin, 1998. Springer.

I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, 1986.

M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society A*, 150(1):1–36, 1987.

M.I. Jordan, P.L. Bartlett, and J.D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, UC Berkeley, 2003.

J. P. Kahane. *Some Random Series of Functions*. Cambridge University Press, 1968.

K. Karhunen. Zur Spektraltheorie stochastischer Prozesse. *Annales Academiae Scientiarum Fennicae*, 37, 1946.

W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939.

H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the* 20^{th} *International Conference on Machine Learning (ICML)*, Washington, DC, United States, 2003.

Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. Kernels on graphs. In K. Tsuda, B. Schölkopf, and J.P. Vert, editors, *Kernels and Bioinformatics*, Cambridge, MA, 2004. MIT Press.

J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

Kwang In Kim, Matthias O. Franz, and Bernhard Schölkopf. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1351–1366, 2005.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

I. R. Kondor and J. D. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the ICML*, 2002.

W. Krauth and M. Mézard. Learning algorithms with optimal stability in neural networks. *Journal of Physics A*, 20:745–752, 1987.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proc. 2^{nd} Berkeley Symposium on Mathematical Statistics and Probabilistics*, pages 481–492, Berkeley, 1951. University of California Press.

J. Lafferty, X. Zhu, and Y. Liu. Kernel conditional random fields: Representation and clique selection. In *Proceedings of the 21st International Conference (ICML)*, 2004.

J. D. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic modeling for segmenting and labeling sequence data. In *18th International Conference on Machine Learning ICML*, 2001.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Q.V. Le, T. Sears, and A.J. Smola. Nonparametric quantile regression. Technical report, National ICT Australia, June 2005. Available at http://sml.nicta.com.au/∼quoc.le.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for SVM protein classification. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 564–575, 2002a.

C. Leslie, E. Eskin, J. Weston, and W. S. Noble. Mismatch string kernels for SVM protein classification. In *Advances in Neural Information Processing Systems 15*, volume 15, Cambridge, MA, 2002b. MIT Press.

H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. Technical Report 2000 - 79, NeuroCOLT, 2000. Published in: T. K. Leen, T. G. Dietterich and V. Tresp (eds.), *Advances in Neural Information Processing Systems 13*, MIT Press, 2001.

M. Loève. *Probability Theory II*. Springer-Verlag, New York, 4 edition, 1978.

D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, Reading, second edition, May 1984. ISBN 0 - 201 - 15794 - 2.

David M. Magerman. Learning grammatical stucture using statistical decision-trees. In *Grammatical Inference: Learning Syntax from Sentences, 3rd International Colloquium, ICGI-96, Montpellier, France, September 25 - 27, 1996, Proceedings*, volume 1147 of *Lecture Notes in Artificial Intelligence*, pages 1–21. Springer, Berlin, 1996.

Y. Makovoz. Random approximants and neural networks. *Journal of Approximation Theory*, 85: 98–109, 1996.

O. L. Mangasarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13:444–452, 1965.

R.J. Martin. A metric for ARMA processes. *IEEE Transactions on Signal Processing*, 48(4): 1164–1170, 2000.

A. McCallum, K. Bellare, and F. Pereira. A conditional random field for discriminatively-trained finite-state string edit distance. In *Conference on Uncertainty in AI (UAI)*, 2005.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983.

C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

T. Melzer, M. Reiter, and H. Bischof. Kernel canonical correlation analysis. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Proceedings of the International Conference on Artificial Neural Networks*, pages 353–360, Berlin, 2001. Springer.

S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 2001. submitted.

S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures on Machine Learning*, number 2600 in LNAI, pages 1–40. Springer, 2003.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, A 209:415–446, 1909.

H. Meschkowski. *Hilbertsche Räume mit Kernfunktion*. Springer-Verlag, Berlin, 1962.

S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and de-noising in feature spaces. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 536–542, Cambridge, MA, 1999. MIT Press.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. J. Smola, and K.-R. Müller. Learning discriminative and invariant nonlinear features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):623–628, 2003.

