Introduction to Support Vector Classification

by Prof. Dr. Dr. Daniel J. Strauss

Contents

Chapter 1

Support Vector Machine Classification

1.1 Inducing Feature Spaces by Reproducing Kernels

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$ containing the data to be classified. We suppose that there exists an underlying unknown function t, the so–called *target function*, which maps $\mathcal{X}$ to the binary set $\{-1, 1\}$. Given a training set

$$\mathcal{A} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\} : i = 1, \dots, M\} \tag{1.1.1}$$

of M associations, we are interested in the construction of a real valued function f defined on $\mathcal{X}$ such that $\operatorname{sgn}(f)$ is a 'good approximation' of t which classifies the training data correctly, i.e., $\operatorname{sgn}(f(\mathbf{x}_i)) = t(\mathbf{x}_i) = y_i$ for all $i = 1, \dots, M$. Here

$$\operatorname{sgn}(f(\mathbf{x})) = \begin{cases} 1 & \text{if } f(\mathbf{x}) \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

We will search for the hypothesis function f in some reproducing kernel Hilbert spaces which we will introduce next.

Positive Define Functions. By $L^2(\mathcal{X})$ we denote the Hilbert space of real valued square integrable functions on $\mathcal{X}$ with inner product $\langle f, g \rangle_{L^2} = \int_{\mathcal{X}} f(x)g(x)\,\mathrm{d}x$. Let $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ be a positive definite symmetric function in $L^2(\mathcal{X} \times \mathcal{X})$. Following

[22], we call a function $K \in L^2(\mathcal{X} \times \mathcal{X})$ *positive definite* if for any finite set of elements $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathcal{X}$, the matrix $(K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ is positive definite. In this chapter, we are only interested in functions K arising from RBFs. In other words, we assume that there exists a real valued function k on $\mathbb{R}$ so that

$$K(\mathbf{x}, \mathbf{y}) = k(||\mathbf{x} - \mathbf{y}||_2), \tag{1.1.2}$$

where $||\cdot||_2$ denotes the Euclidean norm on $\mathbb{R}^d$. In our applications, we will use Gaussian kernels and Wendland's compactly supported RBFs [35]. The latter were not applied in connection with classification tasks up to now.

For a given K, there exists a *reproducing kernel Hilbert space*

$$\mathcal{H}_K = \overline{\operatorname{span}\{K(\tilde{\mathbf{x}}, \cdot) : \tilde{\mathbf{x}} \in \mathcal{X}\}}$$

of real valued functions on $\mathcal{X}$ with inner product determined by

$$\langle K(\tilde{\mathbf{x}}, \mathbf{x}), K(\bar{\mathbf{x}}, \mathbf{x}) \rangle_{\mathcal{H}_K} = K(\tilde{\mathbf{x}}, \bar{\mathbf{x}}) \tag{1.1.3}$$

which has reproducing kernel K, i.e.,

$$\langle f(\cdot), K(\tilde{\mathbf{x}}, \cdot) \rangle_{\mathcal{H}_K} = f(\tilde{\mathbf{x}}), \quad f \in \mathcal{H}_K.$$

By *Mercer's Theorem*, K can be expanded in a uniformly convergent series on $\mathcal{X} \times \mathcal{X}$

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \eta_j \varphi_j(\mathbf{x}) \varphi_j(\mathbf{y}), \tag{1.1.4}$$

where $\eta_j \geq 0$ are the eigenvalues of the integral operator $T_K : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ with $T_K f(\mathbf{y}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) \, \mathrm{d}\mathbf{x}$ and where $\{\varphi_j\}_{j \in \mathbb{N}}$ are the corresponding $L^2(\mathcal{X})$–orthonormalized eigenfunctions.

We introduce a so–called *feature map* $\Phi : \mathcal{X} \to \ell^2$ by

$$\Phi(\cdot) = \left(\sqrt{\eta_j} \varphi_j(\cdot)\right)_{j \in \mathbb{N}}.$$

Let ℓ^2 denote the Hilbert space of real valued quadratic summable sequences $a = (a_i)_{i \in \mathbb{N}}$ with inner product $\langle a, b \rangle_{\ell^2} = \sum_{i \in \mathbb{N}} a_i b_i$. By (1.1.4) we have that $\Phi(\mathbf{x})$ $(\mathbf{x} \in \mathcal{X})$ is an element in ℓ^2 with

$$||\Phi(\mathbf{x})||_{\ell^2}^2 = \sum_{j=1}^{\infty} \eta_j \varphi_j^2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) = k(0).$$

