ARTICLE Communicated by Vladimir Vapnik

Estimating the Support of a High-Dimensional Distribution

Bernhard Schölkopf
Microsoft Research Ltd, Cambridge CB2 3NH, U.K.

John C. Platt
Microsoft Research, Redmond, WA 98052, U.S.A

John Shawe-Taylor
Royal Holloway, University of London, Egham, Surrey TW20 OEX, U.K.

Alex J. Smola
Robert C. Williamson
Department of Engineering, Australian National University, Canberra 0200, Australia

Suppose you are given some data set drawn from an underlying probability distribution P and you want to estimate a "simple" subset S of input space such that the probability that a test point drawn from P lies outside of S equals some a priori specified value between 0 and 1.

We propose a method to approach this problem by trying to estimate a function f that is positive on S and negative on the complement. The functional form of f is given by a kernel expansion in terms of a potentially small subset of the training data; it is regularized by controlling the length of the weight vector in an associated feature space. The expansion coefficients are found by solving a quadratic programming problem, which we do by carrying out sequential optimization over pairs of input patterns. We also provide a theoretical analysis of the statistical performance of our algorithm.

The algorithm is a natural extension of the support vector algorithm to the case of unlabeled data.

1 Introduction

During recent years, a new set of kernel techniques for supervised learning has been developed (Vapnik, 1995; Schölkopf, Burges, & Smola, 1999). Specifically, support vector (SV) algorithms for pattern recognition, regression estimation, and solution of inverse problems have received considerable attention.

There have been a few attempts to transfer the idea of using kernels to compute inner products in feature spaces to the domain of unsupervised learning. The problems in that domain are, however, less precisely speci-

fied. Generally, they can be characterized as estimating functions of the data, which reveal something interesting about the underlying distributions. For instance, kernel principal component analysis (PCA) can be characterized as computing functions that on the training data produce unit variance outputs while having minimum norm in feature space (Schölkopf, Smola, & Müller, 1999). Another kernel-based unsupervised learning technique, regularized principal manifolds (Smola, Mika, Schölkopf, & Williamson, in press), computes functions that give a mapping onto a lower-dimensional manifold minimizing a regularized quantization error. Clustering algorithms are further examples of unsupervised learning techniques that can be kernelized (Schölkopf, Smola, & Müller, 1999).

An extreme point of view is that unsupervised learning is about estimating densities. Clearly, knowledge of the density of P would then allow us to solve whatever problem can be solved on the basis of the data.

The work presented here addresses an easier problem: it proposes an algorithm that computes a binary function that is supposed to capture regions in input space where the probability density lives (its support), that is, a function such that most of the data will live in the region where the function is nonzero (Schölkopf, Williamson, Smola, Shawe-Taylor, 1999). In doing so, it is in line with Vapnik's principle never to solve a problem that is more general than the one we actually need to solve. Moreover, it is also applicable in cases where the density of the data's distribution is not even well defined, for example, if there are singular components.

After a review of some previous work in section 2, we propose SV algorithms for the considered problem. section 4 gives details on the implementation of the optimization procedure, followed by theoretical results characterizing the present approach. In section 6, we apply the algorithm to artificial as well as real-world data. We conclude with a discussion.

2 Previous Work

In order to describe some previous work, it is convenient to introduce the following definition of a (multidimensional) quantile function, introduced by Einmal and Mason (1992). Let $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ be independently and identically distributed (i.i.d.) random variables in a set $\mathcal{X}$ with distribution P. 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 with respect to $(P, \lambda, \mathcal{C})$ is

$$U(\alpha) = \inf\{\lambda(C): P(C) \geq \alpha, C \in \mathcal{C}\} \qquad 0 < \alpha \leq 1.$$

In the special case where P is the empirical distribution ($P_\ell(C) := \frac{1}{\ell} \sum_{i=1}^{\ell} \mathbf{1}_C(\mathbf{x}_i)$), we obtain the empirical quantile function. We denote by $C(\alpha)$ and $C_\ell(\alpha)$ the (not necessarily unique) $C \in \mathcal{C}$ that attains the infimum (when it is achievable). The most common choice of λ is Lebesgue measure, in which case $C(\alpha)$ is the minimum volume $C \in \mathcal{C}$ that contains at least a fraction α

of the probability mass. Estimators of the form $C_\ell(\alpha)$ are called minimum volume estimators.

Observe that for $\mathcal{C}$ being all Borel measurable sets, $C(1)$ is the support of the density p corresponding to P, assuming it exists. (Note that $C(1)$ is well defined even when p does not exist.) For smaller classes $\mathcal{C}$, $C(1)$ is the minimum volume $C \in \mathcal{C}$ containing the support of p.

Turning to the case where $\alpha < 1$, it seems the first work was reported by Sager (1979) and then Hartigan (1987) who considered $\mathcal{X} = \mathbb{R}^2$ with $\mathcal{C}$ being the class of closed convex sets in $\mathcal{X}$. (They actually considered density contour clusters; cf. appendix A for a definition.) Nolan (1991) considered higher dimensions with $\mathcal{C}$ being the class of ellipsoids. Tsybakov (1997) has studied an estimator based on piecewise polynomial approximation of $C(\alpha)$ and has shown it attains the asymptotically minimax rate for certain classes of densities p. Polonik (1997) has studied the estimation of $C(\alpha)$ by $C_\ell(\alpha)$. He derived asymptotic rates of convergence in terms of various measures of richness of $\mathcal{C}$. He considered both VC classes and classes with a log ϵ-covering number with bracketing of order $O\left(\epsilon^{-r}\right)$ for $r > 0$. More information on minimum volume estimators can be found in that work and in appendix A.

A number of applications have been suggested for these techniques. They include problems in medical diagnosis (Tarassenko, Hayton, Cerneaz, & Brady, 1995), marketing (Ben-David & Lindenbaum, 1997), condition monitoring of machines (Devroye & Wise, 1980), estimating manufacturing yields (Stoneking, 1999), econometrics and generalized nonlinear principal curves (Tsybakov, 1997; Korostelev & Tsybakov, 1993), regression and spectral analysis (Polonik, 1997), tests for multimodality and clustering (Polonik, 1995b), and others (Müller, 1992). Most of this work, in particular that with a theoretical slant, does not go all the way in devising practical algorithms that work on high-dimensional real-world-problems. A notable exception to this is the work of Tarassenko et al. (1995).

Polonik, (1995a) has shown how one can use estimators of $C(\alpha)$ to construct density estimators. The point of doing this is that it allows one to encode a range of prior assumptions about the true density p that would be impossible to do within the traditional density estimation framework. He has shown asymptotic consistency and rates of convergence for densities belonging to VC-classes or with a known rate of growth of metric entropy with bracketing.

Let us conclude this section with a short discussion of how the work presented here relates to the above. This article describes an algorithm that finds regions close to $C(\alpha)$. Our class $\mathcal{C}$ is defined implicitly via a kernel k as the set of half-spaces in an SV feature space. We do not try to minimize the volume of C in input space. Instead, we minimize an SV-style regularizer that, using a kernel, controls the smoothness of the estimated function describing C. In terms of multidimensional quantiles, our approach can be thought of as employing $\lambda(C_w) = \|w\|^2$, where $C_w = \{x: f_w(x) \geq \rho\}$. Here,

(w, ρ) are a weight vector and an offset parameterizing a hyperplane in the feature space associated with the kernel.

The main contribution of this work is that we propose an algorithm that has tractable computational complexity, even in high-dimensional cases. Our theory, which uses tools very similar to those used by Polonik, gives results that we expect will be of more use in a finite sample size setting.

3 Algorithms

We first introduce terminology and notation conventions. We consider training data

$$\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{X}, \tag{3.1}$$

where $\ell \in \mathbb{N}$ is the number of observations and $\mathcal{X}$ is some set. For simplicity, we think of it as a compact subset of $\mathbb{R}^N$. Let Φ be a feature map $\mathcal{X} \to F$, that is, a map into an inner product space F such that the inner product in the image of Φ can be computed by evaluating some simple kernel (Boser, Guyon, & Vapnik, (1992), Vapnik, (1995); Schölkopf, Burges, et al., (1999))

$$k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})), \tag{3.2}$$

such as the gaussian kernel

$$k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/c}. \tag{3.3}$$

Indices i and j are understood to range over $1, \ldots, \ell$ (in compact notation: $i, j \in [\ell]$). Boldface Greek letters denote ℓ-dimensional vectors whose components are labeled using a normal typeface.

In the remainder of this section, we develop an algorithm that returns a function f that takes the value $+1$ in a "small" region capturing most of the data points and -1 elsewhere. Our strategy is to map the data into the feature space corresponding to the kernel and to separate them from the origin with maximum margin. For a new point $\mathbf{x}$, the value $f(\mathbf{x})$ is determined by evaluating which side of the hyperplane it falls on in feature space. Via the freedom to use different types of kernel functions, this simple geometric picture corresponds to a variety of nonlinear estimators in input space.

