Learning Bounds for Support Vector Machines with Learned Kernels

Nathan Srebro[1] and Shai Ben-David[2]

[1] University of Toronto Department of Computer Science, Toronto ON, CANADA
[2] University of Waterloo School of Computer Science, Waterloo ON, CANADA
nati@cs.toronto.edu, shai@cs.uwaterloo.ca

Abstract. Consider the problem of learning a kernel for use in SVM classification. We bound the estimation error of a large margin classifier when the kernel, relative to which this margin is defined, is chosen from a family of kernels based on the training sample. For a kernel family with *pseudodimension* d_ϕ, we present a bound of $\sqrt{\tilde{\mathcal{O}}(d_\phi + 1/\gamma^2)/n}$ on the estimation error for SVMs with margin γ. This is the first bound in which the relation between the margin term and the family-of-kernels term is **additive** rather then multiplicative. The pseudodimension of families of linear combinations of base kernels is the number of base kernels. Unlike in previous (multiplicative) bounds, there is no non-negativity requirement on the coefficients of the linear combinations. We also give simple bounds on the pseudodimension for families of Gaussian kernels.

1 Introduction

In support vector machines (SVMs), as well as other similar methods, prior knowledge is represented through a *kernel function* specifying the inner products between an implicit representation of input points in some Hilbert space. A large margin linear classifier is then sought in this implicit Hilbert space. Using a "good" kernel function, appropriate for the problem, is crucial for successful learning: The kernel function essentially specifies the permitted hypothesis class, or at least which hypotheses are preferred.

In the standard SVM framework, one commits to a fixed kernel function apriori, and then searches for a large margin classifier with respect to this kernel. If it turns out that this fixed kernel in inappropriate for the data, it might be impossible to find a good large margin classifier. Instead, one can search for a data-appropriate kernel function, from some class of allowed kernels, permitting large margin classification. That is, search for both a kernel *and* a large margin classifier with respect to the kernel. In this paper we develop bounds for the sample complexity cost of allowing such kernel adaptation.

1.1 Learning the Kernel

As in standard hypothesis learning, the process of learning a kernel is guided by some family of potential kernels. A popular type of kernel family consists of

kernels that are a linear, or convex, combinations of several base kernels [1–3][3]:

$$\mathcal{K}_{\text{linear}}(K_1,\ldots,K_k) \stackrel{\text{def}}{=} \left\{ K_{\boldsymbol{\lambda}} = \sum_{i=1}^{k} \lambda_i K_i \mid K_{\boldsymbol{\lambda}} \succcurlyeq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\} \quad (1)$$

$$\mathcal{K}_{\text{convex}}(K_1,\ldots,K_k) \stackrel{\text{def}}{=} \left\{ K_{\boldsymbol{\lambda}} = \sum_{i=1}^{k} \lambda_i K_i \mid \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\} \quad (2)$$

Such kernel families are useful for integrating several sources of information, each encoded in a different kernel, and are especially popular in bioninformatics applications [4–6, and others].

Another common approach is learning (or "tuning") parameters of a parameterized kernel, such as the covariance matrix of a Gaussian kernel, based on training data [7–10, and others]. This amounts to learning a kernel from a parametric family, such as the family of Gaussian kernels:

$$\mathcal{K}^{\ell}_{\text{Gaussian}} \stackrel{\text{def}}{=} \left\{ K_A : (x_1,x_2) \mapsto e^{-(x_1-x_2)'A(x_1-x_2)} \mid A \in \mathbb{R}^{\ell\times\ell},\ A \succcurlyeq 0 \right\} \quad (3)$$

Infinite-dimensional kernel families have also been considered, either through *hyperkernels* [11] or as convex combinations of a continuum of base kernels (e.g. convex combinations of Gaussian kernels) [12, 13]. In this paper we focus on finite-dimensional kernel families, such as those defined by equations (1)–(3).

Learning the kernel matrix allows for greater flexibility in matching the target function, but this of course comes at the cost of higher estimation error, i.e. a looser bound on the expected error of the learned classifier in terms of its empirical error. Bounding this estimation gap is essential for building theoretical support for kernel learning, and this is the focus of this paper.

1.2 Learning Bounds with Learned Kernels—Previous Work

For standard SVM learning, with a fixed kernel, one can show that, with high probability, the estimation error (gap between the expected error and empirical error) of a learned classifier with margin γ is bounded by $\sqrt{\tilde{\mathcal{O}}(1/\gamma^2)/n}$ where n is the sample size and the $\tilde{\mathcal{O}}()$ notation hides logarithmic factors in its argument, the sample size and the allowed failure probability. That is, the number of samples needed for learning is $\tilde{\mathcal{O}}(1/\gamma^2)$.

Lanckriet *et al.* [1] showed that when a kernel is chosen from a convex combination of k base kernels, the estimation error of the learned classifier is bounded by $\sqrt{\tilde{\mathcal{O}}(k/\gamma^2)/n}$ where γ is the margin of the learned classifier under the learned kernel. Note the multiplicative interaction between the margin complexity term $1/\gamma^2$ and the number of base kernels k. Recently, Micchelli *et al.* [14] derived bounds for the family of Gaussian kernels of equation (3). The dependence of

[3] Lanckriet *et al.* [1] impose a bound on the trace of the Gram matrix of K_λ—this is equivalent to bounding $\sum \lambda_i$ when the base kernels are normalized.

these bounds on the margin and the complexity of the kernel family is also multiplicative—the estimation error is bounded by $\sqrt{\tilde{\mathcal{O}}(C_\ell/\gamma^2)/n}$, where C_ℓ is a constant that depends on the input dimensionality ℓ.

The multiplicative interaction between the margin and the complexity measure of the kernel class is disappointing. It suggests that learning even a few kernel parameters (e.g. the coefficients λ) leads to a multiplicative increase in the required sample size. It is important to understand whether such a multiplicative increase in the number of training samples is in fact necessary.

Bousquet and Herrmann [2, Theorem 2] and Lanckriet *et al.* [1] also discuss bounds for families of convex and linear combinations of kernels that appear to be independent of the number of base kernels. However, we show in the Appendix that these bounds are meaningless: The bound on the expected error is never less than one. We are not aware of any previous work describing meaningful explicit bounds for the family of linear combinations of kernels given in equation (1).

1.3 New, Additive, Learning Bounds

In this paper, we bound the estimation error, when the kernel is chosen from a kernel family $\mathcal{K}$, by $\sqrt{\tilde{\mathcal{O}}(d_\phi + 1/\gamma^2)/n}$, where d_ϕ is the *pseudodimension* of the family $\mathcal{K}$ (Theorem 2; the pseudodimension is defined in Definition 5). This establishes that the bound on the required sample size, $\tilde{\mathcal{O}}\left(d_\phi + 1/\gamma^2\right)$ grows only **additively** with the dimensionality of the allowed kernel family (up to logarithmic factors). This is a much more reasonable price to pay for not committing to a single kernel apriori.

