Feasible Adaptation Criteria
for Hybrid Wavelet - Large Margin Classifiers

Julia Neumann, Christoph Schnörr, Gabriele Steidl

Dept. of Mathematics and Computer Science
University of Mannheim
D-68131 Mannheim, Germany

{jneumann,schnoerr,steidl}@uni-mannheim.de

December 22, 2002

Abstract

In the context of signal classification, this paper assembles and compares criteria to easily judge the discrimination quality of a set of feature vectors. The quality measures are based on the assumption that a Support Vector Machine is used for the final classification. Thus, the ultimate criterion is a large margin separating the two classes. We apply the criteria to control the feature extraction process for signal classification. Adaptive features related to the shape of the signals are extracted by wavelet filtering followed by a nonlinear map. To be able to test many features, the criteria are easily computable while still reliably predicting the classification performance.

We also present a novel approach for computing the radius of a set of points in feature space. The radius, in relation to the margin, forms the most commonly used error bound for Support Vector Machines. For isotropic kernels, the problem of radius computation can be reduced to a common Support Vector Machine classification problem.

1 Introduction

This paper addresses the problem of how to choose an orthogonal compactly supported wavelet to optimally preprocess signals for binary classification. For that purpose, several criteria to judge the discrimination ability of a set of feature vectors are presented and closely examined.

The problem we treat here is to assign a class label to signals that are originally divided into two classes. Therefore, one assumes that labelled training samples are given in advance. The classification problem emerges in the one-dimensional case for medical applications and acoustic signals and in the two-dimensional case for texture images, for example. Some sample signals for the detection of ventricular tachycardia as a medical application are shown in Figure 1. Our wavelet adaptation approach improves the classification accuracy compared to commonly used algorithms for this important problem.

Instead of trying to classify signals directly, i.e. taking single pixels as 'features', a preprocessing step extracting relevant features from the data commonly relies on filter banks ([1, 13, 28, 29, 31]).

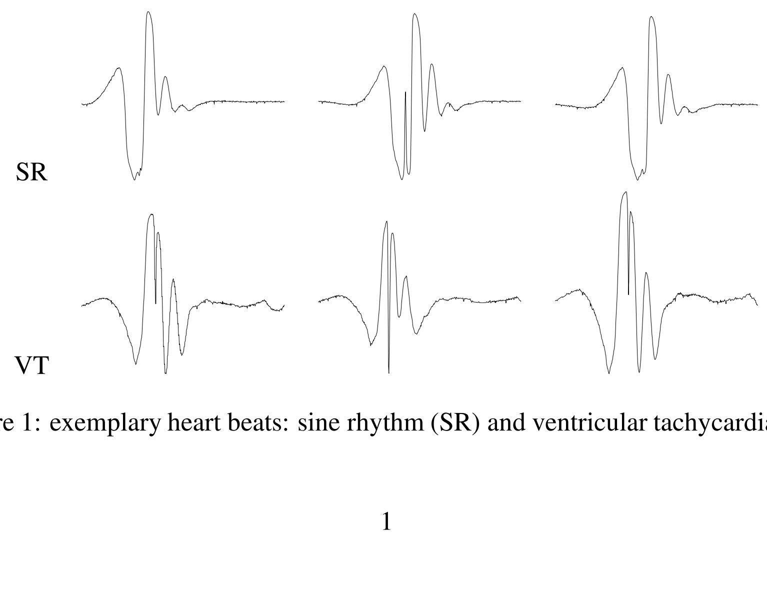


Figure 1: exemplary heart beats: sine rhythm (SR) and ventricular tachycardia(VT)

Those filtering approaches are closely related to wavelet decomposition as used in [27, 43, 23]. To generate low-dimensional feature vectors, we propose to use the norm of the coefficients of the different frequency bands for classification as done in [43].

Still there remains an important question to be answered. Which wavelet should one utilise to perform the wavelet decomposition? As the signal types vary as much as from cardiac signals to texture images, different waveforms are encountered in the classification problem. As already shown by one of the authors in [41], an adaptation of the applied wavelet to the classification problem at hand potentially heavily increases classification performance. Jones et al. ([23]) also claim that a problem specific wavelet choice is promising. In general, it is desirable to adapt the preprocessing or some classifier parameters to the specific classification problem. This brings up the problem of finding a criterion for the wavelet design.