To separate the data set from the origin, we solve the following quadratic program:

$$\min_{w \in F, \boldsymbol{\xi} \in \mathbb{R}^\ell, \rho \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \tag{3.4}$$

$$\text{subject to } (w \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0. \tag{3.5}$$

Here, $\nu \in (0, 1]$ is a parameter whose meaning will become clear later.

Since nonzero slack variables ξ_i are penalized in the objective function, we can expect that if w and ρ solve this problem, then the decision function

$$f(\mathbf{x}) = \operatorname{sgn}((w \cdot \Phi(\mathbf{x})) - \rho) \tag{3.6}$$

will be positive for most examples $\mathbf{x}_i$ contained in the training set,[1] while the SV type regularization term $\|w\|$ will still be small. The actual trade-off between these two goals is controlled by ν.

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$\begin{aligned} L(w, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \\ - \sum_i \alpha_i \left((w \cdot \Phi(\mathbf{x}_i)) - \rho + \xi_i\right) - \sum_i \boldsymbol{\beta}_i \xi_i, \end{aligned} \tag{3.7}$$

and set the derivatives with respect to the primal variables w, $\boldsymbol{\xi}$, ρ equal to zero, yielding

$$w = \sum_i \alpha_i \Phi(\mathbf{x}_i), \tag{3.8}$$

$$\alpha_i = \frac{1}{\nu\ell} - \boldsymbol{\beta}_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \tag{3.9}$$

In equation 3.8, all patterns $\{\mathbf{x}_i : i \in [\ell],\ \alpha_i > 0\}$ are called support vectors. Together with equation 3.2, the SV expansion transforms the decision function, equation 3.6 into a kernel expansion:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho\right). \tag{3.10}$$

Substituting equation 3.8 and equation 3.9 into L (see equation 3.7) and using equation 3.2, we obtain the dual problem:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \text{ subject to } 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \qquad \sum_i \alpha_i = 1. \tag{3.11}$$

One can show that at the optimum, the two inequality constraints, equation 3.5, become equalities if α_i and $\boldsymbol{\beta}_i$ are nonzero, that is, if $0 < \alpha_i < 1/(\nu\ell)$. Therefore, we can recover ρ by exploiting that for any such α_i, the corresponding pattern $\mathbf{x}_i$ satisfies

$$\rho = (w \cdot \Phi(\mathbf{x}_i)) = \sum_j \alpha_j k(\mathbf{x}_j, \mathbf{x}_i). \tag{3.12}$$

[1] We use the convention that $\operatorname{sgn}(z)$ equals 1 for $z \geq 0$ and -1 otherwise.

Note that if ν approaches 0, the upper boundaries on the Lagrange multipliers tend to infinity, that is, the second inequality constraint in equation 3.11 becomes void. The problem then resembles the corresponding hard margin algorithm, since the penalization of errors becomes infinite, as can be seen from the primal objective function (see equation 3.4). It is still a feasible problem, since we have placed no restriction on the offset ρ, so it can become a large negative number in order to satisfy equation 3.5. If we had required $\rho \geq 0$ from the start, we would have ended up with the constraint $\sum_i \alpha_i \geq 1$ instead of the corresponding equality constraint in equation 3.11, and the multipliers α_i could have diverged.

It is instructive to compare equation 3.11 to a Parzen windows estimator. To this end, suppose we use a kernel that can be normalized as a density in input space, such as the gaussian (see equation 3.3). If we use $\nu = 1$, then the two constraints only allow the solution $\alpha_1 = \cdots = \alpha_\ell = 1/\ell$. Thus the kernel expansion in equation 3.10 reduces to a Parzen windows estimate of the underlying density. For $\nu < 1$, the equality constraint in equation 3.11 still ensures that the decision function is a thresholded density; however, in that case, the density will be represented only by a subset of training examples (the SVs)—those that are important for the decision (see equation 3.10) to be taken. Section 5 will explain the precise meaning of ν.

To conclude this section, we note that one can also use balls to describe the data in feature space, close in spirit to the algorithms of Schölkopf, Burges, and Vapnik (1995), with hard boundaries, and Tax and Duin (1999), with "soft margins." Again, we try to put most of the data into a small ball by solving, for $\nu \in (0, 1)$,

$$\begin{aligned} \min_{R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell, c \in F} \quad & R^2 + \frac{1}{\nu \ell} \sum_i \xi_i \\ \text{subject to} \quad & \|\Phi(\mathbf{x}_i) - c\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0 \text{ for } i \in [\ell]. \end{aligned} \tag{3.13}$$

This leads to the dual

$$\min_{\boldsymbol{\alpha}} \quad \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) \tag{3.14}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu \ell}, \quad \sum_i \alpha_i = 1 \tag{3.15}$$

and the solution

$$c = \sum_i \alpha_i \Phi(\mathbf{x}_i), \tag{3.16}$$

corresponding to a decision function of the form

$$f(\mathbf{x}) = \operatorname{sgn}\left(R^2 - \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + 2 \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - k(\mathbf{x}, \mathbf{x}) \right). \tag{3.17}$$

Similar to the above, R^2 is computed such that for any $\mathbf{x}_i$ with $0 < \alpha_i < 1/(\nu\ell)$, the argument of the sgn is zero.

For kernels $k(\mathbf{x}, \mathbf{y})$ that depend on only $\mathbf{x} - \mathbf{y}$, $k(\mathbf{x}, \mathbf{x})$ is constant. In this case, the equality constraint implies that the linear term in the dual target function is constant, and the problem, 3.14 and 3.15, turns out to be equivalent to equation 3.11. It can be shown that the same holds true for the decision function; hence, the two algorithms coincide in that case. This is geometrically plausible. For constant $k(\mathbf{x}, \mathbf{x})$, all mapped patterns lie on a sphere in feature space. Therefore, finding the smallest sphere (containing the points) really amounts to finding the smallest segment of the sphere that the data live on. The segment, however, can be found in a straightforward way by simply intersecting the data sphere with a hyperplane; the hyperplane with maximum margin of separation to the origin will cut off the smallest segment.

4 Optimization

Section 3 formulated quadratic programs (QPs) for computing regions that capture a certain fraction of the data. These constrained optimization problems can be solved using an off-the-shelf QP package to compute the solution. They do, however, possess features that set them apart from generic QPs, most notably the simplicity of the constraints. In this section, we describe an algorithm that takes advantage of these features and empirically scales better to large data set sizes than a standard QP solver with time complexity of order $O(\ell^3)$ (cf. Platt, 1999). The algorithm is a modified version of SMO (sequential minimal optimization), an SV training algorithm originally proposed for classification (Platt, 1999), and subsequently adapted to regression estimation (Smola & Schölkopf, in press).

The strategy of SMO is to break up the constrained minimization of equation 3.11 into the smallest optimization steps possible. Due to the constraint on the sum of the dual variables, it is impossible to modify individual variables separately without possibly violating the constraint. We therefore resort to optimizing over pairs of variables.

4.1 Elementary Optimization Step. For instance, consider optimizing over α_1 and α_2 with all other variables fixed. Using the shorthand $K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, equation 3.11 then reduces to

$$\min_{\alpha_1, \alpha_2} \frac{1}{2} \sum_{i,j=1}^{2} \alpha_i \alpha_j K_{ij} + \sum_{i=1}^{2} \alpha_i C_i + C, \tag{4.1}$$

with $C_i := \sum_{j=3}^{\ell} \alpha_j K_{ij}$ and $C := \sum_{i,j=3}^{\ell} \alpha_i \alpha_j K_{ij}$, subject to

$$0 \le \alpha_1, \alpha_2 \le \frac{1}{\nu\ell}, \quad \sum_{i=1}^{2} \alpha_i = \Delta, \tag{4.2}$$

where $\Delta := 1 - \sum_{i=3}^{\ell} \alpha_i$.

We discard C, which is independent of α_1 and α_2, and eliminate α_1 to obtain

$$\min_{\alpha_2} \frac{1}{2}(\Delta - \alpha_2)^2 K_{11} + (\Delta - \alpha_2)\alpha_2 K_{12} + \frac{1}{2}\alpha_2^2 K_{22} + (\Delta - \alpha_2)C_1 + \alpha_2 C_2, \tag{4.3}$$

with the derivative

$$-(\Delta - \alpha_2)K_{11} + (\Delta - 2\alpha_2)K_{12} + \alpha_2 K_{22} - C_1 + C_2. \tag{4.4}$$

Setting this to zero and solving for α_2, we get

$$\alpha_2 = \frac{\Delta(K_{11} - K_{12}) + C_1 - C_2}{K_{11} + K_{22} - 2K_{12}}. \tag{4.5}$$

Once α_2 is found, α_1 can be recovered from $\alpha_1 = \Delta - \alpha_2$. If the new point (α_1, α_2) is outside of $[0, 1/(\nu\ell)]$, the constrained optimum is found by projecting α_2 from equation 4.5 into the region allowed by the constraints, and then recomputing α_1.

The offset ρ is recomputed after every such step.