The pseudodimension of most kernel families matches our intuitive notion of the dimensionality of the family, and in particular:

- The pseudodimension of a family of linear, or convex, combinations of k base kernels (equations 1,2) is at most k (Lemma 7).
- The pseudodimension of the family $\mathcal{K}^\ell_{\text{Gaussian}}$ of Gaussian kernels (equation 3) for inputs $x \in \mathbb{R}^\ell$, is at most $\ell(\ell+1)/2$ (Lemma 9). If only diagonal covariances are allowed, the pseudodimension is ℓ (Lemma 10). If the covariances (and therefore A) are constrained to be of rank at most k, the pseudodimension is at most $k\ell \log_2(22k\ell)$ (Lemma 11).

1.4 Plan of Attack

For a fixed kernel, it is well known that, with probability at least $1-\delta$, the estimation error of all margin-γ classifiers is at most $\sqrt{\mathcal{O}(1/\gamma^2 - \log\delta)/n}$ [15]. To obtain a bound that holds for all margin-γ classifiers with respect to *any* kernel K in some *finite* kernel family $\mathcal{K}$, consider a union bound over the $|\mathcal{K}|$ events "the estimation error is large for some margin-γ classifier with respect to K" for each $K \in \mathcal{K}$. Using the above bound with δ scaled by the cardinality $|\mathcal{K}|$, the union bound ensures us that with probability at least $1-\delta$, the estimation

error will be bounded by $\sqrt{\mathcal{O}(\log|\mathcal{K}| + 1/\gamma^2 - \log\delta)/n}$ for all margin-γ classifiers with respect to any kernel in the family.

In order to extend this type of result also to infinite-cardinality families, we employ the standard notion of ϵ-nets: Roughly speaking, even though a continuous family $\mathcal{K}$ might be infinite, many kernels in it will be very similar and it will not matter which one we use. Instead of taking a union bound over all kernels in $\mathcal{K}$, we only take a union bound over "essentially different" kernels. In Section 4 we use standard results to show that the number of "essentially different" kernels in a family grows exponentially only with the dimensionality of the family, yielding an additive term (almost) proportional to the dimensionality.

As is standard in obtaining such bounds, our notion of "essentially different" refers to a specific sample and so symmetrization arguments are required in order to make the above conceptual arguments concrete. To do so cleanly and cheaply, we use an ϵ-net of *kernels* to construct an ϵ-net of *classifiers* with respect to the kernels, noting that the size of the ϵ-net increases only multiplicatively relative to the size of an ϵ-net for any one kernel (Section 3). An important component of this construction is the observation that kernels that are close as real-valued functions also yield similar classes of classifiers (Lemma 2). Using our constructed ϵ-net, we can apply standard results bounding the estimation error in terms of the log-size of ϵ-nets, without needing to invoke symmetrization arguments directly.

For the sake of simplicity and conciseness of presentation, the results in this paper are stated for binary classification using a homogeneous large-margin classifier, i.e. not allowing a bias term, and refer to zero-one error. The results can be easily extended to other loss functions and to allow a bias term.

2 Preliminaries

Notation: We use $\|v\|$ to denote the norm of a vector in an abstract Hilbert space. For a vector $v \in \mathbb{R}^n$, $\|v\|$ is the Euclidean norm of v. For a matrix $A \in \mathbb{R}^{n\times n}$, $\|A\|_2 = \max_{\|v\|=1}\|Av\|$ is the L_2 operator norm of A, $|A|_\infty = \max_{ij}|A_{ij}|$ is the l_∞ norm of A and $A \succcurlyeq 0$ indicates that A is positive semi-definite (p.s.d.) and symmetric. We use boldface $\mathbf{x}$ for samples (multisets, though we refer to them simply as sets) of points, where $|\mathbf{x}|$ is the number of points in a sample.

2.1 Support Vector Machines

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a training set of n pairs of input points $x_i \in \mathcal{X}$ and target labels $y_i \in \{\pm 1\}$. Let $\phi : \mathcal{X} \to \mathcal{H}$ be a mapping of input points into a Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$. A vector $w \in \mathcal{H}$ can be used as a predictor for points in $\mathcal{X}$, predicting the label $\mathrm{sign}(\langle w, \phi(x)\rangle)$ for input x. Consider learning by seeking a unit-norm predictor w achieving low empirical hinge loss $\hat{h}^\gamma(w) = \frac{1}{n}\sum_{i=1}^n \max(\gamma - y_i\langle w, \phi(x_i)\rangle, 0)$, relative to a margin $\gamma > 0$.

The Representer Theorem [16, Theorem 4.2] guarantees that the predictor w minimizing $\hat{h}^\gamma(w)$ can be written as $w = \sum_{i=1}^n \alpha_i\phi(x_i)$. For such w, predictions

$\langle w, \phi(x)\rangle = \sum_i \alpha_i \langle \phi(x_i), \phi(x)\rangle$ and the norm $\|w\|^2 = \sum_{ij} \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j)\rangle$ depend only on inner products between mappings of input points. The Hilbert space $\mathcal{H}$ and mapping ϕ can therefore be represented implicitly by a *kernel function* $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ specifying these inner products: $K(x^{\heartsuit}, x^{\spadesuit}) = \langle \phi(x^{\heartsuit}), \phi(x^{\spadesuit})\rangle$.

Definition 1. *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a* **kernel function** *if for some Hilbert space $\mathcal{H}$ and mapping $\phi : \mathcal{X} \to \mathcal{H}$, $K(x^{\heartsuit}, x^{\spadesuit}) = \langle \phi(x^{\heartsuit}), \phi(x^{\spadesuit})\rangle$ for all $x^{\heartsuit}, x^{\spadesuit}$.*

For a set $\mathbf{x} = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ of points, it will be useful to consider their Gram matrix $K_{\mathbf{x}} \in \mathbb{R}^{n \times n}$, $K_{\mathbf{x}}[i, j] = K(x_i, x_j)$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function iff for any finite $\mathbf{x} \subset \mathcal{X}$, the Gram matrix $K_{\mathbf{x}}$ is p.s.d [16].

When specifying the mapping ϕ implicitly through a kernel function, it is useful to think about a predictor as a function $f : \mathcal{X} \to \mathbb{R}$ instead of considering w explicitly. Given a kernel K, learning can then be phrased as choosing a predictor from the class

$$\mathcal{F}_K \stackrel{\text{def}}{=} \{x \mapsto \langle w, \phi(x)\rangle \mid \|w\| \leq 1 \ , \ K(x^{\heartsuit}, x^{\spadesuit}) = \langle \phi(x^{\heartsuit}), \phi(x^{\spadesuit})\rangle\} \tag{4}$$

minimizing

$$\hat{h}^{\gamma}(f) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \max(\gamma - y_i f(x_i), 0). \tag{5}$$

For a set of points $\mathbf{x} = \{x_1, \ldots, x_n\}$, let $f(\mathbf{x}) \in \mathbb{R}^n$ be the vector whose entries are $f(x_i)$. The following restricted variant of the Representer Theorem characterizes the possible prediction vectors $f(\mathbf{x})$ by suggesting the matrix square root of the Gram matrix ($K_{\mathbf{x}}^{1/2} \succcurlyeq 0$ such that $K_{\mathbf{x}} = K_{\mathbf{x}}^{1/2} K_{\mathbf{x}}^{1/2}$) as a possible "feature mapping" for points in $\mathbf{x}$:

Lemma 1. *For any kernel function K and set $\mathbf{x} = \{x_1, \ldots, x_n\}$ of n points:*

$$\{f(\mathbf{x}) \mid f \in \mathcal{F}_K\} = \{K_{\mathbf{x}}^{1/2} \tilde{w} \mid \tilde{w} \in \mathbb{R}^n, \|\tilde{w}\| \leq 1\},$$

Proof. For any $f \in \mathcal{F}_K$ we can write $f(x) = \langle w, \phi(x)\rangle$ with $\|w\| \leq 1$ (equation 4). Consider the projection $w_{\|} = \sum_i \alpha_i \phi(x_i)$ of w onto $\text{span}(\phi(x_1), \ldots, \phi(x_n))$. We have $f(x_i) = \langle w, \phi(x_i)\rangle = \langle w_{\|}, \phi(x_i)\rangle = \sum_j \alpha_j K(x_j, x_i)$ and $1 \geq \|w\|^2 \geq \|w_{\|}\|^2 = \sum_{ij} \alpha_i \alpha_j K(x_i, x_j)$. In matrix form: $f(\mathbf{x}) = K_{\mathbf{x}} \alpha$ and $\alpha' K_{\mathbf{x}} \alpha \leq 1$. Setting $\tilde{w} = K_{\mathbf{x}}^{1/2} \alpha$ we have $f(\mathbf{x}) = K_{\mathbf{x}} \alpha = K_{\mathbf{x}}^{1/2} K_{\mathbf{x}}^{1/2} \alpha = K_{\mathbf{x}}^{1/2} \tilde{w}$ while $\|\tilde{w}\|^2 = \alpha' K_{\mathbf{x}}^{1/2} K_{\mathbf{x}}^{1/2} \alpha = \alpha K_{\mathbf{x}} \alpha \leq 1$. This establishes that the left-hand side is a subset of the right-hand side.

For any $\tilde{w} \in \mathbb{R}^n$ with $\|\tilde{w}\| \leq 1$ we would like to define $w = \sum_i \alpha_i \phi(x_i)$ with $\alpha = K_{\mathbf{x}}^{-1/2} \tilde{w}$ and get $\langle w, \phi(x_i)\rangle = \sum_j \alpha_j \langle \phi(x_j), \phi(x_i)\rangle = K_{\mathbf{x}} \alpha = K_{\mathbf{x}} K_{\mathbf{x}}^{-1/2} \tilde{w} = K_{\mathbf{x}}^{1/2} \tilde{w}$. However, $K_{\mathbf{x}}$ might be singular. Instead, consider the singular value decomposition $K_{\mathbf{x}} = USU'$, with $U'U = I$, where zero singular values have been removed, i.e. S is an all-positive diagonal matrix and U might be rectangular. Set $\alpha = US^{-1/2} U' \tilde{w}$ and consider $w = \sum_i \alpha_i \phi(x_i)$. We can now calculate:

$$\begin{aligned} \langle w, \phi(x_i)\rangle = \sum_j \alpha_j \langle \phi(x_j), \phi(x_i)\rangle = K_{\mathbf{x}} \alpha \\ = USU' \cdot US^{-1/2} U' \tilde{w} = US^{1/2} U' \tilde{w} = K_{\mathbf{x}}^{1/2} \tilde{w} \end{aligned} \tag{6}$$

while $\|w\|^2 = \alpha' K \alpha = \tilde{w}' U S^{-1/2} U' \cdot U S U' \cdot U S^{-1/2} U' \tilde{w} = \tilde{w}' U U' \tilde{w} \leq \|\tilde{w}\|^2 \leq 1$ $\square$

To remove confusion we note some differences between the presentation here and other common, and equivalent, presentations of SVMs. Instead of fixing the margin γ and minimizing the empirical hinge loss, it is common to try to maximize γ while minimizing the loss. The most common combined objective, in our notation, is to minimize $\frac{1}{\gamma^2} + C \cdot \frac{1}{\gamma} \hat{h}^\gamma(w)$ for some trade-off parameter C. This is usually done with a change of variable to $\tilde{w} = w/\gamma$, which results in an equivalent problem where the margin is fixed to one, and the norm of $\tilde{w}$ varies. Expressed in terms of $\tilde{w}$ the objective is $\|\tilde{w}\|^2 + C \cdot \hat{h}^1(\tilde{w})$. Varying the trade-off parameter C is equivalent to varying the margin and minimizing the loss. The variant of the Representer Theorem given in Lemma 1 applies to *any* predictor in $\mathcal{F}_K$, but only describes the behavior of the predictor on the set $\mathbf{x}$. This will be sufficient for our purposes.

2.2 Learning Bounds and Covering Numbers

We derive generalization error bounds in the standard agnostic learning setting. That is, we assume data is generated by some unknown joint distribution $P(X, Y)$ over input points in $\mathcal{X}$ and labels in ± 1. The training set consists of n i.i.d. samples (x_i, y_i) from this joint distribution. We would like to bound the difference $\mathrm{est}^\gamma(f) = \mathrm{err}(f) - \widehat{\mathrm{err}}^\gamma(f)$ (the *estimation error*) between the expected error rate

$$\mathrm{err}(f) = \Pr_{X,Y}(Yf(X) \leq 0), \tag{7}$$

and the empirical *margin* error rate

$$\widehat{\mathrm{err}}^\gamma(f) = \frac{|\{i | y_i f(x_i) < \gamma\}|}{n}. \tag{8}$$

The main challenge of deriving such bounds is bounding the estimation error *uniformly* over all predictors in a class. The technique we employ in this paper to obtain such uniform bounds is bounding the covering numbers of classes.

Definition 2. *A subset $\tilde{A} \subset A$ is an ϵ-net of A under the metric d if for any $a \in A$ there exists $\tilde{a} \in \tilde{A}$ with $d(a, \tilde{a}) \leq \epsilon$. The* **covering number** *$\mathcal{N}_d(A, \epsilon)$ is the size of the smallest ϵ-net of A.*

We will study coverings of classes of predictors under the sample-based l_∞ metric, which depends on a sample $\mathbf{x} = \{x_1, \ldots, x_n\}$:

$$d_\infty^{\mathbf{x}}(f_1, f_2) = \max_{i=1}^{n} |f_1(x_i) - f_2(x_i)| \tag{9}$$

Definition 3. *The* **uniform l_∞ covering number** *$\mathcal{N}_n(\mathcal{F}, \epsilon)$ of a predictor class $\mathcal{F}$ is given by considering all possible samples $\mathbf{x}$ of size n:*

$$\mathcal{N}_n(\mathcal{F}, \epsilon) = \sup_{|\mathbf{x}|=n} \mathcal{N}_{d_\infty^{\mathbf{x}}}(\mathcal{F}, \epsilon)$$

The uniform l_∞ covering number can be used to bound the estimation error uniformly. For a predictor class $\mathcal{F}$ and fixed $\gamma > 0$, with probability at least $1-\delta$ over the choice of a training set of size n [17, Theorem 10.1]:

$$\sup_{f\in\mathcal{F}} \mathrm{est}^\gamma(f) \leq \sqrt{8\frac{1+\log\mathcal{N}_{2n}(\mathcal{F},\gamma/2)-\log\delta}{n}} \tag{10}$$

The uniform covering number of the class $\mathcal{F}_K$ (unit-norm predictors corresponding to a kernel function K; recall eq. (4)), with $K(x,x) \leq B$ for all x, can be bounded by applying Theorems 14.21 and 12.8 of Anthony and Bartlett [17]:

$$\mathcal{N}_n(\mathcal{F},\epsilon) \leq 2\left(\frac{4nB}{\epsilon^2}\right)^{\frac{16B}{\epsilon^2}\log_2\left(\frac{\epsilon e n}{4\sqrt{B}}\right)} \tag{11}$$

yielding $\sup_{f\in\mathcal{F}_K} \mathrm{est}^\gamma(f) = \sqrt{\tilde{\mathcal{O}}(B/\gamma^2)/n}$ and implying that $\tilde{\mathcal{O}}\big(B/\gamma^2\big)$ training examples are enough to guarantee that the estimation error diminishes.