As to the final classification, there are many possible classifier choices. The 'Support Vector Machine' (SVM) is an algorithm to solve two-class classification problems. The SVM as a relatively new tool is described in [46, 6, 32] and is already widely accepted due to its simplicity and flexibility. The approach is based on *Structural Risk Minimisation* ([45]) and is a generalised linear classifier that tries to maximise the margin between the two classes. There are two variants of the classifier concerning its invariance to noise. The 'hard margin SVM' claims that all training points are separated by the hyperplane with maximal margin. The noise insensitive variant, the 'soft margin SVM' allows some outliers falling within the margin ([6]). In practise, the SVM is also used for multi-class classification problems, most effectively by using a sequence of binary SVM classifiers to find the likeliest class ([18, 21]).

It is obvious that adequate adaptation criteria can be obtained from the classifier's objectives and derived classification error bounds. The most frequently used error bound for SVM classifiers is the radius - margin bound ([45]), where the margin is the objective of the SVM. We will show that, for a special class of kernels, the radius of the smallest sphere enclosing the feature vectors, the second quantity used in the bound, can be computed by solving another standard SVM problem again. This bound has for example been successfully applied by Weston et al. [49] to apply gradient descent methods for feature selection. But despite the computational convenience coming from our radius computation problem reduction, this method is not applicable here as it still includes repeated minimisation of quadratic programs. This is computationally very expensive because wavelet adaptation criteria typically have many local minima and hence need to be evaluated many times for different parameter values. This brings up the problem of finding reliable criteria that are still fast to evaluate to rank a given feature set. We will compare some simple criteria for the example of the one-dimensional signal classification by wavelet decomposition. The results apply to the two-dimensional setting as well because wavelet decomposition can be separably applied to each dimension and the resulting wavelet features then may be used to analyse texture in images, for instance ([31, 27]). The proposed criteria can also be used for feature selection, which aims at discarding features from a predetermined set. Apart from a computational gain, this may also improve classifier accuracy ([4, 49]).

Our experiments show that there exist simple criteria that well approximate the classification error, assessed by the radius - margin error bound. Applied to our wavelet adaptation problem, these criteria establish an easy way to find the wavelet that best discriminates the signal classes.

We will first introduce the particular classification problem we are interested in. To this end, Section 2 will deal with the feature extraction process, especially with the wavelet parameterisation, and Section 3 will present the SVM classifier to be applied. Motivated by the measures of quality for the SVM, we will then come to the main part. Section 4 presents selected criteria for feature adaptation and discusses their relations and applicability. To see how the criteria fare in practise, Section 5 examines their imposed ranking for the problem of the wavelet choice. Finally, the results of the paper are summarised in Section 6. Additional and background material is given in the appendix. Appendices A and B give some mathematical backgrounds about filter banks and discrete wavelets and SVMs. Appendix C gives the interesting relation between the radius computation problem and the standard SVM classification problem.