Additional insight can be obtained by rewriting equation 4.5 in terms of the outputs of the kernel expansion on the examples $\mathbf{x}_1$ and $\mathbf{x}_2$ before the optimization step. Let α_1^*, α_2^* denote the values of their Lagrange parameter before the step. Then the corresponding outputs (cf. equation 3.10) read

$$O_i := K_{1i}\alpha_1^* + K_{2i}\alpha_2^* + C_i. \tag{4.6}$$

Using the latter to eliminate the C_i, we end up with an update equation for α_2, which does not explicitly depend on α_1^*,

$$\alpha_2 = \alpha_2^* + \frac{O_1 - O_2}{K_{11} + K_{22} - 2K_{12}}, \tag{4.7}$$

which shows that the update is essentially the fraction of first and second derivative of the objective function along the direction of ν-constraint satisfaction.

Clearly, the same elementary optimization step can be applied to any pair of two variables, not just α_1 and α_2. We next briefly describe how to do the overall optimization.

4.2 Initialization of the Algorithm. We start by setting a fraction ν of all α_i, randomly chosen, to $1/(\nu\ell)$. If $\nu\ell$ is not an integer, then one of the examples is set to a value in $(0, 1/(\nu\ell))$ to ensure that $\sum_i \alpha_i = 1$. Moreover, we set the initial ρ to $\max\{O_i : i \in [\ell],\ \alpha_i > 0\}$.

4.3 Optimization Algorithm. We then select a first variable for the elementary optimization step in one of the two following ways. Here, we use the shorthand SV_{nb} for the indices of variables that are not at bound, that is, $SV_{nb} := \{i : i \in [\ell],\ 0 < \alpha_i < 1/(\nu\ell)\}$. At the end, these correspond to points that will sit exactly on the hyperplane and that will therefore have a strong influence on its precise position.

We scan over the entire data set[2] until we find a variable violating a Karush-kuhn-Tucker (KKT) condition (Bertsekas, 1995), that is, a point such that $(O_i - \rho) \cdot \alpha_i > 0$ or $(\rho - O_i) \cdot (1/(\nu\ell) - \alpha_i) > 0$. Once we have found one, say α_i, we pick α_j according to

$$j = \arg\max_{n \in SV_{nb}} |O_i - O_n|. \tag{4.8}$$

We repeat that step, but the scan is performed only over SV_{nb}.

In practice, one scan of the first type is followed by multiple scans of the second type, until there are no KKT violators in SV_{nb}, whereupon the optimization goes back to a single scan of the first type. If it finds no KKT violators, the optimization terminates.

In unusual circumstances, the choice heuristic, equation 4.8, cannot make positive progress. Therefore, a hierarchy of other choice heuristics is applied to ensure positive progress. These other heuristics are the same as in the case of pattern recognition (cf. Platt, 1999), and have been found to work well in the experiments we report below.

In our experiments with SMO applied to distribution support estimation, we have always found it to converge. However, to ensure convergence even in rare pathological conditions, the algorithm can be modified slightly (Keerthi, Shevade, Bhattacharyya, & Murthy, 1999).

We end this section by stating a trick that is of importance in practical implementations. In practice, one has to use a nonzero accuracy tolerance when checking whether two quantities are equal. In particular, comparisons of this type are used in determining whether a point lies on the margin. Since we want the final decision function to evaluate to 1 for points that lie on the margin, we need to subtract this constant from the offset ρ at the end.

In the next section, it will be argued that subtracting something from ρ is actually advisable also from a statistical point of view.

[2] This scan can be accelerated by not checking patterns that are on the correct side of the hyperplane by a large margin, using the method of Joachims (1999)

5 Theory

We now analyze the algorithm theoretically, starting with the uniqueness of the hyperplane (proposition 1). We then describe the connection to pattern recognition (proposition 2), and show that the parameter ν characterizes the fractions of SVs and outliers (proposition 3). Following that, we give a robustness result for the soft margin (proposition 4) and finally we present a theoretical result on the generalization error (theorem 1). Some of the proofs are given in appendix B.

In this section, we use italic letters to denote the feature space images of the corresponding patterns in input space, that is,

$$x_i := \Phi(\mathbf{x}_i). \tag{5.1}$$

Definition 1. *A data set*

$$x_1, \ldots, x_\ell \tag{5.2}$$

is called separable if there exists some $w \in F$ such that $(w \cdot x_i) > 0$ for $i \in [\ell]$.

If we use a gaussian kernel (see equation 3.3), then any data set $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ is separable after it is mapped into feature space. To see this, note that $k(\mathbf{x}_i, \mathbf{x}_j) > 0$ for all i, j; thus all inner products between mapped patterns are positive, implying that all patterns lie inside the same orthant. Moreover, since $k(\mathbf{x}_i, \mathbf{x}_i) = 1$ for all i, they all have unit length. Hence, they are separable from the origin.

Proposition 1 (supporting hyperplane). *If the data set, equation 5.2, is separable, then there exists a unique supporting hyperplane with the properties that (1) it separates all data from the origin, and (2) its distance to the origin is maximal among all such hyperplanes. For any $\rho > 0$, it is given by*

$$\min_{w \in F} \frac{1}{2}\|w\|^2 \text{ subject to } (w \cdot x_i) \geq \rho, \quad i \in [\ell]. \tag{5.3}$$

The following result elucidates the relationship between single-class classification and binary classification.

Proposition 2 (connection to pattern recognition). *(i) Suppose (w, ρ) parameterizes the supporting hyperplane for the data in equation 5.2. Then $(w, 0)$ parameterizes the optimal separating hyperplane for the labeled data set,*

$$\{(x_1, 1), \ldots, (x_\ell, 1), (-x_1, -1), \ldots, (-x_\ell, -1)\}. \tag{5.4}$$

(ii) Suppose $(w, 0)$ *parameterizes the optimal separating hyperplane passing through the origin for a labeled data set,*

$$\{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}, \quad (y_i \in \{\pm 1\} \text{ for } i \in [\ell]), \tag{5.5}$$

aligned such that $(w \cdot x_i)$ *is positive for* $y_i = 1$*. Suppose, moreover, that* $\rho/\|w\|$ *is the margin of the optimal hyperplane. Then* (w, ρ) *parameterizes the supporting hyperplane for the unlabeled data set,*

$$\{y_1 x_1, \ldots, y_\ell x_\ell\}. \tag{5.6}$$

Note that the relationship is similar for nonseparable problems. In that case, margin errors in binary classification (points that are either on the wrong side of the separating hyperplane or fall inside the margin) translate into outliers in single-class classification, (points that fall on the wrong side of the hyperplane). Proposition 2 then holds, cum grano salis, for the training sets with margin errors and outliers, respectively, removed.

The utility of proposition 2 lies in the fact that it allows us to recycle certain results proven for binary classification (Schölkopf, Smola, Williamson, & Bartlett, 2000) for use in the single-class scenario. The following, explaining the significance of the parameter ν, is such a case.

Proposition 3 ν-property. *Assume the solution of equation 3.4 and 3.5 satisfies* $\rho \neq 0$*. The following statements hold:*

i. ν *is an upper bound on the fraction of outliers, that is, training points outside the estimated region.*

ii. ν *is a lower bound on the fraction of SVs.*

iii. Suppose the data (see equation 5.2) were generated independently from a distribution $P(\mathbf{x})$*, which does not contain discrete components. Suppose, moreover, that the kernel is analytic and nonconstant. With probability 1, asymptotically,* ν *equals both the fraction of SVs and the fraction of outliers.*

Note that this result also applies to the soft margin ball algorithm of Tax and Duin (1999), provided that it is stated in the ν-parameterization given in section 3.

Proposition 4 (resistance). *Local movements of outliers parallel to* w *do not change the hyperplane.*

Note that although the hyperplane does not change, its parameterization in w and ρ does.

We now move on to the subject of generalization. Our goal is to bound the probability that a novel point drawn from the same underlying distribution

lies outside the estimated region. We present a "marginalized" analysis that in fact provides a bound on the probability that a novel point lies outside the region slightly larger than the estimated one.

Definition 2. *Let f be a real-valued function on a space $\mathcal{X}$. Fix $\theta \in \mathbb{R}$. For $\mathbf{x} \in \mathcal{X}$ let $d(\mathbf{x}, f, \theta) = \max\{0, \theta - f(\mathbf{x})\}$. Similarly for a training sequence $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_\ell)$, we define*

$$\mathcal{D}(\mathbf{X}, f, \theta) = \sum_{\mathbf{x} \in \mathbf{X}} d(\mathbf{x}, f, \theta).$$

In the following, log denotes logarithms to base 2 and ln denotes natural logarithms.