2.3 Learning the Kernel

Instead of committing to a fixed kernel, we consider a family $\mathcal{K} \subseteq \{K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}\}$ of allowed kernels and the corresponding predictor class:

$$\mathcal{F}_\mathcal{K} = \cup_{K\in\mathcal{K}}\mathcal{F}_K \tag{12}$$

The learning problem is now one of minimizing $\hat{h}^\gamma(f)$ for $f \in \mathcal{F}_\mathcal{K}$. We are interested in bounding the estimation error uniformly for the class $\mathcal{F}_\mathcal{K}$ and will do so by bounding the covering numbers of the class. The bounds will depend on the "dimensionality" of $\mathcal{K}$, which we will define later, the margin γ, and a bound B such that $K(x,x) \leq B$ for all $K \in \mathcal{K}$ and all x. We will say that such a kernel family is *bounded by* B. Note that $\sqrt{B}$ is the radius of a ball (around the origin) containing $\phi(x)$ in the implied Hilbert space, and scaling ϕ scales both $\sqrt{B}$ and γ linearly. Our bounds will therefore depend on the *relative margin* $\gamma/\sqrt{B}$.

3 Covering Numbers with Multiple Kernels

In this section, we will show how to use bounds on covering numbers of a family $\mathcal{K}$ of kernels to obtain bounds on the covering number of the class $\mathcal{F}_\mathcal{K}$ of predictors that are low-norm linear predictors under some kernel $K \in \mathcal{K}$. We will show how to combine an ϵ-net of $\mathcal{K}$ with ϵ-nets for the classes $\mathcal{F}_K$ to obtain an ϵ-net for the class $\mathcal{F}_\mathcal{K}$. In the next section, we will see how to bound the covering numbers of a kernel family $\mathcal{K}$ and will then be able to apply the main result of this section to get a bound on the covering number of $\mathcal{F}_\mathcal{K}$.

In order to state the main result of this section, we will need to consider covering numbers of kernel families. We will use the following sample-based metric between kernels. For a sample $\mathbf{x} = \{x_1,\ldots,x_n\}$:

$$D^\mathbf{x}_\infty(K,\tilde{K}) \stackrel{\mathrm{def}}{=} \max_{i,j=1}^n |K(x_i,x_j) - \tilde{K}(x_i,x_j)| = \left|K_\mathbf{x} - \tilde{K}_\mathbf{x}\right|_\infty \tag{13}$$

Definition 4. *The* **uniform l_∞ kernel covering number** $\mathcal{N}_n^D(\mathcal{K},\epsilon)$ *of a kernel class $\mathcal{K}$ is given by considering all possible samples* $\mathbf{x}$ *of size n:*

$$\mathcal{N}_n^D(\mathcal{K},\epsilon) = \sup_{|\mathbf{x}|=n} \mathcal{N}_{D_\infty^{\mathbf{x}}}(\mathcal{K},\epsilon)$$

Theorem 1. *For a family $\mathcal{K}$ of kernels bounded by B and any $\epsilon < 1$:*

$$\mathcal{N}_n(\mathcal{F}_\mathcal{K},\epsilon) \le 2\cdot\mathcal{N}_n^D(\mathcal{K},\tfrac{\epsilon^2}{4n})\cdot\left(\tfrac{16nB}{\epsilon^2}\right)^{\frac{64B}{\epsilon^2}\log\left(\frac{\epsilon en}{8\sqrt{B}}\right)}$$

In order to prove Theorem 1, we will first show how all the predictors of one kernel can be approximated by predictors of a nearby kernel. Roughly speaking, we do so by showing that the possible "feature mapping" $K_{\mathbf{x}}^{1/2}$ of Lemma 1 does not change too much:

Lemma 2. *Let $K,\tilde{K}$ be two kernel functions. Then for any predictor $f\in\mathcal{F}_K$ there exists a predictor $\tilde{f}\in\mathcal{F}_{\tilde{K}}$ with $d_\infty^{\mathbf{x}}(f,\tilde{f}) \le \sqrt{nD_\infty^{\mathbf{x}}(K,\tilde{K})}$.*

Proof. Let $w\in\mathbb{R}^n$, $\|w\|=1$ such that $f(\mathbf{x}) = K_{\mathbf{x}}^{1/2}w$, as guaranteed by Lemma 1. Consider the predictor $\tilde{f}\in\mathcal{F}_{\tilde{K}}$ such that $\tilde{f}(\mathbf{x}) = \tilde{K}_{\mathbf{x}}^{1/2}w$, guaranteed by the reverse direction of Lemma 1:

$$d_\infty^{\mathbf{x}}(f,\tilde{f}) = \max_i\left|f(x_i)-\tilde{f}(x_i)\right| \le \left\|f(\mathbf{x})-\tilde{f}(\mathbf{x})\right\| \tag{14}$$

$$= \left\|K_{\mathbf{x}}^{1/2}w - \tilde{K}_{\mathbf{x}}^{1/2}w\right\| \le \left\|K_{\mathbf{x}}^{1/2}-\tilde{K}_{\mathbf{x}}^{1/2}\right\|_2\|w\| \le \sqrt{\left\|K_{\mathbf{x}}-\tilde{K}_{\mathbf{x}}\right\|_2}\cdot 1 \tag{15}$$

$$\le \sqrt{n\left|K_{\mathbf{x}}-\tilde{K}_{\mathbf{x}}\right|_\infty} = \sqrt{nD_\infty^{\mathbf{x}}(K,\tilde{K})} \tag{16}$$