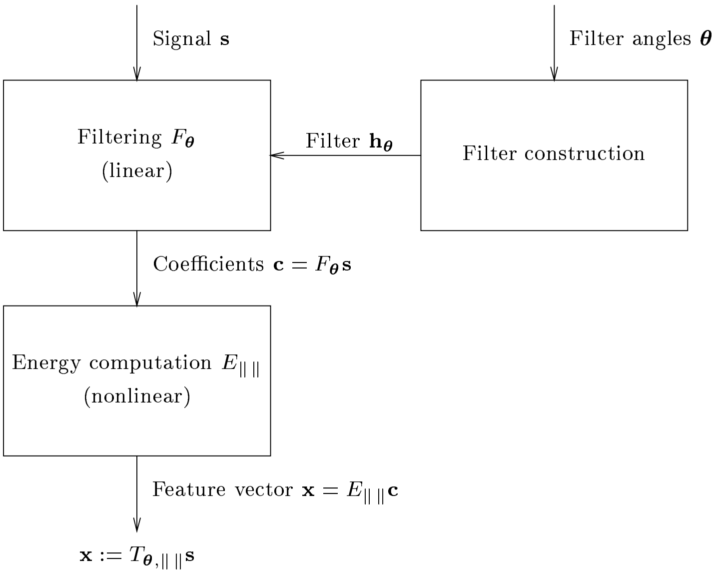


Figure 2: feature extraction: from the signal to the feature vector

Notation. Throughout the paper, we denote vectors and matrices by bold face lower and upper case letters, respectively. The matrix $\mathbf{I}$ denotes the identity matrix in appropriate dimensions. The vector $\mathbf{0}$ signifies a vector of zeros in the respective space, $\mathbf{e}$ a vector of ones. All vectors will be column vectors unless transposed by the superior symbol ^{T}. If $\mathbf{x} \in \mathbb{R}^n$ denotes a vector, in general, we will indicate its components by x_i ($i = 1, \ldots, n$). We assume vector inequalities coming up in optimisation problems to hold componentwise. Further, ℓ_2 denotes the Hilbert space of real valued quadratic summable sequences $\mathbf{a} = (a_i)_{i\in\mathbb{N}}$ with inner product $\langle \mathbf{a}, \mathbf{b}\rangle_{\ell_2} = \sum_{i\in\mathbb{N}} a_i b_i$ and corresponding norm $\|\mathbf{a}\|_{\ell_2} = \left(\sum_{i\in\mathbb{N}} a_i^2\right)^{1/2}$.

2 Feature Extraction by Discrete Wavelet Transform

This section describes the feature extraction process for our application of signal classification. We briefly summarise the mathematical definition of 'feature vectors' used in this paper, and their parametrisation. This provides the basis for feature adaptation and classification.

Figure 2 illustrates the feature extraction process from an input signal $\mathbf{s} \in \mathbb{R}^l$ to its corresponding feature vector $\mathbf{x} \in \mathbb{R}^d$, where $d \ll l$. The feature extraction process consists of two successive steps, namely filtering and energy computation of the bandpass coefficients. For the filtering we use orthonormal filter banks. As illustrated on the right hand side of the diagram, these filters can be determined by some filter angles $\boldsymbol{\theta}$ which will be the main parameters of our feature extraction process. Therefore the filtering operator is denoted by $F_{\boldsymbol{\theta}}$. Then the features are generated using the norm of the resulting coefficients at each decomposition level. As different norms $\|\ \|$ will be used, the corresponding operator is called $E_{\|\,\|}$. In summary, the feature extraction operator is given by $T_{\boldsymbol{\theta},\|\,\|} := E_{\|\,\|}F_{\boldsymbol{\theta}}$. The single steps will be more closely looked at in the following.

For filtering, we apply the concept of filter banks. For a short introduction see Appendix A.1. Fundamental for the parametrisation of our feature extraction process is the representation of orthonormal filter banks in a lattice structure composed of rotations and delays. According to [44, Theorem 14.3.1], [40, Theorem 4.7], a two-channel FIR filter bank with filter length $2L+2$ is orthogonal (paraunitary) if and only if, up to the sign of the high-pass filter, the corresponding polyphase matrix $\mathbf{H}_{\mathrm{pol}}(z)$ can be decomposed into

$$\mathbf{H}_{\mathrm{pol}}(z) = \left[\prod_{l=0}^{L-1} \begin{pmatrix} \cos\theta_l & \sin\theta_l \\ -\sin\theta_l & \cos\theta_l \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix}\right] \begin{pmatrix} \cos\theta_L & \sin\theta_L \\ -\sin\theta_L & \cos\theta_L \end{pmatrix}, \tag{1}$$

where $\theta_L \in [0, 2\pi)$ and $\theta_l \in [0, \pi)$ $l = 0, \ldots, L-1$. If the filter bank's high-pass filter has at least one vanishing moment, the filter bank angles add up to $\frac{\pi}{4}$ (see also [40, Theorem 4.6]):

$$\sum_{l=0}^{L} \theta_l = \frac{\pi}{4} \pmod{2\pi}.$$

The resulting parameter space $\{\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{L-1}) : \theta_l \in [0, \pi)\ l = 0, \ldots, L-1\}$ is π-periodic as the angles θ_i can be interpreted as rotation angles: A rotation of $\theta_i + \pi$ implies half a rotation extra because the result just alters sign. As the last angle is computed as $\theta_L = \frac{\pi}{4} - \sum_{i=0}^{L-1} \theta_i$, it is that amount smaller. Hence, the final filter output is just the same as before.