M. Minsky and S. Papert. *Perceptrons: An Introduction To Computational Geometry*. MIT Press, Cambridge, MA, 1969.

V. A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer, 1984.

K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks ICANN'97*, volume 1327 of *Lecture Notes in Computer Science*, pages 999–1004, Berlin, 1997. Springer-Verlag.

K. Murphy, Y. Weiss, and M.I. Jordan. Loopy belief propagation for approximate inference: an empirical study. In *Proceedings Uncertainty in AI (UAI)*, 1999.

M. K. Murray and J.W. Rice. *Differential geometry and statistics*, volume 48 of *Monographs on statistics and applied probability*. Chapman and Hall, 1993.

G. Navarro and M. Raffinot. Fast regular expression search. In *Proceedings of WAE*, number 1668 in LNCS, pages 198–212. Springer, 1999.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, A*, 135:370–384, 1972.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, 1999.

D. Nolan. The excess mass ellipsoid. *Journal of Multivariate Analysis*, 39:348–371, 1991.

N. Oliver, B. Schölkopf, and A. J. Smola. Natural regularization in SVMs. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 51–60, Cambridge, MA, 2000. MIT Press.

F. O'Sullivan, B. Yandell, and W. Raynor. Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.*, 81:96–103, 1986.

E. Parzen. Statistical inference on time series by rkhs methods. In R. Pyke, editor, *Proceedings 12th Biennial Seminar*, Montreal, 1970. Canadian Mathematical Congress. 1 - 37.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufman, 1988.

K. Pearson. On lines and planes of closest fit to points in space. *Philosophical Magazine*, 2 (sixth series):559–572, 1901.

A. Pinkus. Strictly positive definite functions on a real inner product space. *Advances in Computational Mathematics*, 20:263–271, 2004.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - -Support Vector Learning*, pages 185–208, Cambridge, MA, 1999. MIT Press.

T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19:201–209, 1975.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78 (9), September 1990.

W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69:1–24, 1997.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C. The Art of Scientific Computation*. Cambridge University Press, 1994.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, February 1989.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

G. Rätsch. *Robust Boosting via Convex Optimization: Theory and Applications*. PhD thesis, University of Potsdam, 2001.

G. Rätsch, T. Onoda, and K. R. Müller. Soft margins for adaboost. *Machine Learning*, 42(3): 287–320, 2001.

G. Rätsch, S. Mika, and A. J. Smola. Adapting codes and embeddings for polychotomies. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 513–520. MIT Press, Cambridge, MA, 2003.

A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10:441–451, 1959.

R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, 1970.

S. Romdhani, P. H. S. Torr, B. Schölkopf, and A. Blake. Efficient face detection by a cascaded support vector machine expansion. *Proceedings of the Royal Society A*, 460:3283–3297, 2004.

W. Rudin. *Fourier analysis on groups*. Interscience Publishers, New York, 1962.

P. Ruján. A fast method for calculating the perceptron with maximal stability. *Journal de Physique I France*, 3:277–290, 1993.

T. W. Sager. An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association*, 74(366):329–339, 1979.

R. Schapire, Y. Freund, P. L. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39: 811–841, 1938.

B. Schölkopf. *Support Vector Learning*. R. Oldenbourg Verlag, Munich, 1997. Download: http://www.kernel-machines.org.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

B. Schölkopf and A. J. Smola. Support vector machines. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1119–1125. MIT Press, 2nd edition, 2003.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 416–426, 2001.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 2004.

R. Schwarz and Y.L. Chow. The n-best algorithm: An efficient and exact procedure for finding the n most likely hypotheses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 81–84, 1990.

F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*, pages 213–220. Association for Computational Linguistics, 2003.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

John Shawe-Taylor, Peter L. Bartlett, R. C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44 (5):1926–1940, 1998.