Theorem 1 (generalization error bound). *Suppose we are given a set of ℓ examples $\mathbf{X} \in \mathcal{X}^\ell$ generated i.i.d. from an unknown distribution P, which does not contain discrete components. Suppose, moreover, that we solve the optimization problem, equations 3.4 and 3.5 (or equivalently equation 3.11) and obtain a solution f_w given explicitly by equation (3.10). Let $R_{w,\rho} := \{\mathbf{x}: f_w(\mathbf{x}) \geq \rho\}$ denote the induced decision region. With probability $1 - \delta$ over the draw of the random sample $\mathbf{X} \in \mathcal{X}^\ell$, for any $\gamma > 0$,*

$$P\left\{\mathbf{x}': \mathbf{x}' \notin R_{w,\rho-\gamma}\right\} \leq \frac{2}{\ell}\left(k + \log\frac{\ell^2}{2\delta}\right), \tag{5.7}$$

where

$$k = \frac{c_1 \log\left(c_2 \hat{\gamma}^2 \ell\right)}{\hat{\gamma}^2} + \frac{2\mathcal{D}}{\hat{\gamma}} \log\left(e\left(\frac{(2\ell - 1)\hat{\gamma}}{2\mathcal{D}} + 1\right)\right) + 2, \tag{5.8}$$

$c_1 = 16c^2$, $c_2 = \ln(2)/\left(4c^2\right)$, $c = 103$, $\hat{\gamma} = \gamma / \|w\|$, $\mathcal{D} = \mathcal{D}(\mathbf{X}, f_{w,0}, \rho) = \mathcal{D}(\mathbf{X}, f_{w,\rho}, 0)$, and ρ is given by equation (3.12).

The training sample $\mathbf{X}$ defines (via the algorithm) the decision region $R_{w,\rho}$. We expect that new points generated according to P will lie in $R_{w,\rho}$. The theorem gives a probabilistic guarantee that new points lie in the larger region $R_{w,\rho-\gamma}$.

The parameter ν can be adjusted when running the algorithm to trade off incorporating outliers versus minimizing the "size" of $R_{w,\rho}$. Adjusting ν will change the value of $\mathcal{D}$. Note that since $\mathcal{D}$ is measured with respect to ρ while the bound applies to $\rho - \gamma$, any point that is outside the region that the bound applies to will make a contribution to $\mathcal{D}$ that is bounded away from 0. Therefore, equation 5.7 does *not* imply that asymptotically, we will always estimate the complete support.

The parameter γ allows one to trade off the confidence with which one wishes the assertion of the theorem to hold against the size of the predictive region $R_{w,\rho-\gamma}$: one can see from equation 5.8 that k and hence the right-hand side of equation 5.7 scales inversely with γ. In fact, it scales inversely with $\hat{\gamma}$, that is, it increases with w. This justifies measuring the complexity of the estimated region by the size of w, and minimizing $\|w\|^2$ in order to find a region that will generalize well. In addition, the theorem suggests not to use the offset ρ returned by the algorithm, which would correspond to $\gamma = 0$, but a smaller value $\rho - \gamma$ (with $\gamma > 0$).

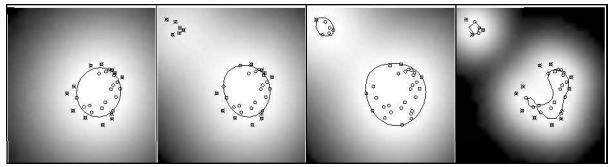
We do not claim that using theorem 1 directly is a practical means to determine the parameters ν and γ explicitly. It is loose in several ways. We suspect c is too large by a factor of more than 50. Furthermore, no account is taken of the smoothness of the kernel used. If that were done (by using refined bounds on the covering numbers of the induced class of functions as in Williamson, Smola, and Schölkopf (1998)), then the first term in equation 5.8 would increase much more slowly when decreasing γ. The fact that the second term would not change indicates a different trade-off point. Nevertheless, the theorem gives one some confidence that ν and γ are suitable parameters to adjust.

6 Experiments

We apply the method to artificial and real-world data. Figure 1 displays two-dimensional (2D) toy examples and shows how the parameter settings influence the solution. Figure 2 shows a comparison to a Parzen windows estimator on a 2D problem, along with a family of estimators that lie "in between" the present one and the Parzen one.

Figure 3 shows a plot of the outputs $(w \cdot \Phi(\mathbf{x}))$ on training and test sets of the U.S. Postal Service (USPS) database of handwritten digits. The database contains 9298 digit images of size $16 \times 16 = 256$; the last 2007 constitute the test set. We used a gaussian kernel (see equation 3.3) that has the advantage that the data are always separable from the origin in feature space (cf. the comment following definition 1). For the kernel parameter c, we used $0.5 \cdot 256$. This value was chosen a priori, it is a common value for SVM classifiers on that data set (cf. Schölkopf et al., 1995).[3] We fed our algorithm with the training instances of digit 0 only. Testing was done on both digit 0 and all other digits. We present results for two values of ν, one large, one small; for values in between, the results are qualitatively similar. In the first experiment, we used $\nu = 50\%$, thus aiming for a description of "0-ness," which

[3] Hayton, Schölkopf, Tarassenko, and Anuzis (in press), use the following procedure to determine a value of c. For small c, all training points will become SVs; the algorithm just memorizes the data and will not generalize well. As c increases, the number of SVs initially drops. As a simple heuristic, one can thus start with a small value of c and increase it until the number of SVs does not decrease any further.



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|w\|$	0.84	0.70	0.62	0.48

Figure 1: (First two pictures) A single-class SVM applied to two toy problems; $\nu = c = 0.5$, domain: $[-1, 1]^2$. In both cases, at least a fraction of ν of all examples is in the estimated region (cf. Table 1). The large value of ν causes the additional data points in the upper left corner to have almost no influence on the decision function. For smaller values of ν, such as 0.1 (third picture) the points cannot be ignored anymore. Alternatively, one can force the algorithm to take these outliers (OLs) into account by changing the kernel width (see equation 3.3). In the *fourth picture*, using $c = 0.1, \nu = 0.5$, the data are effectively analyzed on a different length scale, which leads the algorithm to consider the outliers as meaningful points.

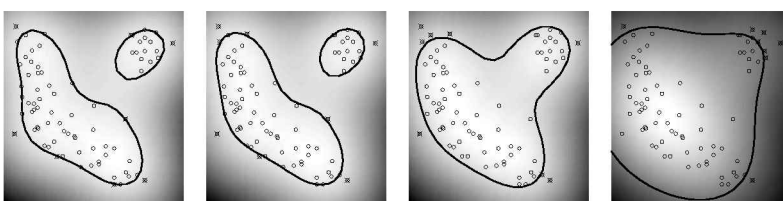


Figure 2: A single-class SVM applied to a toy problem; $c = 0.5$, domain: $[-1, 1]^2$, for various settings of the offset ρ. As discussed in section 3, $\nu = 1$ yields a Parzen windows expansion. However, to get a Parzen windows estimator of the distribution's support, we must in that case not use the offset returned by the algorithm (which would allow all points to lie outside the estimated region). Therefore, in this experiment, we adjusted the offset such that a fraction $\nu' = 0.1$ of patterns would lie outside. From left to right, we show the results for $\nu \in \{0.1, 0.2, 0.4, 1\}$. The right-most picture corresponds to the Parzen estimator that uses all kernels; the other estimators use roughly a fraction of ν kernels. Note that as a result of the averaging over all kernels, the Parzen windows estimate does not model the shape of the distribution very well for the chosen parameters.

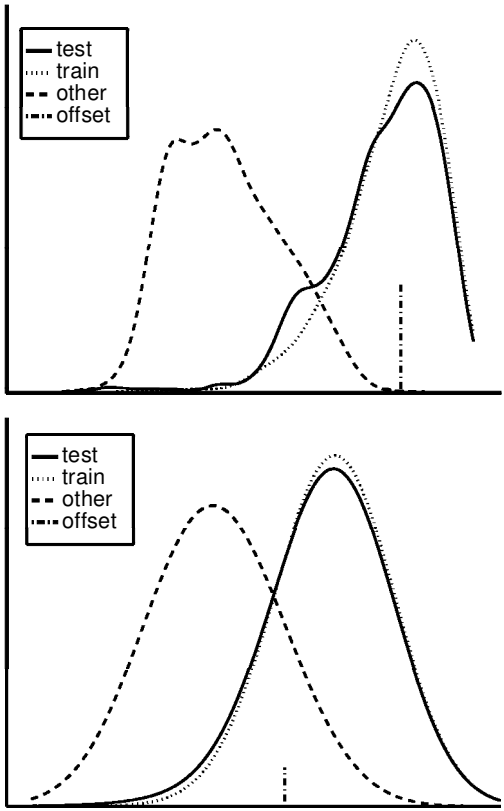


Figure 3: Experiments on the U.S. Postal Service OCR data set. Recognizer for digit 0; output histogram for the exemplars of 0 in the training/test set, and on test exemplars of other digits. The x-axis gives the output values, that is, the argument of the sgn function in equation 3.10. For $\nu = 50\%$ (top), we get 50% SVs and 49% outliers (consistent with proposition 3), 44% true positive test examples, and zero false positives from the "other" class. For $\nu = 5\%$ (bottom), we get 6% and 4% for SVs and outliers, respectively. In that case, the true positive rate is improved to 91%, while the false-positive rate increases to 7%. The offset ρ is marked in the graphs. Note, finally, that the plots show a Parzen windows density estimate of the output histograms. In reality, many examples sit exactly at the threshold value (the nonbound SVs). Since this peak is smoothed out by the estimator, the fractions of outliers in the training set appear slightly larger than it should be.