In our experiments we will assume a filter length of 6 corresponding to $L = 2$, i.e., a two-dimensional parameter space. In most instances, according to our experiments filters of length 6 are sufficient to describe the waveforms and they have the advantage that they are conveniently depicted in 2D for comparison.

For a given $\boldsymbol{\theta}$ in the parameter space, the synthesis filters are determined by the polyphase matrix. Moreover these filters provide us with an orthonormal basis

$$\{\tilde{\mathbf{v}}^{\mathbf{J}}_{\mathbf{k}} : k = 1, \ldots, \frac{l}{2^J}\} \cup \{\tilde{\mathbf{w}}^{\mathbf{j}}_{\mathbf{k}} : j = 1, \ldots, J; k = 1, \ldots, \frac{l}{2^j}\}$$

of $\mathbb{R}^l$ consisting of the periodic discrete–time scaling sequences $\tilde{\mathbf{v}}^{\mathbf{J}}_{\mathbf{k}}$ and discrete–time wavelets $\tilde{\mathbf{w}}^{\mathbf{j}}_{\mathbf{k}}$ (see Appendix A.2.) The representation of the signal $\mathbf{s} \in \mathbb{R}^l$ with respect to this basis is given by

$$\mathbf{s} = \sum_{k=1}^{l/2^J} c^J_k \tilde{\mathbf{v}}^{\mathbf{J}}_{\mathbf{k}} + \sum_{j=1}^{J} \sum_{k=1}^{l/2^j} d^j_k \tilde{\mathbf{w}}^{\mathbf{j}}_{\mathbf{k}}$$

with the wavelet coefficients

$$\begin{aligned}
\mathbf{c}^{\mathbf{J}} &:= (c^J_k)_{k=1,\ldots,l/2^J} := (\langle \mathbf{s}, \tilde{\mathbf{v}}^{\mathbf{J}}_{\mathbf{k}} \rangle_{\ell_2})_{k=1,\ldots,l/2^J}\,, \\
\mathbf{d}^{\mathbf{j}} &:= (d^j_k)_{k=1,\ldots,l/2^j} := (\langle \mathbf{s}, \tilde{\mathbf{w}}^{\mathbf{j}}_{\mathbf{k}} \rangle_{\ell_2})_{k=1,\ldots,l/2^j}\,,\ j = 1, \ldots, J.
\end{aligned}$$

The filtering operator $F_{\boldsymbol{\theta}}$ then performs a wavelet analysis by an octave–band filter bank:

$$F_{\boldsymbol{\theta}} : \mathbb{R}^l \to \mathbb{R}^l\,,\ \mathbf{s} \mapsto \begin{pmatrix} \mathbf{c}^{\mathbf{d}} \\ \mathbf{d}^{\mathbf{d}} \\ \vdots \\ \mathbf{d}^{\mathbf{1}} \end{pmatrix} = \begin{pmatrix} (\langle \mathbf{s}, \tilde{\mathbf{v}}^{\mathbf{J}}_{\mathbf{k}} \rangle_{\ell_2})_{k=1,\ldots,l/2^d} \\ (\langle \mathbf{s}, \tilde{\mathbf{w}}^{\mathbf{J}}_{\mathbf{k}} \rangle_{\ell_2})_{k=1,\ldots,l/2^d} \\ \vdots \\ (\langle \mathbf{s}, \tilde{\mathbf{w}}^{\mathbf{1}}_{\mathbf{k}} \rangle_{\ell_2})_{k=1,\ldots,l/2} \end{pmatrix} = \mathbf{F}_{\boldsymbol{\theta}} \mathbf{s}.$$

The matrix $\mathbf{F}_{\boldsymbol{\theta}} \in \mathbb{R}^{l \times l}$ is orthogonal and consequently preserves the norm of the signals

$$||\mathbf{F}_{\boldsymbol{\theta}} \mathbf{s}||_{\ell_2} = ||\mathbf{s}||_{\ell_2}. \tag{2}$$