We define the *feature space* $\mathcal{F}_K \subset \ell^2$ by the ℓ^2–closure of all finite linear combinations of elements $\Phi(\mathbf{x})$ $(\mathbf{x} \in \mathcal{X})$

$$\mathcal{F}_K = \overline{\text{span}\,\{\Phi(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}}.$$

Then $\mathcal{F}_K$ is a Hilbert space with $||\cdot||_{\mathcal{F}_K} = ||\cdot||_{\ell^2}$. The feature space $\mathcal{F}_K$ and the reproducing kernel Hilbert space $\mathcal{H}_K$ are isometrically isomorph with isometry $\iota : \mathcal{F}_K \to \mathcal{H}_K$ defined by

$$\iota(\mathbf{w}) = f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle_{\ell^2} = \sum_{j=1}^{\infty} w_j \sqrt{\eta_j} \varphi_j(\mathbf{x}). \tag{1.1.5}$$

In particular, we have that

$$||f_{\mathbf{w}}||_{\mathcal{H}_K} = ||\mathbf{w}||_{\mathcal{F}_K}. \tag{1.1.6}$$

1.2 Regularization in RKHS and the Optimal Hyperplane

Let us turn to our classification task. For a given training set (1.1.1) we intend to construct a function $f \in \mathcal{H}_K$ which minimizes

$$\lambda \sum_{i=1}^{M} (1 - y_i f(\mathbf{x}_i))_+ + \frac{1}{2}||f||^2_{\mathcal{H}_K}, \tag{1.2.7}$$

where

$$(\tau)_+ = \begin{cases} \tau & \text{if } \tau \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that we can also look for functions of the form $f = h + b$ $(h \in \mathcal{H}_K)$ with a so–called bias term $b \in \mathbb{R}$. We omit the bias term b here, because its explicit consideration does not lead to an improvement of our numerical results in this chapter.

The unconstrained optimization problem (1.2.7) is equivalent to the following constraint optimization problem: find $f \in \mathcal{H}_K$ and u_i $(i = 1, \ldots, M)$ to minimize

$$\lambda \left(\sum_{i=1}^{M} u_i \right) + \frac{1}{2}||f||^2_{\mathcal{H}_K}, \tag{1.2.8}$$

subject to

$$\begin{aligned} y_i f(\mathbf{x}_i) &\geq 1 - u_i, \quad i = 1, \ldots, M, \\ u_i &\geq 0, \quad i = 1, \ldots, M. \end{aligned}$$

Every function $f \in \mathcal{H}_K$ corresponds uniquely to a sequence $\mathbf{w} \in \mathcal{F}_K$. Thus, by (1.1.5) and (1.1.6), the optimization problem (1.2.8) can be rewritten as follows: find $\mathbf{w} \in \mathcal{F}_K$ and u_i $(i = 1, \ldots, M)$ to minimize

$$\lambda \left(\sum_{i=1}^{M} u_i \right) + \frac{1}{2} ||\mathbf{w}||^2_{\mathcal{F}_K}, \tag{1.2.9}$$

subject to

$$\begin{aligned} y_i \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{F}_K} &\geq 1 - u_i \quad (i = 1, \ldots, M), \\ u_i &\geq 0 \quad (i = 1, \ldots, M). \end{aligned} \tag{1.2.10}$$