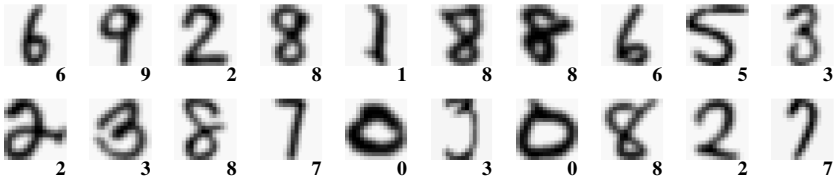


Figure 4: Subset of 20 examples randomly drawn from the USPS test set, with class labels.

captures only half of all zeros in the training set. As shown in Figure 3, this leads to zero false positives (although the learning machine has not seen any non-0's during training, it correctly identifies all non-0's as such), while still recognizing 44% of the digits 0 in the test set. Higher recognition rates can be achieved using smaller values of ν: for $\nu = 5\%$, we get 91% correct recognition of digits 0 in the test set, with a fairly moderate false-positive rate of 7%.

Although this experiment leads to encouraging results, it did not really address the actual task the algorithm was designed for. Therefore, we next focused on a problem of novelty detection. Again, we used the USPS set; however, this time we trained the algorithm on the test set and used it to identify outliers. It is folklore in the community that the USPS test set (see Figure 4) contains a number of patterns that are hard or impossible to classify, due to segmentation errors or mislabeling (Vapnik, 1995). In this experiment, we augmented the input patterns by 10 extra dimensions corresponding to the class labels of the digits. The rationale is that if we disregarded the labels, there would be no hope to identify mislabeled patterns as outliers. With the labels, the algorithm has the chance of identifying both unusual patterns and usual patterns with unusual labels. Figure 5 shows the 20 worst outliers for the USPS test set, respectively. Note that the algorithm indeed extracts patterns that are very hard to assign to their respective classes. In the experiment, we used the same kernel width as above and a ν value of 5%. The latter was chosen roughly to reflect our expectations as to how many "bad" patterns there are in the test set. Most good learning algorithms achieve error rates of 3 to 5% on the USPS benchmark (for a list of results, cf. Vapnik, 1995). Table 1 shows that ν lets us control the fraction of outliers.

In the last experiment, we tested the run-time scaling behavior of the proposed SMO solver, which is used for training the learning machine (see Figure 6). It was found to depend on the value of ν used. For the small values of ν that are typically used in outlier detection tasks, the algorithm scales very well to larger data sets, with a dependency of training times on the sample size, which is at most quadratic.

In addition to the experiments reported above, the algorithm has since

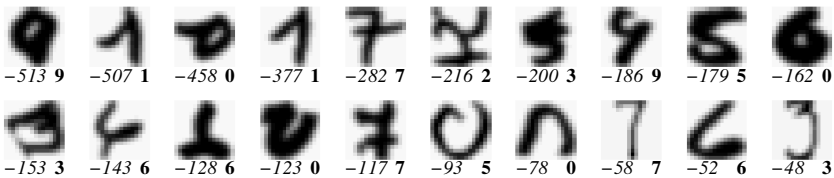


Figure 5: Outliers identified by the proposed algorithm, ranked by the negative output of the SVM (the argument of equation 3.10). The outputs (for convenience in units of 10^{-5}) are written underneath each image in italics; the (alleged) class labels are given in boldface. Note that most of the examples are "difficult" in that they are either atypical or even mislabeled.

Table 1: Experimental Results for Various Values of the Outlier Control Constant ν, USPS Test Set, Size $\ell = 2007$.

ν	Fraction of OLs	Fraction of SVs	Training Time (CPU sec)
1%	0.0%	10.0%	36
2%	0.0	10.0	39
3%	0.1	10.0	31
4%	0.6	10.1	40
5%	**1.4**	**10.6**	**36**
6%	1.8	11.2	33
7%	2.6	11.5	42
8%	4.1	12.0	53
9%	5.4	12.9	76
10%	6.2	13.7	65
20%	16.9	22.6	193
30%	27.5	31.8	269
40%	37.1	41.7	685
50%	47.4	51.2	1284
60%	58.2	61.0	1150
70%	68.3	70.7	1512
80%	78.5	80.5	2206
90%	89.4	90.1	2349

Notes: ν bounds the fractions of outliers and support vectors from above and below, respectively (cf. Proposition 3). As we are not in the asymptotic regime, there is some slack in the bounds; nevertheless, ν can be used to control the above fractions. Note, moreover, that training times (CPU time in seconds on a Pentium II running at 450 MHz) increase as ν approaches 1. This is related to the fact that almost all Lagrange multipliers will be at the upper bound in that case (cf. section 4). The system used in the outlier detection experiments is shown in boldface.

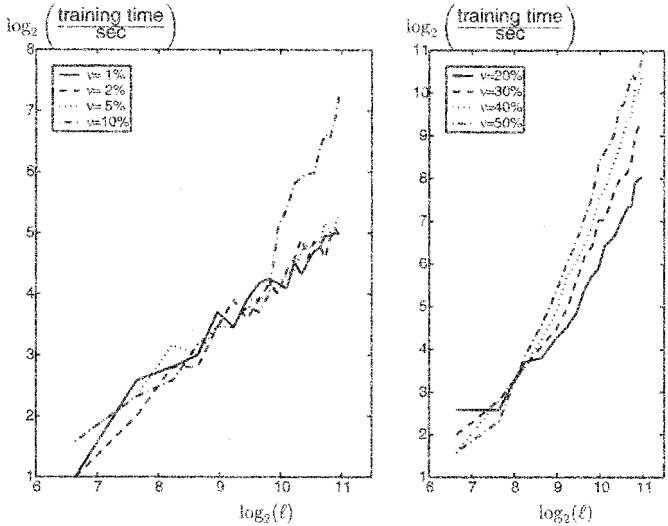


Figure 6: Training times versus data set sizes ℓ (both axes depict logs at base 2; CPU time in seconds on a Pentium II running at 450 MHz, training on subsets of the USPS test set); $c = 0.5 \cdot 256$. As in Table 1, it can be seen that larger values of ν generally lead to longer training times (note that the plots use different y-axis ranges). However, they also differ in their scaling with the sample size. The exponents can be directly read off from the slope of the graphs, as they are plotted in log scale with equal axis spacing. For small values of ν ($\leq 5\%$), the training times were approximately linear in the training set size. The scaling gets worse as ν increases. For large values of ν, training times are roughly proportional to the sample size raised to the power of 2.5 (right plot). The results should be taken only as an indication of what is going on; they were obtained using fairly small training sets, the largest being 2007, the size of the USPS test set. As a consequence, they are fairly noisy and refer only to the examined regime. Encouragingly, the scaling is better than the cubic one that one would expect when solving the optimization problem using all patterns at once (cf. section 4). Moreover, for small values of ν, that is, those typically used in outlier detection (in Figure 5, we used $\nu = 5\%$), our algorithm was particularly efficient.

been applied in several other domains, such as the modeling of parameter regimes for the control of walking robots (Still & Schölkopf, in press), and condition monitoring of jet engines (Hayton et al., in press).

7 Discussion

One could view this work as an attempt to provide a new algorithm in line with Vapnik's principle never to solve a problem that is more general

than the one that one is actually interested in. For example, in situations where one is interested only in detecting novelty, it is not always necessary to estimate a full density model of the data. Indeed, density estimation is more difficult than what we are doing, in several respects.

Mathematically, a density will exist only if the underlying probability measure possesses an absolutely continuous distribution function. However, the general problem of estimating the measure for a large class of sets, say, the sets measureable in Borel's sense, is not solvable (for a discussion, see Vapnik, 1998). Therefore we need to restrict ourselves to making a statement about the measure of some sets. Given a small class of sets, the simplest estimator that accomplishes this task is the empirical measure, which simply looks at how many training points fall into the region of interest. Our algorithm does the opposite. It starts with the number of training points that are supposed to fall into the region and then estimates a region with the desired property. Often there will be many such regions. The solution becomes unique only by applying a regularizer, which in our case enforces that the region be small in a feature space associated with the kernel.

Therefore, we must keep in mind that the measure of smallness in this sense depends on the kernel used, in a way that is no different from any other method that regularizes in a feature space. A similar problem, however, appears in density estimation already when done in input space. Let p denote a density on $\mathcal{X}$. If we perform a (nonlinear) coordinate transformation in the input domain $\mathcal{X}$, then the density values will change; loosely speaking, what remains constant is $p(\mathbf{x}) \cdot d\mathbf{x}$, while $d\mathbf{x}$ is transformed too. When directly estimating the probability measure of regions, we are not faced with this problem, as the regions automatically change accordingly.