To generate a handy number of features that still make the signals well distinguishable, we introduce the energy operator

$$E_{||\ ||} : \mathbb{R}^l \to \mathbb{R}^d\,,\ \begin{pmatrix} \mathbf{c}^{\mathbf{d}} \\ \mathbf{d}^{\mathbf{d}} \\ \vdots \\ \mathbf{d}^{\mathbf{1}} \end{pmatrix} \mapsto \begin{pmatrix} ||\mathbf{d}^{\mathbf{d}}|| \\ \vdots \\ ||\mathbf{d}^{\mathbf{1}}|| \end{pmatrix},$$

where we restrict our attention to two significant norms $||\ ||$, namely the ℓ_2-norm and the weighted ℓ_2-norm

$$\mathbf{c} \mapsto \frac{1}{\sqrt{n}} ||\mathbf{c}||_{\ell_2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} c_i^2}, \qquad \mathbf{c} \in \mathbb{R}^n.$$

The weighted ℓ_2-norm seizes the average power or channel variance as proposed by Unser ([43]). The normalisation balances the contribution of the different channels and, as its expectation is independent of the coefficient vector length, makes the feature values for different size signals comparable. In the translation invariant case, if the applied high-pass filter has at least one vanishing moment, the expectation of the coefficients for each high-pass channels is zero, so that the above norm effectively represents the channel variance. For sub-sampled decomposition, this choice of the norm still provides a variance estimate. Besides the two proposed norms, the ℓ_p-norms for $p \geq 1$ may also be chosen. Especially the ℓ_1-norm behaves robustly with respect to different signals ([41]).

Now, for a signal $\mathbf{s} \in \mathbb{R}^l$, the corresponding *feature vector* $\mathbf{x}$ is defined by

$$\mathbf{x} := T_{\boldsymbol{\theta},\|\,\|}\mathbf{s} = E_{\|\,\|}F_{\boldsymbol{\theta}}\mathbf{s}.$$

The norm preserving property (2) implies

$$\|T_{\boldsymbol{\theta},\|\,\|}\mathbf{s}\|_{\ell_2} \leq \|\mathbf{s}\|_{\ell_2}. \tag{3}$$

As a consequence, if we deal with ℓ_2-normed input signals $\mathbf{s}$, then the feature vectors lie within or on a sphere in $\mathbb{R}^d$ centred at the origin. In our experiments we deal with signals having average value zero and apply the full wavelet decomposition, i.e., $l/2^d = 1$. Then it is easy to check that $\mathbf{c}^d = 0$ (see Lemma 1 in Appendix A.2). If we further use the ℓ_2-norm in $E_{\|\,\|}$, then we have equality in (3).

In the rest of the paper, we will drop the subscript ℓ_2 for norms and inner products if that does not cause confusion.

The relationship just mentioned reveals some important structure of the feature vector set. But to rate a set of feature vectors according to their classification ability, it is essential to take into account the classifier in use. The Support Vector Machine, which will be described next, intends to maximise the 'margin' between the feature vectors of both classes in some 'feature space'. The classifiers target term, the margin as well as potential classification error bounds may motivate possible adaptation criteria.

3 Support Vector Machine Classification

In this section we give a short introduction to Support Vector Machine classification.

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

$$\mathcal{Z} := \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\} : i = 1, \ldots, n\} \tag{4}$$

of n 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. f classifies the training data correctly if $\operatorname{sgn}(f(\mathbf{x}_i)) = t(\mathbf{x}_i) = y_i$ for all $i = 1, \ldots, n$. Here

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

Support Vector Machine Classification combines the simplicity of a linear learning machine with the high generalisation ability only found in nonlinear classifiers. To this end, the so–called *feature map* $\boldsymbol{\phi} : \mathcal{X} \to \ell_2$ non-linearly mapping the input vectors into some generally higher-dimensional space. We will then search for f as a linear function in the feature vectors. The mathematical details with respect to the nonlinear feature map and the final optimisation problem are provided in the appendix.

It is possible to state linear learning machines only in terms of inner products between the input vectors. As only inner products need to be evaluated and we would like to have fast computation, one can directly compute the inner products instead of explicitly carrying out the feature map. This is done by means of a kernel function. The kernel function K induces a 'reproducing kernel Hilbert space'

$\mathcal{H}_K$ defined in Appendix B.1, where the details concerning the feature map and kernel functions are given. In our applications we will use Gaussian kernels

$$K(\mathbf{x}, \mathbf{y}) = e^{-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}} \tag{5}$$

which are known to have reasonable performance ([36]).