See, e.g., Theorem X.1.1 of Bhatia [18] for the third inequality in (15). □

Proof of Theorem 1: Set $\epsilon_{\mathrm{K}} = \frac{\epsilon^2}{4n}$ and $\epsilon_{\mathrm{F}} = \epsilon/2$. Let $\tilde{\mathcal{K}}$ be an ϵ_{K}-net of $\mathcal{K}$. For each $\tilde{K}\in\tilde{\mathcal{K}}$, let $\tilde{\mathcal{F}}_{\tilde{K}}$ be an ϵ_{F}-net of $\mathcal{F}_{\tilde{K}}$. We will show that

$$\widetilde{\mathcal{F}_\mathcal{K}} \stackrel{\text{def}}{=} \cup_{\tilde{K}\in\tilde{\mathcal{K}}}\tilde{\mathcal{F}}_{\tilde{K}} \tag{17}$$

is an ϵ-net of $\mathcal{F}_\mathcal{K}$. For any $f\in\mathcal{F}_\mathcal{K}$ we have $f\in\mathcal{F}_K$ for some $K\in\mathcal{K}$. The kernel K is covered by some $\tilde{K}\in\tilde{\mathcal{K}}$ with $D_\infty^{\mathbf{x}}(K,\tilde{K})\le\epsilon_{\mathrm{K}}$. Let $\tilde{f}\in\mathcal{F}_{\tilde{K}}$ be a predictor with $d_\infty^{\mathbf{x}}(f,\tilde{f}) \le \sqrt{nD_\infty^{\mathbf{x}}(K,\tilde{K})} \le \sqrt{n\epsilon_{\mathrm{K}}}$ guaranteed by Lemma 2, and $\tilde{\tilde{f}}\in\tilde{\mathcal{F}}_{\tilde{K}}$ such that $d_\infty^{\mathbf{x}}(\tilde{f},\tilde{\tilde{f}})\le\epsilon_{\mathrm{F}}$. Then $\tilde{\tilde{f}}\in\widetilde{\mathcal{F}_\mathcal{K}}$ is a predictor with:

$$d_\infty^{\mathbf{x}}(f,\tilde{\tilde{f}}) \le d_\infty^{\mathbf{x}}(f,\tilde{f}) + d_\infty^{\mathbf{x}}(\tilde{f},\tilde{\tilde{f}}) \le \sqrt{n\epsilon_{\mathrm{K}}}+\epsilon_{\mathrm{F}} = \epsilon \tag{18}$$

This establishes that $\widetilde{\mathcal{F}_\mathcal{K}}$ is indeed an ϵ-net. Its size is bounded by

$$\left|\widetilde{\mathcal{F}_\mathcal{K}}\right| \le \sum_{\tilde{K}\in\tilde{\mathcal{K}}}\left|\tilde{\mathcal{F}}_{\tilde{K}}\right| \le \left|\tilde{\mathcal{K}}\right|\cdot\max_K\left|\tilde{\mathcal{F}}_{\tilde{K}}\right| \le \mathcal{N}_n^D(\mathcal{K},\tfrac{\epsilon^2}{4n})\cdot\max_K\mathcal{N}_n(\mathcal{F}_K,\epsilon/2). \tag{19}$$

Substituting in (11) yields the desired bound. □

4 Learning Bounds in terms of the Pseudodimension

We saw that if we could bound the covering numbers of a kernel family $\mathcal{K}$, we could use Theorem 1 to obtain a bound on the covering numbers of the class $\mathcal{F}_{\mathcal{K}}$ of predictors that are low-norm linear predictors under some kernel $K \in \mathcal{K}$. We could then use (10) to establish a learning bound. In this section, we will see how to bound the covering numbers of a kernel family by its *pseudodimension*, and use this to state learning bounds in terms of this measure. To do so, we will use well-known results bounding covering numbers in terms of the pseudodimension, paying a bit of attention to the subtleties of the differences between Definition 4 of uniform *kernel* covering numbers, and the standard Definition 3 of uniform covering numbers.

To define the pseudodimension of a kernel family we will treat kernels as functions from pairs of points to the reals:

Definition 5. *Let $\mathcal{K} = \{K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}\}$ be a kernel family. The class $\mathcal{K}$* **pseudo-shatters** *a set of n pairs of points $(x_1^{\heartsuit}, x_1^{\spadesuit}), \ldots, (x_n^{\heartsuit}, x_n^{\spadesuit})$ if there exist thresholds $t_1, \ldots, t_n \in \mathbb{R}$ such that for any $b_1, \ldots, b_n \in \{\pm 1\}$ there exists $K \in \mathcal{K}$ with $\operatorname{sign}(K(x_i^{\heartsuit}, x_i^{\spadesuit}) - t_i) = b_i$. The* **pseudodimension** *$d_\phi(\mathcal{K})$ is the largest n such that there exists a set of n pairs of points that are pseudo-shattered by $\mathcal{K}$.*

The uniform l_∞ covering numbers of a class G of real-valued functions taking values in $[-B, B]$ can be bounded in terms of its pseudodimension. Let d_ϕ be the pseudodimension of G; then for any $n > d_\phi$ and $\epsilon > 0$ [17, Theorem 12.2]:

$$\mathcal{N}_n(G, \epsilon) \leq \left(\frac{enB}{\epsilon d_\phi}\right)^{d_\phi} \tag{20}$$

We should be careful here, since the covering numbers $\mathcal{N}_n(\mathcal{K}, \epsilon)$ are in relation to the metrics:

$$d_\infty^{\mathbf{x}^{\heartsuit\spadesuit}}(K, \tilde{K}) = \max_{i=1}^{n} |K(x_i^{\heartsuit}, x_i^{\spadesuit}) - \tilde{K}(x_i^{\heartsuit}, x_i^{\spadesuit})| \tag{21}$$

defined for a sample $\mathbf{x}^{\heartsuit\spadesuit} \subset \mathcal{X} \times \mathcal{X}$ of *pairs* of points $(x_i^{\heartsuit}, x_i^{\spadesuit})$. The supremum in Definition 3 of $\mathcal{N}_n(\mathcal{K}, \epsilon)$ should then be taken over all samples of n *pairs* of points. Compare with (13) where the kernels are evaluated over the n^2 pairs of points (x_i, x_j) arising from a sample of n points.

However, for any sample of n points $\mathbf{x} = \{x_1, \ldots, x_n\} \subset \mathcal{X}$, we can always consider the n^2 point pairs $\mathbf{x}^2 = \{(x_i, x_j) | i, j = 1..n\}$ and observe that $D_\infty^{\mathbf{x}}(K, \tilde{K}) = d_\infty^{\mathbf{x}^2}(K, \tilde{K})$ and so $\mathcal{N}_{D_\infty^{\mathbf{x}}}(\mathcal{K}, \epsilon) = \mathcal{N}_{d_\infty^{\mathbf{x}^2}}(\mathcal{K}, \epsilon)$. Although such sets of point pairs do not account for all sets of n^2 point pairs in the supremum of Definition 3, we can still conclude that for any $\mathcal{K}, n, \epsilon > 0$:

$$\mathcal{N}_n^D(\mathcal{K}, \epsilon) \leq \mathcal{N}_{n^2}(\mathcal{K}, \epsilon) \tag{22}$$

Combining (22) and (20):

Lemma 3. *For any kernel family $\mathcal{K}$ bounded by B with pseudodimension d_ϕ:*

$$\mathcal{N}_n^D(\mathcal{K}, \epsilon) \leq \left(\frac{en^2B}{\epsilon d_\phi}\right)^{d_\phi}$$

Using Lemma 3 and relying on (10) and Theorem 1 we have:

Theorem 2. *For any kernel family $\mathcal{K}$, bounded by B and with pseudodimension d_ϕ, and any fixed $\gamma > 0$, with probability at least $1-\delta$ over the choice of a training set of size n:*

$$\sup_{f \in \mathcal{F}_\mathcal{K}} est^\gamma(f) \leq \sqrt{8\frac{2 + d_\phi \log \frac{128en^3B}{\gamma^2 d_\phi} + 256\frac{B}{\gamma^2}\log\frac{\gamma en}{8\sqrt{B}}\log\frac{128nB}{\gamma^2} - \log\delta}{n}}$$

Theorem 2 is stated for a fixed margin but it can also be stated uniformly over all margins, at the price of an additional $|\log\gamma|$ term (e.g. [15]). Also, instead of bounding $K(x,x)$ for all x, it is enough to bound it only on average, i.e. require $\mathbf{E}[K(X,X)] \leq B$. This corresponds to bounding the trace of the Gram matrix as was done by Lanckriet *et al.*. In any case, we can set $B = 1$ without loss of generality and scale the kernel and margin appropriately. The learning setting investigated here differs slightly from that of Lanckriet *et al.*, who studied transduction, but learning bounds can easily be translated between the two settings.