Optimal Hyperplane. In general the feature space $\mathcal{F}_K \subset \ell^2$ is infinitely dimensional. For a better illustration of (1.2.9) we assume for a moment that $\mathcal{F}_K \subset \mathbb{R}^n$. Then the function $\tilde{f}_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{w}, \mathbf{v} \rangle_{\mathcal{F}_K}$ defines a hyperplane $h_{\mathbf{w}} = \{\mathbf{v} \in \mathcal{F}_K : \tilde{f}_{\mathbf{w}}(\mathbf{v}) = 0\}$ in $\mathbb{R}^n$ through the origin and an arbitrary point $\mathbf{v}_i \in \mathcal{F}_K$ has the distance $\langle \mathbf{w}, \mathbf{v}_i \rangle_{\mathcal{F}_K} / ||\mathbf{w}||_{\mathcal{F}_K}$ from $h_{\mathbf{w}}$. Note that $\tilde{f}_{\mathbf{w}}(\Phi(\mathbf{x})) = f_{\mathbf{w}}(\mathbf{x})$. Thus the constraints $y_i \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{F}_K} / ||\mathbf{w}||_{\mathcal{F}_K} \geq 1/||\mathbf{w}||_{\mathcal{F}_K} - u_i/||\mathbf{w}||_{\mathcal{F}_K}$ $(i = 1, \ldots, M)$ in (1.2.10) require that every $\Phi(\mathbf{x}_i)$ must at least have the distance $1/||\mathbf{w}||_{\mathcal{F}_K} - u_i/||\mathbf{w}||_{\mathcal{F}_K}$ from the hyperplane.
If there exists $\mathbf{w} \in \mathcal{F}_K$ so that (1.2.10) can be fulfilled with $u_i = 0$ $(i = 1, \ldots, M)$, then we say that our training set is linearly separable in $\mathcal{F}_K$. Of course, for Gaussian kernels or kernels arising from Wendland's compactly supported RBFs every finite training set is linearly separable in $\mathcal{F}_K$, e.g., see [3] and [31]. Then the optimization problem (1.2.9) can be further simplified to: find $\mathbf{w} \in \mathcal{F}_K$ to minimize

$$\frac{1}{2} ||\mathbf{w}||^2_{\mathcal{F}_K} \tag{1.2.11}$$

subject to

$$y_i \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{F}_K} \geq 1 \quad (i = 1, \ldots, M).$$

Given $\mathcal{H}_K$ and $\mathcal{A}$, the optimization problem above has a unique solution $f_{\mathbf{w}^*}$. In our hyperplane context $h_{\mathbf{w}^*} = \{\mathbf{v} \in \mathcal{F}_K : \tilde{f}_{\mathbf{w}^*}(\mathbf{v}) = 0\}$ is exactly the hyperplane which has maximal distance γ from the training data, where

$$\gamma = \frac{1}{||\mathbf{w}^*||_{\mathcal{F}_K}} = \frac{1}{||f_{\mathbf{w}^*}||_{\mathcal{H}_K}} = \max_{\mathbf{w} \in \mathcal{F}_K} \min_{i=1,\ldots,M} \left\{ \frac{|\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{F}_K}|}{||\mathbf{w}||_{\mathcal{F}_K}} \right\}. \tag{1.2.12}$$

The value γ is called the *margin* of $h_{\mathbf{w}^*}$ with respect the training set $\mathcal{A}$. In this context, the solutions of the optimizations problems (1.2.9) and (1.2.11) are called *soft margin* and *hard margin SVM*, respectively. See Figure 1.1 for an illustration of the hard margin case in $\mathbb{R}^2$.

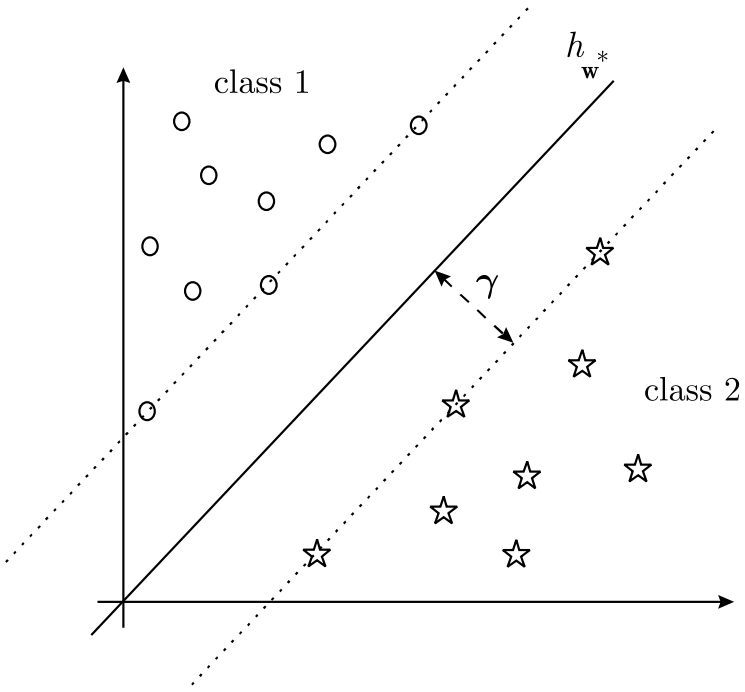

Figure 1.1: The separation of two classes by an optimal hyperplane $h_{\mathbf{w}^*}$ with margin γ.