Let us turn to our classification task. For a given training set (4) we intend to construct a function $f \in \mathcal{H}_K$ which minimises

$$C \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \frac{1}{2} ||f||^2_{\mathcal{H}_K} \tag{6}$$

where

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

for some constant $C \in \mathbb{R}_+$ controlling the trade-off between the approximation error and the regularisation term. For the choice $C = \infty$, the resulting classifier is called *hard margin SVM*, otherwise *soft margin SVM*. The 'margin' is the minimal distance of a training point $\mathbf{x}_i$ to the hyperplane separating both classes for a linear classifier.

As argued in Appendix B.2, the minimiser of (6) can be found by solving the following quadratic problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \; & -\frac{1}{2} \boldsymbol{\alpha}^T \mathbf{YKY} \boldsymbol{\alpha} + \mathbf{e}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{e}, \end{aligned} \tag{7}$$

where $\mathbf{Y} := \operatorname{diag}(y_1, \ldots, y_n)$ and the kernel matrix $\mathbf{K} := (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ represents the inner products of the feature vectors.

The *support vectors* (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 $I := \{i \in \{1, \ldots, n\} : \alpha_i \neq 0\}$ then the function f has the representation

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

which only depends on the SVs.

We obtain the maximal margin from the solution by

$$\rho = \left(\sum_{i \in I} y_i \alpha_i f(\mathbf{x}_i) \right)^{-1/2}.$$

In case of hard margin classification, as reasoned in the appendix, this expression simplifies to

$$\rho = \left(\sum_{i \in I} \alpha_i \right)^{-\frac{1}{2}}. \tag{8}$$

4 Criteria for Feature Adaptation

So far, an application problem, a feature extraction method and the SVM classifier have been described. This main section will propose several criteria for rating sets of feature vectors according to their classification capability with respect to an SVM classifier. In the following, the criteria will be presented and some properties and relationships between them will be indicated.

Here we consider the task of having to rate sets of labelled feature vectors $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\} \subset \mathcal{X} \times \{-1, 1\}$, $\mathcal{X} \subset \mathbb{R}^d$. For our signal classification problem especially, all feature vectors $\mathbf{x}_i$ lie in or even on a sphere centred at the origin. The goal is to find a measure that allows for fast comparison of different sets of feature vectors based on maximising the classifier performance.

Possible criteria for adaptation are obtained by bounds for the *generalisation error* $\mathrm{err}(f)$ and approximations thereof, i.e., the probability that $\mathrm{sgn}(f_{\mathbf{w}^*}(\mathbf{x})) \neq y$ for a randomly chosen example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, 1\}$. Although there exist many proven bounds for the error risk or its expectation in the literature (see, for example, [5, 19]), in essence, most of them rely either on the number of support vectors ([45, Theorem 5.2], [16] and [19, Section 5.2.1]) or on the size of the margin separating the classes normalised by a measure of the feature vector variation such as their radius ([45, Theorem 5.2], [20]). This means that minimising the error bound is equivalent to maximising the margin resp. minimising the number of support vectors. Unfortunately, both objectives imply solving a quadratic program which, for our purpose, is impracticable. Besides, the resolution of error bounds relying on the number of support vectors is too low. Since only a few values are possible, many settings may not be comparable. Some additional quantities used in the error bounds are for example the eigenvalues of the kernel matrix ([34]) or the normalised margin ([20]). However, they suffer from the same computational costs as do margin and number of support vectors. Moreover, many bounds are not tight, e.g., the stability bound proposed by Bousquet and Elisseeff ([3, 19]). At the worst, if the bounds value is above 1, one cannot say that a decrease improves the expected classifier performance as there is no conclusion at all possible. This motivates to evaluate the performance of simplified criteria which can be more efficiently evaluated.

In the following, some possible adaptation criteria for the choice of the optimal filter are presented and discussed.