5 The Pseudodimension of Common Kernel Families

In this section, we analyze the pseudodimension of several kernel families in common use. Most pseudodimension bounds we present follow easily from well-known properties of the pseudodimension of function families, which we review at the beginning of the section. The analyses in this section serve also as examples of how the pseudodimension of other kernel families can be bounded.

5.1 Preliminaries

We review some basic properties of the pseudodimension of a class of functions:

Fact 4 *If $G' \subseteq G$ then $d_\phi(G') \leq d_\phi(G)$.*

Fact 5 ([17, Theorem 11.3]) *Let G be a class of real-valued functions and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ a monotone function. Then $d_\phi(\{\sigma \circ g \mid g \in G\}) \leq d_\phi(G)$.*

Fact 6 ([17, Theorem 11.4]) *The pseudodimension of a k-dimensional vector space of real-valued functions is k.*

We will also use a classic result of Warren that is useful, among other things, for bounding the pseudodimension of classes involving low-rank matrices. We say that the real-valued functions $(g_1, g_2, \ldots, g_m)$ *realize* a sign vector $b \in \{\pm 1\}^m$ iff there exists an input x for which $b_i = \operatorname{sign} g_i(x)$ for all i. The number of sign vectors realizable by m polynomials of degree at most d over $\mathbb{R}^n$, where $m \geq n$, is at most $(4edm/n)^n$ [19].

5.2 Combination of Base Kernels

Since families of linear or convex combinations of k base kernels are subsets of k-dimensional vector spaces of functions, we can easily bound their pseudodimension by k. Note that the pseudodimension depends only on the *number* of base kernels, but does not depend on the particular choice of base kernels.

Lemma 7. *For any finite set of kernels* $S = \{K_1, \dots K_k\}$,

$$d_\phi(\mathcal{K}_{convex}(S)) \leq d_\phi(\mathcal{K}_{linear}(S)) \leq k$$

Proof. We have $\mathcal{K}_{\text{convex}} \subseteq \mathcal{K}_{\text{linear}} \subseteq \operatorname{span} S$ where $\operatorname{span} S = \{\sum_i \lambda_i K_i | \lambda_i \in \mathbb{R}\}$ is a vector space of dimensionality $\leq k$. The bounds follow from Facts 4 and 6. □

5.3 Gaussian Kernels with a Learned Covariance Matrix

Before considering the family $\mathcal{K}_{\text{Gaussian}}$ of Gaussian kernels, let us consider a single-parameter family that generalizes tuning a single scale parameter (i.e. variance) of a Gaussian kernel. For a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, consider the class

$$\mathcal{K}_{\text{scale}}(d) \stackrel{\text{def}}{=} \left\{ K^d_\lambda : (x_1, x_2) \mapsto e^{-\lambda d(x_1,x_2)} \mid \lambda \in \mathbb{R}^+ \right\}. \tag{23}$$

The family of spherical Gaussian kernels is obtained with $d(x_1, x_2) = \|x_1 - x_2\|^2$.

Lemma 8. *For any function* d, $d_\phi(\mathcal{K}_{scale}(d)) \leq 1$.

Proof. The set $\{-\lambda d \mid \lambda \in \mathbb{R}^+\}$ of functions over $\mathcal{X} \times \mathcal{X}$ is a subset of a one-dimensional vector space and so has pseudodimension at most one. Composing them with the monotone exponentiation function and using Fact 5 yields the desired bound. □

In order to analyze the pseudodimension of more general families of Gaussian kernels, we will use the same technique of analyzing the functions in the exponent and then composing them with the exponentiation function. Recall that class $\mathcal{K}^\ell_{\text{Gaussian}}$ of Gaussian kernels over $\mathbb{R}^\ell$ defined in (3).

Lemma 9. $d_\phi(\mathcal{K}^\ell_{Gaussian}) \leq \ell(\ell+1)/2$

Proof. Consider the functions at the exponent: $\{(x_1, x_2) \mapsto -(x_1 - x_2)A(x_1 - x_2) \mid A \in \mathbb{R}^{\ell \times \ell}\ ,\ A \succcurlyeq 0\} \subset \operatorname{span}\{(x_1, x_2) \mapsto (x_1 - x_2)[i] \cdot (x_1 - x_2)[j] \mid i \leq j \leq \ell\}$ where $v[i]$ denotes the i^{th} coordinate of a vector in $\mathbb{R}^\ell$. This is a vector space of dimensionality $\ell(\ell+1)$ and the result follows by composition with the exponentiation function. □

We next analyze the pseudodimension of the family of Gaussian kernels with a diagonal covariance matrix, i.e. when we apply an arbitrary scaling to input coordinates:

$$\mathcal{K}^{(\ell-\text{diag})}_{\text{Gaussian}} = \left\{ K_{\bar\lambda} : (x_1, x_2) \mapsto e^{-(\bar\lambda'(x_1 - x_2))^2} \mid \bar\lambda \in \mathbb{R}^\ell \right\} \tag{24}$$

Lemma 10. $d_\phi(\mathcal{K}^{(\ell-diag)}_{Gaussian}) \leq \ell$

Proof. We use the same arguments. The exponents are spanned by the ℓ functions $(x_1, x_2) \mapsto ((x_1 - x_2)[i])^2$. □

As a final example, we analyze the pseudodimension of the family of Gaussian kernels with a low-rank covariance matrix, corresponding to a low-rank A in our notation:

$$\mathcal{K}^{\ell,k}_{\text{Gaussian}} = \left\{ (x_1, x_2) \mapsto e^{-(x_1-x_2)' A (x_1-x_2)} \mid A \in \mathbb{R}^{\ell\times\ell},\ A \succcurlyeq 0,\ \text{rank}\, A \leq k \right\}$$

This family corresponds to learning a dimensionality reducing linear transformation of the inputs that is applied before calculating the Gaussian kernel.