An attractive property of the measure of smallness that we chose to use is that it can also be placed in the context of regularization theory, leading to an interpretation of the solution as maximally smooth in a sense that depends on the specific kernel used. More specifically, let us assume that k is Green's function of P^*P for an operator P mapping into some inner product space (Smola, Schölkopf, & Müller, 1998; Girosi, 1998), and take a look at the dual objective function that we minimize,

$$\begin{aligned}
\sum_{i,j} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i,j} \alpha_i \alpha_j \left(k(\mathbf{x}_i, \cdot) \cdot \delta_{\mathbf{x}_j}(\cdot)\right) \\
&= \sum_{i,j} \alpha_i \alpha_j \left(k(\mathbf{x}_i, \cdot) \cdot \left(P^*Pk\right)(\mathbf{x}_j, \cdot)\right) \\
&= \sum_{i,j} \alpha_i \alpha_j \left((Pk)(\mathbf{x}_i, \cdot) \cdot (Pk)(\mathbf{x}_j, \cdot)\right) \\
&= \left(\left(P \sum_i \alpha_i k\right)(\mathbf{x}_i, \cdot) \cdot \left(P \sum_j \alpha_j k\right)(\mathbf{x}_j, \cdot)\right) \\
&= \|Pf\|^2,
\end{aligned}$$

using $f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x})$. The regularization operators of common kernels can be shown to correspond to derivative operators (Poggio & Girosi, 1990); therefore, minimizing the dual objective function corresponds to maximizing the smoothness of the function f (which is, up to a thresholding operation, the function we estimate). This, in turn, is related to a prior $p(f) \sim e^{-\|Pf\|^2}$ on the function space.

Interestingly, as the minimization of the dual objective function also corresponds to a maximization of the margin in feature space, an equivalent interpretation is in terms of a prior on the distribution of the unknown other class (the "novel" class in a novelty detection problem); trying to separate the data from the origin amounts to assuming that the novel examples lie around the origin.

The main inspiration for our approach stems from the earliest work of Vapnik and collaborators. In 1962, they proposed an algorithm for characterizing a set of unlabeled data points by separating it from the origin using a hyperplane (Vapnik & Lerner, 1963; Vapnik & Chervonenkis, 1974). However, they quickly moved on to two-class classification problems, in terms of both algorithms and the theoretical development of statistical learning theory, which originated in those days.

From an algorithmic point of view, we can identify two shortcomings of the original approach, which may have caused research in this direction to stop for more than three decades. Firstly, the original algorithm in Vapnik and Chervonenkis (1974) was limited to linear decision rules in input space; second, there was no way of dealing with outliers. In conjunction, these restrictions are indeed severe; a generic data set need not be separable from the origin by a hyperplane in input space.

The two modifications that we have incorporated dispose of these shortcomings. First, the kernel trick allows for a much larger class of functions by nonlinearly mapping into a high-dimensional feature space and thereby increases the chances of a separation from the origin being possible. In particular, using a gaussian kernel (see equation 3.3), such a separation is always possible, as shown in section 5. The second modification directly allows for the possibility of outliers. We have incorporated this softness of the decision rule using the ν-trick (Schölkopf, Platt, & Smola, 2000) and thus obtained a direct handle on the fraction of outliers.

We believe that our approach, proposing a concrete algorithm with well-behaved computational complexity (convex quadratic programming) for a problem that so far has mainly been studied from a theoretical point of view, has abundant practical applications. To turn the algorithm into an easy-to-use black-box method for practitioners, questions like the selection of kernel parameters (such as the width of a gaussian kernel) have to be tackled. It is our hope that theoretical results as the one we have briefly outlined and the one of Vapnik and Chapelle (2000) will provide a solid foundation for this formidable task. This, alongside with algorithmic extensions such as the

possibility of taking into account information about the "abnormal" class (Schölkopf, Platt, & Smola, 2000), is subject of current research.

Appendix A: Supplementary Material for Section 2

A.1 Estimating the Support of a Density. The problem of estimating $C(1)$ appears to have first been studied by Geffroy (1964) who considered $\mathcal{X} = \mathbb{R}^2$ with piecewise constant estimators. There have been a number of works studying a natural nonparametric estimator of $C(1)$ (e.g., Chevalier, 1976; Devroye & Wise, 1980; see Gayraud, 1997 for further references). The nonparametric estimator is simply

$$\hat{C}_\ell = \bigcup_{i=1}^{\ell} B(\mathbf{x}_i, \epsilon_\ell), \tag{A.1}$$

where $B(\mathbf{x}, \epsilon)$ is the $l_2(\mathcal{X})$ ball of radius ϵ centered at $\mathbf{x}$ and $(\epsilon_\ell)_\ell$ is an appropriately chosen decreasing sequence. Devroye & Wise (1980) showed the asymptotic consistency of (A.1) with respect to the symmetric difference between $C(1)$ and $\hat{C}_\ell$. Cuevas and Fraiman (1997) studied the asymptotic consistency of a plug-in estimator of $C(1)$: $\hat{C}^{\text{plug-in}} = \{\mathbf{x}: \hat{p}_\ell(\mathbf{x}) > 0\}$ where $\hat{p}_\ell$ is a kernel density estimator. In order to avoid problems with $\hat{C}^{\text{plug-in}}$, they analyzed

$\hat{C}_\beta^{\text{plug-in}} := \{\mathbf{x}: \hat{p}_\ell(\mathbf{x}) > \beta_\ell\}$, where $(\beta_\ell)_\ell$ is an appropriately chosen sequence. Clearly for a given distribution, α is related to β, but this connection cannot be readily exploited by this type of estimator.

The most recent work relating to the estimation of $C(1)$ is by Gayraud (1997), who has made an asymptotic minimax study of estimators of functionals of $C(1)$. Two examples are $\operatorname{vol} C(1)$ or the center of $C(1)$. (See also Korostelev & Tsybakov, 1993, chap. 8).

A.2 Estimating High Probability Regions ($\alpha \neq 1$). Polonik (1995b) has studied the use of the "excess mass approach" (Müller, 1992) to construct an estimator of "generalized α-clusters" related to $C(\alpha)$.

Define the excess mass over $\mathcal{C}$ at level α as

$$E_{\mathcal{C}}(\alpha) = \sup\{H_\alpha(C): C \in \mathcal{C}\},$$

where $H_\alpha(\cdot) = (P - \alpha\lambda)(\cdot)$ and again λ denotes Lebesgue measure. Any set $\Gamma_{\mathcal{C}}(\alpha) \in \mathcal{C}$ such that

$$E_{\mathcal{C}}(\alpha) = H_\alpha(\Gamma_{\mathcal{C}}(\alpha))$$

is called a generalized α-cluster in $\mathcal{C}$. Replace P by P_ℓ in these definitions to obtain their empirical counterparts $E_{\ell,\mathcal{C}}(\alpha)$ and $\Gamma_{\ell,\mathcal{C}}(\alpha)$. In other words, his estimator is

$$\Gamma_{\ell,\mathcal{C}}(\alpha) = \arg\max\{(P_\ell - \alpha\lambda)(C): C \in \mathcal{C}\},$$

where the max is not necessarily unique. Now while $\Gamma_{\ell,C}(\alpha)$ is clearly different from $C_\ell(\alpha)$, it is related to it in that it attempts to find small regions with as much excess mass (which is similar to finding small regions with a given amount of probability mass). Actually $\Gamma_{\ell,C}(\alpha)$ is more closely related to the determination of density contour clusters at level α:

$$c_p(\alpha) := \{\mathbf{x}\colon p(\mathbf{x}) \geq \alpha\}.$$

Simultaneously, and independently, Ben-David & Lindenbaum (1997) studied the problem of estimating $c_p(\alpha)$. They too made use of VC classes but stated their results in a stronger form, which is meaningful for finite sample sizes.

Finally we point out a curious connection between minimum volume sets of a distribution and its differential entropy in the case that $\mathcal{X}$ is one-dimensional. Suppose X is a one-dimensional random variable with density p. Let $S = C(1)$ be the support of p and define the differential entropy of X by $h(X) = -\int_S p(x)\log p(x)dx$. For $\epsilon > 0$ and $\ell \in \mathbb{N}$, define the typical set $A_\epsilon^{(\ell)}$ with respect to p by

$$A_\epsilon^{(\ell)} = \left\{(x_1, \ldots, x_\ell) \in S^\ell\colon \left| -\tfrac{1}{\ell}\log p(x_1, \ldots, x_\ell) - h(X)\right| \leq \epsilon\right\},$$

where $p(x_1, \ldots, x_\ell) = \prod_{i=1}^{\ell} p(x_i)$. If $(a_\ell)_\ell$ and $(b_\ell)_\ell$ are sequences, the notation $a_\ell \doteq b_\ell$ means $\lim_{\ell\to\infty} \frac{1}{\ell}\log\frac{a_\ell}{b_\ell} = 0$. Cover and Thomas, (1991) show that for all $\epsilon, \delta < \frac{1}{2}$, then

$$\operatorname{vol} A_\epsilon^{(\ell)} \doteq \operatorname{vol} C_\ell(1-\delta) \doteq 2^{\ell h}.$$

They point out that this result "indicates that the volume of the smallest set that contains most of the probability is approximately $2^{\ell h}$. This is a ℓ-dimensional volume, so the corresponding side length is $(2^{\ell h})^{1/\ell} = 2^h$. This provides an interpretation of differential entropy" (p. 227).

Appendix B: Proofs of Section 5

Proof. (Proposition 1). Due to the separability, the convex hull of the data does not contain the origin. The existence and uniqueness of the hyperplane then follow from the supporting hyperplane theorem (Bertsekas, 1995).