Large Margin Theorem. To obtain a classifier that generalizes well, we have not only to minimize the error on the training data (the *empirical risk*) but also to adjust the capacity[1] of a learning machine, e.g., the number of hidden neurons for FFBNs, appropriately to the complexity of the data. In fact, this is one major result of statistical learning theory. Roughly speaking, it states that we should select the simplest model to explain the dependence of the training associations. SVMs control the capacity by an increasing margin and do not depend on parameters related to the input and feature space dimensionality. Therefore SVMs are often considered to be independent from the dimensionality. In particular, by [7, Theorem 4.18] we have that the generalization

[1]To measure the capacity, several quantities are known such as the *Vapnik Chervonenkis (VC) dimension* used in *structural risk minimization* [33], the *fat–shattering dimension* related to *data–dependent structural risk minimization* [7] as well as the *gamma–dimension* [10].

error of the hard margin SVM classifier decreases if the margin γ increases. In other words: *the larger the margin γ the better generalization of the SVM can be expected.* We call this the *large margin theorem* in our following discussions. Note that there exist also estimates for the generalization error of soft margin SVM classifiers which involve the margin of the unknown target function, see, e.g., [7, Theorem 4.21] and [31].

A major advantage of SVMs compared to FFBNs is that they adapt the complexity automatically (as we will see) and circumvent the unintentional freedom of choosing an appropriate complexity, i.e., determining the capacity, by hand. In general, there are only a few parameters to adjust.

1.3 Solution of the RKHS Regularization Problem

Next we consider the solution of regularization problems above, where we follow mainly the lines of [34]. Here the notation *support vector* (SV) comes into the play.

By the *Representer Theorem* ([17, 34]), the minimizer of (1.2.9), i.e., the hypothesis function, has the form

$$f(\mathbf{x}) = \sum_{j=1}^{M} c_j K(\mathbf{x}, \mathbf{x}_j). \tag{1.3.13}$$

Setting $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_M))^T$, $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^M$ and $\mathbf{c} = (c_1, \ldots, c_M)^T$ we obtain that

$$\mathbf{f} = \mathbf{K}\mathbf{c}.$$

Note that $\mathbf{K}$ is positive definite. Further, let $\mathbf{Y} = \mathrm{diag}(y_1, \ldots, y_M)$ and $\mathbf{u} = (u_1, \ldots, u_M)^T$. By $\mathbf{0}$ and $\mathbf{e}$ we denote the vectors with M entries 0 and 1, respectively. Then the optimization problem (1.2.9) can be rewritten as

$$\min_{\mathbf{u},\mathbf{c}} \ \lambda \mathbf{e}^T \mathbf{u} + \frac{1}{2} \mathbf{c}^T \mathbf{K} \mathbf{c} \tag{1.3.14}$$

subject to

$$\begin{aligned} \mathbf{u} &\geq \mathbf{e} - \mathbf{Y}\mathbf{K}\mathbf{c}, \\ \mathbf{u} &\geq \mathbf{0}. \end{aligned}$$

The dual problem with Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)^T$ and $\boldsymbol{\beta} = (\beta_1, \dots \beta_M)^T$ reads

$$\max_{\mathbf{c},\mathbf{u},\boldsymbol{\alpha},\boldsymbol{\beta}} L(\mathbf{c}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

where

$$L(\mathbf{c}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \lambda \mathbf{e}^T \mathbf{u} + \frac{1}{2}\mathbf{c}^T \mathbf{K}\mathbf{c} - \boldsymbol{\beta}^T \mathbf{u} + \boldsymbol{\alpha}^T \mathbf{e} - \boldsymbol{\alpha}^T \mathbf{Y}\mathbf{K}\mathbf{c} - \boldsymbol{\alpha}^T \mathbf{u}$$

subject to

$$\frac{\partial L}{\partial \mathbf{c}} = \mathbf{0},\ \frac{\partial L}{\partial \mathbf{u}} = \mathbf{0},\ \boldsymbol{\alpha} \geq \mathbf{0},\ \boldsymbol{\beta} \geq \mathbf{0}.$$