Lemma 11. $d_\phi(\mathcal{K}^{\ell,k}_{Gaussian}) \leq kl \log_2(8ek\ell)$

Proof. Any $A \succcurlyeq 0$ of rank at most k can be written as $A = U'U$ with $U \in \mathbb{R}^{k\times\ell}$. Consider the set $G = \{(x^\heartsuit, x^\spadesuit) \mapsto -(x^\heartsuit - x^\spadesuit)'U'U(x^\heartsuit - x^\spadesuit) \mid U \in \mathbb{R}^{k\times\ell}\}$ of functions at the exponent. Assume G pseudo-shatters a set of m point pairs $S = \{(x^\heartsuit_1, x^\spadesuit_1) \ldots, (x^\heartsuit_m, x^\spadesuit_m)\}$. By the definition of pseudo-shattering, we get that there exist $t_1, \ldots, t_m \in \mathbb{R}$ so that for every $b \in \{\pm 1\}^m$ there exist $U_b \in \mathbb{R}^{k\times\ell}$ with $b_i = \text{sign}\,(-(x^\heartsuit_i - x^\spadesuit_i)'U'U(x^\heartsuit_i - x^\spadesuit_i) - t_i)$ for all $i \leq m$. Viewing each $p_i(U) \stackrel{\text{def}}{=} -(x^\heartsuit_i - x^\spadesuit_i)'U'U(x^\heartsuit_i - x^\spadesuit_i) - t_i$ as a quadratic polynomial in the $k\ell$ entries of U, where $x^\heartsuit_i - x^\spadesuit_i$ and t_i determine the coefficients of p_i, we get a set of m quadratic polynomials over $k\ell$ variables which realize all 2^m sign vectors. Applying Warren's bound [19] discussed above we get $2^m \leq (8em/k\ell)^{k\ell}$ which implies $m \leq kl \log_2(8ek\ell)$. This is a bound on the number of points that can be pseudo-shattered by G, and hence on the pseudodimension of G, and by composition with exponentiation we get the desired bound. □

6 Conclusion and Discussion

Learning with a *family* of allowed kernel matrices has been a topic of significant interest and the focus of considerable body of research in recent years, and several attempts have been made to establish learning bounds for this setting. In this paper we establish the first generalization error bounds for kernel-learning SVMs where the margin complexity term and the dimensionality of the kernel family interact *additively* rather then *multiplicatively* (up to log factors). The additive interaction yields stronger bounds. We believe that the implied additive bounds on the sample complexity represent its correct behavior (up to log factors), although this remains to be proved.

The results we present significantly improve on previous results for convex combinations of base kernels, for which the only previously known bound had a multiplicative interaction [1], and for Gaussian kernels with a learned covariance matrix, for which only a bound with a multiplicative interaction and an unspecified dependence on the input dimensionality was previously shown [14]. We

also provide the first explicit non-trivial bound for linear combinations of base kernels—a bound that depends only on the (relative) margin and the number of base kernels. The techniques we introduce for obtaining bounds based on the pseudodimension of the class of kernels should readily apply to straightforward derivation of bounds for many other classes.

We note that previous attempts at establishing bounds for this setting [1, 2, 14] relied on bounding the Rademacher complexity [15] of the class $\mathcal{F}_\mathcal{K}$. However, generalization error bounds derived solely from the Rademacher complexity $\boldsymbol{\mathcal{R}}[\mathcal{F}_\mathcal{K}]$ of the class $\mathcal{F}_\mathcal{K}$ *must* have a multiplicative dependence on $\sqrt{B}/\gamma$: The Rademacher complexity $\boldsymbol{\mathcal{R}}[\mathcal{F}_\mathcal{K}]$ scales linearly with the scale $\sqrt{B}$ of functions in $\mathcal{F}_\mathcal{K}$, and to obtain an estimation error bound it is multiplied by the Lipschitz constant $1/\gamma$ [15]. This might be avoidable by clipping predictors in $\mathcal{F}_\mathcal{K}$ to the range $[-\gamma, \gamma]$:

$$\mathcal{F}_\mathcal{K}^\gamma \stackrel{\text{def}}{=} \{f_{[\pm\gamma]} \mid f \in \mathcal{F}_\mathcal{K}\}, \quad f_{[\pm\gamma]}(x) = \begin{cases} \gamma & \text{if } f(x) \geq \gamma \\ f(x) & \text{if } \gamma \geq f(x) \geq -\gamma \\ -\gamma & \text{if } -\gamma \geq f(x) \end{cases} \tag{25}$$

When using the Rademacher complexity $\boldsymbol{\mathcal{R}}[\mathcal{F}_\mathcal{K}]$ to obtain generalization error bounds in terms of the margin error, the class is implicitly clipped and only the Rademacher complexity of $\mathcal{F}_\mathcal{K}^\gamma$ is actually relevant. This Rademacher complexity $\boldsymbol{\mathcal{R}}[\mathcal{F}_\mathcal{K}^\gamma]$ is bounded by $\boldsymbol{\mathcal{R}}[\mathcal{F}_\mathcal{K}]$. In our case, it seems that this last bound is loose. It is possible though, that covering numbers of $\mathcal{K}$ can be used to bound $\boldsymbol{\mathcal{R}}[\mathcal{F}_\mathcal{K}^\gamma]$ by $\mathcal{O}\Big(\gamma \log \mathcal{N}_{2n}^D(\mathcal{K}, 4B/n^2) + \sqrt{B}\Big)/\sqrt{n}$, yielding a generalization error bound with an additive interaction, and perhaps avoiding the log factors of the margin complexity term $\tilde{\mathcal{O}}\big(B/\gamma^2\big)$ of Theorem 2.

References

1. Lanckriet, G.R., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. J Mach Learn Res **5** (2004) 27–72
2. Bousquet, O., Herrmann, D.J.L.: On the complexity of learning the kernel matrix. In: Adv. in Neural Information Processing Systems 15. (2003)
3. Crammer, K., Keshet, J., Singer, Y.: Kernel design using boosting. In: Advances in Neural Information Processing Systems 15. (2003)
4. Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M.I., Noble, W.S.: A statistical framework for genomic data fusion. Bioinformatics **20** (2004)
5. Sonnenburg, S., Rätsch, G., Schafer, C.: Learning interpretable SVMs for biological sequence classification. In: Research in Computational Molecular Biology. (2005)
6. Ben-Hur, A., Noble, W.S.: Kernel methods for predicting protein-protein interactions. Bioinformatics **21** (2005)
7. Cristianini, N., Campbell, C., Shawe-Taylor, J.: Dynamically adapting kernels in support vector machines. In: Adv. in Neural Information Proceedings Systems 11. (1999)
8. Chapelle, O., Vapnik, V., Bousquet, O., Makhuerjee, S.: Choosing multiple parameters for support vector machines. Machine Learning **46** (2002) 131–159

9. Keerthi, S.S.: Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms. IEEE Tran. on Neural Networks **13** (2002) 1225–1229
10. Glasmachers, T., Igel, C.: Gradient-based adaptation of general gaussian kernels. Neural Comput. **17** (2005) 2099–2105
11. Ong, C.S., Smola, A.J., Williamson, R.C.: Learning the kernel with hyperkernels. J. Mach. Learn. Res. **6** (2005)
12. Micchelli, C.A., Pontil, M.: Learning the kernel function via regularization. J. Mach. Learn. Res. **6** (2005)
13. Argyriou, A., Micchelli, C.A., Pontil, M.: Learning convex combinations of continuously parameterized basic kernels. In: 18th Annual Conf. on Learning Theory. (2005)
14. Micchelli, C.A., Pontil, M., Wu, Q., Zhou, D.X.: Error bounds for learning the kernel. Research Note RN/05/09, University College London Dept. of Computer Science (2005)
15. Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. Ann. Statist. **30** (2002)
16. Smola, A.J., Schölkopf, B.: Learning with Kernels. MIT Press (2002)
17. Anthony, M., Bartlett, P.L.: Neural Networks Learning: Theoretical Foundations. Cambridge University Press (1999)
18. Bhatia, R.: Matrix Analysis. Springer (1997)
19. Warren, H.E.: Lower bounds for approximation by nonlinear manifolds. T. Am. Math. Soc. **133** (1968) 167–178

A Analysis of Previous Bounds

We show that some of the previously suggested bounds for SVM kernel learning can never lead to meaningful bounds on the expected error.