Moreover, separability implies that there actually exists some $\rho > 0$ and $w \in F$ such that $(w \cdot x_i) \geq \rho$ for $i \in [\ell]$ (by rescaling w, this can be seen to work for arbitrarily large ρ). The distance of the hyperplane $\{z \in F : (w \cdot z) = \rho\}$ to the origin is $\rho/\|w\|$. Therefore the optimal hyperplane is obtained by minimizing $\|w\|$ subject to these constraints, that is, by the solution of equation 5.3.

Proof. (Proposition 2). *Ad (i).* By construction, the separation of equation 5.6 is a point-symmetric problem. Hence, the optimal separating hyperplane passes through the origin, for, if it did not, we could obtain another

optimal separating hyperplane by reflecting the first one with respect to the origin this would contradict the uniqueness of the optimal separating hyperplane (Vapnik, 1995).

Next, observe that $(-w, \rho)$ parameterizes the supporting hyperplane for the data set reflected through the origin and that it is parallel to the one given by (w, ρ). This provides an optimal separation of the two sets, with distance 2ρ, and a separating hyperplane $(w, 0)$.

Ad (ii). By assumption, w is the shortest vector satisfying $y_i(w \cdot x_i) \geq \rho$ (note that the offset is 0). Hence, equivalently, it is also the shortest vector satisfying $(w \cdot y_i x_i) \geq \rho$ for $i \in [\ell]$.

Proof. Sketch (Proposition 3). When changing ρ, the term $\sum_i \xi_i$ in equation 3.4 will change proportionally to the number of points that have a nonzero ξ_i (the outliers), plus, when changing ρ in the positive direction, the number of points that are just about to get a nonzero ρ, that is, which sit on the hyperplane (the SVs). At the optimum of equation 3.4, we therefore have parts i and ii. Part iii can be proven by a uniform convergence argument showing that since the covering numbers of kernel expansions regularized by a norm in some feature space are well behaved, the fraction of points that lie exactly on the hyperplane is asymptotically negligible (Schölkopf, Smola, et al., 2000).

Proof. (Proposition 4). Suppose x_o is an outlier, that is, $\xi_o > 0$, hence by the KKT conditions (Bertsekas, 1995) $\alpha_o = 1/(\nu\ell)$. Transforming it into $x'_o := x_o + \delta \cdot w$, where $|\delta| < \xi_o / \|w\|$, leads to a slack that is still nonzero, that is, $\xi'_o > 0$; hence, we still have $\alpha_o = 1/(\nu\ell)$. Therefore, $\boldsymbol{\alpha}' = \boldsymbol{\alpha}$ is still feasible, as is the primal solution $(w', \boldsymbol{\xi}', \rho')$. Here, we use $\xi'_i = (1 + \delta \cdot \alpha_o)\xi_i$ for $i \neq o$, $w' = (1 + \delta \cdot \alpha_o)w$, and ρ' as computed from equation 3.12. Finally, the KKT conditions are still satisfied, as still $\alpha'_o = 1/(\nu\ell)$. Thus (Bertsekas, 1995), $\boldsymbol{\alpha}$ is still the optimal solution.

We now move on to the proof of theorem 1. We start by introducing a common tool for measuring the capacity of a class $\mathcal{F}$ of functions that map $\mathcal{X}$ to $\mathbb{R}$.

Definition 3. *Let (X, d) be a pseudometric space, and let A be a subset of X and $\epsilon > 0$. A set $U \subseteq X$ is an* ϵ-cover *for A if, for every $a \in A$, there exists $u \in U$ such that $d(a, u) \leq \epsilon$. The* ϵ-covering number *of A, $\mathcal{N}(\epsilon, A, d)$, is the minimal cardinality of an ϵ-cover for A (if there is no such finite cover, then it is defined to be ∞).*

Note that we have used "less than or equal to" in the definition of a cover. This is somewhat unconventional, but will not change the bounds we use. It is, however, technically useful in the proofs.

The idea is that U should be finite but approximate all of A with respect to the pseudometric d. We will use the l_∞ norm over a finite sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_\ell)$ for the pseudonorm on $\mathcal{F}$,

$$\|f\|_{l_\infty^{\mathbf{X}}} := \max_{\mathbf{x} \in \mathbf{X}} |f(\mathbf{x})|. \tag{B.2}$$

(The (pseudo)-norm induces a (pseudo)-metric in the usual way.) Suppose $\mathcal{X}$ is a compact subset of X. Let $\mathcal{N}(\epsilon, \mathcal{F}, \ell) = \max_{\mathbf{X} \in \mathcal{X}^\ell} \mathcal{N}(\epsilon, \mathcal{F}, l_\infty^{\mathbf{X}})$ (the maximum exists by definition of the covering number and compactness of $\mathcal{X}$).

We require a technical restriction on the function class, referred to as *sturdiness*, which is satisfied by standard classes such as kernel-based linear function classes and neural networks with bounded weights (see Shawe-Taylor & Wiliamson, 1999, and Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 1999, for further details.)

Below, the notation $\lceil t \rceil$ denotes the smallest integer $\geq t$. Several of the results stated claim that some event occurs with probability $1 - \delta$. In all cases, $0 < \delta < 1$ is understood, and δ is presumed in fact to be small.

Theorem 2. *Consider any distribution P on $\mathcal{X}$ and a sturdy real-valued function class $\mathcal{F}$ on $\mathcal{X}$. Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ are generated i.i.d. from P. Then with probability $1 - \delta$ over such an ℓ-sample, for any $f \in \mathcal{F}$ and for any $\gamma > 0$,*

$$P\left\{\mathbf{x} : f(\mathbf{x}) < \min_{i \in [\ell]} f(\mathbf{x}_i) - 2\gamma\right\} \leq \epsilon(\ell, k, \delta) := \frac{2}{\ell}\left(k + \log\frac{\ell}{\delta}\right),$$

where $k = \lceil \log \mathcal{N}(\gamma, \mathcal{F}, 2\ell) \rceil$.

The proof, which is given in Schölkopf, Platt, et al. (1999), uses standard techniques of VC theory: symmetrization via a ghost sample, application of the union bound, and a permutation argument.

Although theorem 2 provides the basis of our analysis, it suffers from a weakness that a single, perhaps unusual point may significantly reduce the minimum $\min_{i \in [\ell]} f(\mathbf{x}_i)$. We therefore now consider a technique that will enable us to ignore some of the smallest values in the set $\{f(\mathbf{x}_i) : \mathbf{x}_i \in \mathbf{X}\}$ at a corresponding cost to the complexity of the function class. The technique was originally considered for analyzing soft margin classification in Shawe-Taylor and Cristianini (1999), but we will adapt it for the unsupervised problem we are considering here.

We will remove minimal points by increasing their output values. This corresponds to having a nonzero slack variable ξ_i in the algorithm, where we use the class of linear functions in feature space in the application of the theorem. There are well-known bounds for the log covering numbers of this class. We measure the increase in terms of $\mathcal{D}$ (definition 2).

Let $\mathcal{X}$ be an inner product space. The following definition introduces a new inner product space based on $\mathcal{X}$.

Definition 4. *Let $L(\mathcal{X})$ be the set of real valued nonnegative functions f on $\mathcal{X}$ with support* $\mathrm{supp}(f)$ *countable, that is, functions in $L(\mathcal{X})$ are nonzero for at most countably many points. We define the inner product of two functions $f, g \in L(\mathcal{X})$, by*

$$f \cdot g = \sum_{\mathbf{x} \in \mathrm{supp}(f)} f(\mathbf{x}) g(\mathbf{x}).$$

The 1-norm on $L(\mathcal{X})$ is defined by $\|f\|_1 = \sum_{\mathbf{x} \in \mathrm{supp}(f)} f(\mathbf{x})$ and let $L^B(\mathcal{X}) := \{f \in L(\mathcal{X}) : \|f\|_1 \leq B\}$. Now we define an embedding of $\mathcal{X}$ into the inner product space $\mathcal{X} \times L(\mathcal{X})$ via $\tau : \mathcal{X} \to \mathcal{X} \times L(\mathcal{X})$, $\tau : \mathbf{x} \mapsto (\mathbf{x}, \delta_x)$, where $\delta_x \in L(\mathcal{X})$ is defined by

$$\delta_x(\mathbf{y}) = \begin{cases} 1, & \text{if } \mathbf{y} = \mathbf{x}; \\ 0, & \text{otherwise.} \end{cases}$$

For a function $f \in \mathcal{F}$, a set of training examples $\mathbf{X}$, and $\theta \in \mathbb{R}$, we define the function $g_f \in L(\mathcal{X})$ by

$$g_f(\mathbf{y}) = g_f^{\mathbf{X},\theta}(\mathbf{y}) = \sum_{\mathbf{x} \in \mathbf{X}} d(\mathbf{x}, f, \theta) \delta_x(\mathbf{y}).$$

Theorem 3. *Fix $B > 0$. Consider a fixed but unknown probability distribution P which has no atomic components on the input space $\mathcal{X}$ and a sturdy class of real-valued functions $\mathcal{F}$. Then with probability $1 - \delta$ over randomly drawn training sequences $\mathbf{X}$ of size ℓ, for all $\gamma > 0$ and any $f \in \mathcal{F}$ and any θ such that $g_f = g_f^{\mathbf{X},\theta} \in L^B(\mathcal{X})$, (i.e., $\sum_{\mathbf{x} \in \mathbf{X}} d(\mathbf{x}, f, \theta) \leq B$),*

$$P\{\mathbf{x} : f(\mathbf{x}) < \theta - 2\gamma\} \leq \frac{2}{\ell}\left(k + \log \frac{\ell}{\delta}\right), \tag{B.3}$$

where $k = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2\ell) + \log \mathcal{N}(\gamma/2, L^B(\mathcal{X}), 2\ell) \right\rceil$.