Now $\mathbf{0} = \frac{\partial L}{\partial \mathbf{c}} = \mathbf{K}\mathbf{c} - \mathbf{K}\mathbf{Y}\boldsymbol{\alpha}$ yields

$$\mathbf{c} = \mathbf{Y}\boldsymbol{\alpha}. \tag{1.3.15}$$

Further we have by $\frac{\partial L}{\partial \mathbf{u}} = \mathbf{0}$ that $\boldsymbol{\beta} = \lambda \mathbf{e} - \boldsymbol{\alpha}$. Thus our optimization problem becomes

$$\max_{\boldsymbol{\alpha}} \left(-\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Y}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha} + \mathbf{e}^T \boldsymbol{\alpha} \right) \tag{1.3.16}$$

subject to

$$\mathbf{0} \leq \boldsymbol{\alpha} \leq \lambda \mathbf{e}.$$

Quadratic Programming. This quadratic programming (QP) problem is usually solved in the SVM literature. For a moderate number of associations some standard QP routines can be used and for a large number of associations, e.g., $|\mathcal{A}| > 4000$, specifically designed large scale algorithms should be applied, e.g., SVM*light* [15]. Since such QP problems offer a global solution, they cannot be trapped into local minima during learning as FFBNs based on the backpropagation algorithm.

The SVs are those training patterns $\mathbf{x}_i$ for which α_i does not vanish. Let I denote the index set of the support vectors $I := \{i \in \{1, \dots, M\} : \alpha_i \neq 0\}$ then by (1.3.13) and (1.3.15), the function f has the sparse representation

$$f(\mathbf{x}) = \sum_{i \in I} c_i K(\mathbf{x}_i, \mathbf{x}) = \sum_{i \in I} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x})$$

which depends only on the SVs.

With respect to the margin we obtain by (1.2.12) and (1.1.3) that

$$\gamma = (||f||_{\mathcal{H}_K})^{-1} = (\mathbf{c}^T \mathbf{K}\mathbf{c})^{-1/2} = \left(\sum_{i \in I} y_i \alpha_i f(\mathbf{x}_i) \right)^{-1/2}.$$

Due to the Kuhn–Tucker conditions [11] the solution f of the QP problem (1.3.14) has to fulfill

$$\alpha_i(1 - y_i f(\mathbf{x}_i) - u_i) = 0, \quad i = 1 \dots, M.$$

In case of hard margin classification with $u_i = 0$ this implies that $y_i f(\mathbf{x}_i) = 1$ $(i \in I)$ so that we obtain the following simple expression for the margin

$$\gamma = \left(\sum_{i \in I} \alpha_i \right)^{-\frac{1}{2}}. \tag{1.3.17}$$

Support Vector Count Theorem. By [7, Theorem 6.8], the number of SVs can also be used to give an upper bound of the generalization error. *The fewer the number of support vectors the better generalization of the SVM can be expected.* Note that this theorem is in good accordance to our claims of building small models to describe dependencies as we have discussed in the previous chapters. Note also that Burges has given a nice and pictorially interpretation of this theorem in his tutorial [3].

Bibliography

[1] J. Benedetto, *The theory of multiresolution analysis frames and applications to filter banks*, J. of Applied Computation and Harmonic Analysis, 5 (1998), pp. 389–427.

[2] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, *Frame-theoretic analysis of oversampled filter banks*, IEEE Trans. Signal Processing, 46 (1998), pp. 3256–3268.

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

[4] ——, *Geometry and invariance in kernel based methods*, in Advances in Kernel Methods – Support Vector Learning, B. Schölkopf, C. Burges, and A. J. Smola, eds., Cambridge, MA, 1999, pp. 89–116.

[5] C. Burges and B. Schölkopf, *Improving the accuracy and speed of support vector learning machines*, in Advances in Neural Information Processing Systems, M. Mozer, M. Jordan, and T. Petsche, eds., MIT Press, 1997, pp. 375–381.

[6] R. R. Coifman and D. Donoho, *Translation invariant de–noising*, in Wavelet in Statistics, A. Antoniadis, ed., Lecture Notes, Springer Verlag, New York, 1995, pp. 125–150.

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

[8] Z. Cvetković and M. Vetterli, *Oversampled filter banks*, IEEE Trans. on Signal Processing, 46 (1998), pp. 1245–1255.

[9] P. F. C. de Rivaz, *Complex Wavelet Based Image Analysis and Synthesis*, PhD thesis, Department of Engineering, University of Cambridge, March 2001.

[10] T. Evgeniou, M. Pontil, and T. Poggio, *Regularization networks and support vector machines*, Advances in Computational Mathematics, 1 (2000), pp. 1–50.