Y. Shibata. On the tree representation of chordal graphs. *Journal of Graph Theory*, 12(3): 421–428, 1988.

Kristy Sim. Context kernels for text categorization. Master's thesis, The Australian National University, Canberra, Australia, ACT 0200, June 2001.

A. J. Smola. Regression estimation with support vector learning machines. Diplomarbeit, Technische Universität München, 1996.

A. J. Smola and I. R. Kondor. Kernels and regularization on graphs. In B. Schölkopf and M. K. Warmuth, editors, *Proceedings of the Annual Conference on Computational Learning Theory*, Lecture Notes in Computer Science, pages 144–158. Springer, 2003.

A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In P. Langley, editor, *Proceedings of the International Conference on Machine Learning*, pages 911–918, San Francisco, 2000. Morgan Kaufmann Publishers.

A. J. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998.

A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11(5):637–649, 1998.

A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002a.

I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18:768–791, 2002b.

J. Stewart. Positive definite functions and generalizations, an historical survey. *Rocky Mountain J. Math.*, 6:409–434, 1976.

M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, and J. Weston. Support vector regression with ANOVA decomposition kernels. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - -Support Vector Learning*, pages 285–292, Cambridge, MA, 1999. MIT Press.

B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, 2003.

B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning. Max-margin parsing. In *Empirical Methods in Natural Language Processing*, 2004.

D. M. J. Tax and R. P. W. Duin. Data domain description by support vectors. In M. Verleysen, editor, *Proceedings ESANN*, pages 251–256, Brussels, 1999. D Facto.

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML '04: Twenty-first international conference on Machine learning*, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-828-5.

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 2005.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Annals of Statistics*, 25(3): 948–969, 1997.

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32 (1):135–166, 2003.

C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2 edition, 1979.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.

V. Vapnik, S. Golowich, and A. J. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 281–287, Cambridge, MA, 1997. MIT Press.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Berlin, 1982.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

V. N. Vapnik and A. Ya. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3): 283–305, 1991.

S. V. N. Vishwanathan and A. J. Smola. Fast kernels for string and tree matching. In K. Tsuda, B. Schölkopf, and J.P. Vert, editors, *Kernels and Bioinformatics*, Cambridge, MA, 2004a. MIT Press. URL `http://users.rsise.anu.edu.au/ vishy/papers/VisSmo04.pdf`.

S. V. N. Vishwanathan and A. J. Smola. Binet-Cauchy kernels. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, 2004b. URL `http://users.rsise.anu.edu.au/ vishy/papers/VisSmo04a.pdf`.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Annals of Statistics*, 23:1865–1895, 1995.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Department of Statistics, September 2003a.

Martin J. Wainwright and Michael I. Jordan. Semidefinite relaxations for approximate inference on graphs with cycles. Technical Report UCB/CSD-3-1226, University of California, Berkeley, 2003b.

C. Watkins. Dynamic alignment kernels. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 39–50, Cambridge, MA, 2000. MIT Press.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik. Kernel dependency estimation. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 873–880. MIT Press, Cambridge, MA, 2003.

C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning and Inference in Graphical Models*, pages 599–621. Kluwer Academic, 1998.

C. K. I. Williams and M. Seeger. Using the Nystrom method to speed up kernel machines. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, 2000.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization bounds for regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transaction on Information Theory*, 47(6):2516–2532, 2001.

L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Jounal of Machine Learning Research*, 4:913–931, 2003.

P. Wolfe. A duality theorem for nonlinear programming. *Quarterly of Applied Mathematics*, 19: 239–244, 1961.

J.S. Yedidia, W.T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, chapter 8, pages 239–236. Science & Technology Books, 2003.

L. S. Zettlemoyer and M. Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Uncertainty in Artificial Intelligence UAI*, 2005.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.

D. Zhou, J. Huang, and B. Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *International Conference on Machine Learning*, 2005.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *Bioinformatics*, 16(9): 799–807, 2000.