Lanckriet *et al.* [1, Theorem 24] show that for any class $\mathcal{K}$ and margin γ, with probability at least $1-\delta$, every $f \in \mathcal{F}_{\mathcal{K}}$ satisfies:

$$\operatorname{err}(f) \leq \widehat{\operatorname{err}}^{\gamma}(f) + \tfrac{1}{\sqrt{n}} \left(4 + \sqrt{2\log(1/\delta)} + \sqrt{\tfrac{\mathcal{C}(\mathcal{K})}{n\gamma^2}} \right) \tag{26}$$

Where $\mathcal{C}(\mathcal{K}) = \mathbf{E}_{\sigma}[\max_{K\in\mathcal{K}} \sigma' K_{\mathbf{x}} \sigma]$, with σ chosen uniformly from $\{\pm 1\}^{2n}$ and $\mathbf{x}$ being a set of n training and n test points. The bound is for a transductive setting and the Gram matrix of both training and test data is considered. We continue denoting the empirical margin error, on the n training points, by $\widehat{\operatorname{err}}^{\gamma}(f)$, but now $\operatorname{err}(f)$ is the test error on the specific n test points.

The expectation $\mathcal{C}(\mathcal{K})$ is not easy to compute in general, and Lanckriet *et al.* provide specific bounds for families of linear, and convex, combinations of base kernels.

A.1 Bound for linear combinations of base kernels

For the family $\mathcal{K} = \mathcal{K}_{\text{linear}}$ of linear combinations of base kernels (equation (1)), Lanckriet *et al.* note that $\mathcal{C}(\mathcal{K}) \leq c \cdot n$, where $c = \max_{K\in\mathcal{K}} \operatorname{tr} K_{\mathbf{x}}$ is an upper

bound on the trace of the possible Gram matrices. Substituting this explicit bound on $\mathcal{C}(\mathcal{K})$ in (26) results in:

$$\mathrm{err}(f) \leq \widehat{\mathrm{err}}^{\gamma}(f) + \frac{1}{\sqrt{n}}\left(4 + \sqrt{2\log(1/\delta)} + \sqrt{\frac{c}{\gamma^2}}\right) \tag{27}$$

However, the following lemma shows that if a kernel allows classifying much of the training points within a large margin, then the trace of its Gram matrix cannot be too small:

Lemma 12. *For all* $f \in \mathcal{F}_K$*:* $\operatorname{tr} K_{\mathbf{x}} \geq \gamma^2(1 - \widehat{err}^{\gamma}(f))n$

Proof. Let $f(x) = \langle w, \phi(x)\rangle$, $\|w\| = 1$. Then for any i for which $y_i f(x_i) = y_i\langle w, \phi(x_i)\rangle \geq \gamma$ we must have $\sqrt{K(x_i,x_i)} = \|\phi(x_i)\| \geq \gamma$. Hence $\operatorname{tr} K_{\mathbf{x}} \geq \sum_{i|y_i f(x_i)\geq\gamma} K(x_i,x_i) \geq |\{i|y_i f(x_i) \geq \gamma\}| \cdot \gamma^2 = (1 - \widehat{\mathrm{err}}^{\gamma}(f))n \cdot \gamma^2$. □

Using Lemma 12 we get that the right-hand side of (27) is at least:

$$\widehat{\mathrm{err}}^{\gamma}(f) + \frac{4+\sqrt{2\log(1/\delta)}}{\sqrt{n}} + \sqrt{\frac{\gamma^2(1-\widehat{\mathrm{err}}^{\gamma}(f))n}{n\gamma^2}} \;>\; \widehat{\mathrm{err}}^{\gamma}(f) + \sqrt{1 - \widehat{\mathrm{err}}^{\gamma}(f)} \;\geq\; 1 \tag{28}$$

A.2 Bound for convex combinations of base kernels

For the family $\mathcal{K} = \mathcal{K}_{\mathrm{convex}}$ of convex combinations of base kernels (equation (2)), Lanckriet *et al.* bound $\mathcal{C}(\mathcal{K}) \leq c \cdot \min\left(m, n\max_{K_i} \frac{\|(K_i)_{\mathbf{x}}\|_2}{\mathrm{tr}((K_i)_{\mathbf{x}})}\right)$, where m is the number of base kernels, $c = \max_{K\in\mathcal{K}} \mathrm{tr}(K_{\mathbf{x}})$ as before, and the maximum is over the base kernels K_i. The first minimization argument yields a non-trivial generalization bound that is multiplicative in the number of base kernels, and is discussed in Section 1.2. The second argument yields the following bound, which was also obtained by Bousquet and Herrmann [2]:

$$\mathrm{err}(f) \leq \widehat{\mathrm{err}}^{\gamma}(f) + \frac{1}{\sqrt{n}}\left(4 + \sqrt{2\log(1/\delta)} + \sqrt{\frac{c\cdot b}{\gamma^2}}\right) \tag{29}$$

where $b = \max_{K_i} \|(K_i)_{\mathbf{x}}\|_2 / \operatorname{tr}(K_i)_{\mathbf{x}}$. This implies $\|K_{\mathbf{x}}\|_2 \leq b \cdot \operatorname{tr} K_{\mathbf{x}} \leq b \cdot c$ for all base kernels and so (by convexity) also for all $K \in \mathcal{K}$. However, similar to the bound on the trace of Gram matrices in Lemma 12, we can also bound the L_2 operator norm required for classification of most points with a margin:

Lemma 13. *For all* $f \in \mathcal{F}_K$*:* $\|K_{\mathbf{x}}\|_2 \geq \gamma^2(1 - \widehat{err}^{\gamma}(f))n$

Proof. From Lemma 1 we have $f(\mathbf{x}) = K_{\mathbf{x}}^{1/2} w$ for some w such that $\|w\| \leq 1$, and so $\|K_{\mathbf{x}}\|_2 = \|K_{\mathbf{x}}^{1/2}\|_2^2 \geq \|K_{\mathbf{x}}^{1/2} w\|^2 = \|f(\mathbf{x})\|^2$. To bound the right-hand side, consider that for $(1-\widehat{\mathrm{err}}^{\gamma}(f))n$ of the points in $\mathbf{x}$ we have $|f(x_i)| = |y_i f(x_i)| \geq \gamma$, and so $\|f(\mathbf{x})\|^2 = \sum_i f(x_i)^2 \geq (1 - \widehat{\mathrm{err}}^{\gamma}(f))n \cdot \gamma^2$. □

Lemma 13 implies $bc \geq \gamma^2(1-\widehat{\mathrm{err}}^{\gamma}(f))n$ and a calculation similar to (28) reveals that the right-hand side of (29) is always greater than one.