[11] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, New York, 1990.

[12] M. S. Floater and A. Iske, *Multistep scattered data interpolation using compactly supported radial basis functions*, J. of Computational and Applied Mathematics, 73 (1996), pp. 65–78.

[13] S. Hoth, *Relationship between parameters of evoked otoacoustic emissions and hearing loss*, Audiologische Akustik, 1 (1995), pp. 20–29.

[14] A. Iske, *Hierarchical scattered data filtering for multilevel interpolation schemes*, in Mathematical Methods for Curves and Surfaces, T. Lyche and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, 2001, pp. 211–220.

[15] T. Joachims, *Making large–scale support vector machine learning practical*, in Advances in Kernel Methods – Support Vector Learning, B. Schölkopf, C. Burges, and A. J. Smola, eds., Cambridge, MA, 1999, pp. 169–184.

[16] D. T. Kemp, *Stimulated acoustic emissions from within the human auditory system*, J. Acoust. Soc. Am., 64 (1978), pp. 1386–1391.

[17] G. S. Kimeldorf and G. Wahba, *Some results on tchebycheffian spline functions*, J. Anal. Applic., 33 (1971), pp. 82–95.

[18] N. G. Kingsbury, *Complex wavelets for shift invariant analysis and filtering of signals*, Submitted to: J. of Applied Computation and Harmonic Analysis, (2000).

[19] J. Liang and T. W. Parks, *A translation–invariant wavelet representation algorithm with applications*, IEEE Trans. on Signal Processing, 44 (1996), pp. 225–232.

[20] J.-C. Pesquet, H. Krim, and H. Carfantan, *Time–invariant orthonormal wavelet representations*, IEEE Trans. on Signal Processing, 44 (1996), pp. 1964–1970.

[21] R. Probst, B. L. Lonsbury-Martin, and G. K. Martin, *A review of otoacoustic emissions*, J. Acoust. Soc. Am., 89 (1991), pp. 2027–2067.

[22] R. Schaback, *Creating surfaces from scattered data using radial basis functions*, in Mathematical Methods in Computer Aided Geometric Design III, M. Dæhlen, T. Lyche, and L. L. Schumaker, eds., Vanderbilt Univ. Press, 1995, pp. 477–496.

[23] ——, *Reconstruction of multivariate functions from scattered data*, Monograph, Institute for Numerical and Applied Mathematics, University of Göttingen, (1997).

[24] B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, *Input space vs. feature space in kernel-based methods*, IEEE Trans. on Neural Networks, 10 (1999), pp. 1000–1017.

[25] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, *Comparing support vector machines with gaussian kernels to radial basis function classifiers*, IEEE Trans. Signal Processing, 45 (1997), pp. 2758–2765.

[26] I. Selesnick, *The double-density dual-tree dwt*, Preprint, Department of Electrical Engineering, Polytechnic University, Brooklyn, NY, 2001.

[27] ——, *Smooth wavelet tight frames with zero moments*, J. Applied and Computational Harmonic Analysis, 10 (2001), pp. 163–181.

[28] M. Shensa, *The discrete wavelet transform: Wedding the à trous and mallat algorithms*, IEEE Trans. Signal Processing, 40 (1992), pp. 2464–2482.

[29] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Hegger, *Shiftable multiscale transforms*, IEEE Trans. on Information Theory, 38 (1992), pp. 587–608.

[30] A. J. Smola and B. Schölkopf, *Sparse greedy matrix approximation for machine learning*, in International Conference on Machine Learning, 2000.

[31] I. Steinwart, *On the influence of the kernel on the generalization ability of support vector machines*, Technical Report 01-01, 2001, Institut of Mathematics, University of Jena, 2001.

[32] G. Tognola, F. Grandori, and P. Ravazzani, *Wavelet analysis of click–evoked otoacoustic emissions*, IEEE Trans. on Biomedical Engineering, 45 (1998), pp. 686–697.

[33] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, NY, 1995.

[34] G. Wahba, *Support vector machines, reproducing kernel hilbert spaces and the randomized GACV*, in Advanced in Kernel Methods – Support Vector Learning, B. Schölkopf, C. Burges, and A. J. Smola, eds., Cambridge, MA, 1999, pp. 293–306.

[35] H. Wendland, *Piecewise polynomial, positive definite functions and compactly supported functions radial basis functions of minimal degree*, Adv. in Comp. Math., 4 (1995), pp. 389–396.