The assumption on P can in fact be weakened to require only that there is no point $\mathbf{x} \in \mathbf{X}$ satisfying $f(\mathbf{x}) < \theta - 2\gamma$ that has discrete probability.) The proof is given in Schölkopf, Platt, et al. (1999).

The theorem bounds the probability of a new example falling in the region for which $f(\mathbf{x})$ has value less than $\theta - 2\gamma$, this being the complement of the estimate for the support of the distribution. In the algorithm described in this article, one would use the hyperplane shifted by 2γ toward the origin to define the region. Note that there is no restriction placed on the class of functions, though these functions could be probability density functions.

The result shows that we can bound the probability of points falling outside the region of estimated support by a quantity involving the ratio of the

log covering numbers (which can be bounded by the fat-shattering dimension at scale proportional to γ) and the number of training examples, plus a factor involving the 1-norm of the slack variables. The result is stronger than related results given by Ben-David and Lindenbaum, (1997): their bound involves the square root of the ratio of the Pollard dimension (the fat-shattering dimension when γ tends to 0) and the number of training examples.

In the specific situation considered in this article, where the choice of function class is the class of linear functions in a kernel-defined feature space, one can give a more practical expression for the quantity k in theorem 3. To this end, we use a result of Williamson, Smola, and Schölkopf (2000) to bound $\log \mathcal{N}(\gamma/2, \mathcal{F}, 2\ell)$, and a result of Shawe-Taylor and Cristianini (2000) to bound $\log \mathcal{N}(\gamma/2, L^B(\mathcal{X}), 2\ell)$. We apply theorem 3, stratifying over possible values of B. The proof is given in Schölkopf, Platt, et al. (1999) and leads to theorem 1 stated in the main part of this article.

Acknowledgments

This work was supported by the ARC, the DFG (#Ja 379/9-1 and Sm 62/1-1), and by the European Commission under the Working Group Nr. 27150 (NeuroCOLT2). Parts of it were done while B. S. and A. S. were with GMD FIRST, Berlin. Thanks to S. Ben-David, C. Bishop, P. Hayton, N. Oliver, C. Schnörr, J. Spanjaard, L. Tarassenko, and M. Tipping for helpful discussions.

References

Ben-David, S., & Lindenbaum, M. (1997). Learning distributions by their density levels: A paradigm for learning without a teacher. *Journal of Computer and System Sciences, 55*, 171–182.

Bertsekas, D. P. (1995). *Nonlinear programming*. Belmont, MA: Athena Scientific.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). Pittsburgh, PA: ACM Press.

Chevalier, J. (1976). Estimation du support et du contour du support d'une loi de probabilité. *Annales de l'Institut Henri Poincaré. Section B. Calcul des Probabilités et Statistique. Nouvelle Série, 12*(4),: 339–364.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Cuevas, A., & Fraiman, R. (1997). A plug-in approach to support estimation. *Annals of Statistics*, *25*(6), 2300–2312. 1997.

Devroye, L., & Wise, G. L. (1980). Detection of abnormal behaviour via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, *38*(3), 480–488.

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

Gayraud, G. (1997). Estimation of functional of density support. *Mathematical Methods of Statistics*, *6*(1), 26–46.

Geffroy, J. (1964). Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, *13*, 191–210.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation*, *10*(6), 1455–1480.

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

Hayton, P., Schölkopf, B., Tarassenko, L., & Anuzis, P. (in press). Support vector novelty detection applied to jet engine vibration spectra. In T. Leen, T. Dietterich, & V. Tresp (Eds)., *Advances in neural information processing systems, 13*.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods—Support vector learning* (pp. 169–184). Cambridge, MA: MIT Press.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (1999). Improvements to Platt's SMO algorithm for SVM classifier design (Tech. Rep. No. CD-99-14). Singapore: Department of Mechanical and Production Engineering, National University of Singapore.

Korostelev, A. P., & Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*. New York: Springer-Verlag.

Müller, D. W. (1992). *The excess mass approach in statistics*. Heidelberg: Beiträge zur Statistik, Universität Heidelberg.

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

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods—Support vector learning*, (pp. 185–208). Cambridge, MA: MIT Press.

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

Polonik, W. (1995a). Density estimation under qualitative assumptions in higher dimensions. *Journal of Multivariate Analysis, 55*(1), 61–81.

Polonik, W. (1995b). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Annals of Statistics, 23*(3), 855–881.

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

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

Schölkopf, B., Burges, C. J. C., & Smola, A. J. (1999). *Advances in kernel methods—Support vector learning*. Cambridge, MA: MIT Press.

Schölkopf, B., Burges, C., & Vapnik, V. (1995). Extracting support data for a given task. In U. M. Fayyad & R. Uthurusamy (Eds.), *Proceedings, First International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (1999). *Estimating the support of a high-dimensional distribution* (Tech. Rep. No. 87). Redmond, WA: Microsoft Research. Available online at: http://www.research.microsoft.com/scripts/pubs/view.asp?TR_ID=MSR-TR-99-87.

Schölkopf, B., Platt, J., & Smola, A. J. (2000). *Kernel method for percentile feature extraction* (Tech. Rep. No. 22). Redmond, WA: Microsoft Research.

Schölkopf, B., Smola, A., & Müller, K. R. (1999). Kernel principal component analysis. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods—Support vector learning* (pp. 327–352). Cambridge, MA: MIT Press.

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

Schölkopf, B., Williamson, R., Smola, A., & Shawe-Taylor, J. (1999). Single-class support vector machines. In J. Buhmann, W. Maass, H. Ritter, & N. Tishby, editors (Eds.) *Unsupervised learning* (Rep. No. 235) (pp. 19–20). Dagstuhl, Germany.

Shawe-Taylor, J., & Cristianini, N. (1999). Margin distribution bounds on generalization. In *Computational Learning Theory: 4th European Conference* (pp. 263–273). New York: Springer-Verlag.

Shawe-Taylor, J., & Cristianini, N. (2000). On the generalisation of soft margin algorithms. *IEEE Transactions on Information Theory*. Submitted.

Shawe-Taylor, J., & Williamson, R. C. (1999). Generalization performance of classifiers in terms of observed covering numbers. In *Computational Learning Theory: 4th European Conference* (pp. 274–284). New York: Springer-Verlag.

Smola, A., & Schölkopf, B. (in press). A tutorial on support vector regression. *Statistics and Computing*.

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

Smola, A., Mika, S., Schölkopf, B., & Williamson, R. C. (in press). Regularized principal manifolds. *Machine Learning*.

Still, S., & Schölkopf, B. (in press). Four-legged walking gait control using a neuromorphic chip interfaced to a support vector learning algorithm. In T. Leen, T. Diettrich, & P. Anuzis (Eds.), *Advances in neural information processing systems, 13*. Cambridge, MA: MIT Press.

Stoneking, D. (1999). Improving the manufacturability of electronic designs. *IEEE Spectrum, 36*(6), 70–76.

Tarassenko, L., Hayton, P., Cerneaz, N., & Brady, M. (1995). Novelty detection for the identification of masses in mammograms. In *Proceedings Fourth IEE International Conference on Artificial Neural Networks* (pp. 442–447). Cambridge.

Tax, D. M. J., & Duin, R. P. W. (1999). Data domain description by support vectors. In M. Verleysen (Ed.), *Proceedings ESANN* (pp. 251–256). Brussels: D Facto.

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

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V., & Chapelle, O. (2000). Bounds on error expectation for SVM. In A. J. Smola, P. L. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 261 – 280). Cambridge, MA: MIT Press.

Vapnik, V., & Chervonenkis, A. (1974). *Theory of pattern recognition*. Nauka, Moscow. [In Russian] (German translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control, 24*.

Williamson, R. C., Smola, A. J., & Schölkopf, B. (1998). *Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators* (Tech. Rep. No. 19) NeuroCOLT. Available online at: http://www.neurocolt.com, 1998. Also: *IEEE Transactions on Information Theory* (in press).

Williamson, R. C., Smola, A. J., & Schölkopf, B. (2000). Entropy numbers of linear function classes. In N. Cesa-Bianchi & S. Goldman (Eds.), *Proceedings of the 13th Annual Conference on Computational Learning Theory* (pp. 309–319). San Mateo, CA: Morgan Kaufman.

Received November 19, 1999; accepted September 22, 2000.