4.1 Margin

For the hard margin case, the margin ρ itself (obtained by equation (8) from the solution of the optimisation problem (7) in Section 3) as the SVM objective criterion may be a first guess for a criterion for the wavelet choice. Indeed, our experiments indicate that if training and test data have the same underlying distribution, the margin behaves much like the classification error. The major disadvantage of taking the margin as adaptation criterion is that for every possible wavelet to examine, a quadratic optimisation problem has to be solved. As the optimal wavelet can only be found by search heuristics due to the complexity of the feature extraction process, i.e., the multi-level wavelet transform and energy computation, and the resulting non-convex objective function, the margin criterion will be a time-consuming criterion. Furthermore, the size of the margin depends only on few data points, precisely on the support vectors. Thus, the size of the margin is not a 'smooth' function of all input vectors. The same main drawback, the complexity of the evaluation, holds for the soft margin optimisation criterion $C \sum_{i=1}^{n} \xi_i + \frac{1}{2}||\mathbf{w}||^2_{\mathcal{F}_K}$, the equivalent of the margin, even though the optimisation functional in the soft margin case is smoother because of the limited influence of single points (cf. the dual constraint $\boldsymbol{\alpha} \leq C\mathbf{e}$ in (7)).

4.2 Radius-Margin

According to [46, Theorem 10.6], the expectation of the quotient

$$\frac{1}{n} \frac{R^2}{\rho^2} \tag{9}$$

forms an upper bound on the SVMs generalisation error, where expectation is meant over all training samples with equal size assuming the same underlying distribution. Thereby, R is the radius of the smallest sphere enclosing the points in feature space and can be computed by a single-class SVM, i.e., another quadratic program. See Appendix C for a detailed derivation of the final computation algorithm in Theorem 2.

In the soft margin case, there also exists a radius margin bound. According to [11], the expectation of the generalisation error of the SVM is bounded from above by the expectation of the term

$$\frac{1}{n}(4R^2 \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \xi_i)$$

where $\boldsymbol{\alpha}$ comes from the solution of the problem (7) and $\xi_i := (1 - y_i f(\mathbf{x}_i))_+$ is the resulting error term from problem (6).

Due to the property (3), if the SVM input vectors are normalised, the radius R is bounded. Consequently, the margin or the soft margin minimisation functional by themselves provide an error bound and a justified criterion.

Although this quantity is quite meaningful, it suffers from the same problem as the margin or even worse: For every criterion evaluation, two quadratic optimisation problems have to be solved. But as the bound is relatively tight to the error (most evaluations at least lead to bounds below the trivial $\frac{1}{2}$, see Section 5), we selected it for comparison as a representative for all error bounds which are close to the generalisation error but are too computationally intensive for feature adaptation.

4.3 Alignment

For kernel problems, a new characteristic has been proposed ([8, 25]): The sample alignment

$$\hat{A}(\mathbf{K}_1, \mathbf{K}_2) := \frac{\langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1 \rangle_F \langle \mathbf{K}_2, \mathbf{K}_2 \rangle_F}}$$

was proposed as a measure of conformance between kernels. Especially, the kernel matrix $\mathbf{y}\mathbf{y}^T$, where $\mathbf{y}$ denotes the vector of class labels, is viewed as the optimal kernel matrix for two-class classification. This leads to maximising the criterion

$$\hat{A}(\mathbf{K}, \mathbf{y}\mathbf{y}^T) = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^T \rangle_F}{n\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F}} = \frac{\mathbf{y}^T \mathbf{K} \mathbf{y}}{n\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F}}$$

which, by the inequality of Cauchy-Schwarz, only takes values in $[0, 1]$ as the kernel matrix is always positive definite.

According to all our experiments, its denominator $n\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F}$ doesn't influence the alignment much for the Gaussian kernel with a fixed kernel width. The change of the alignment is dominated by its numerator $\langle \mathbf{K}, \mathbf{y}\mathbf{y}^T \rangle_F$ that varies about 200% whereas its denominator only varies about 20% for different wavelets. This may result from the norm preservation (3) which implies that the kernel matrix is more or less normalised. For normed feature vectors $\|\mathbf{x}_i\| = c\,,\ i = 1, \ldots, n$ as guaranteed by using the ℓ_2-norm for the energy computation, the Gaussian kernel reads

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}} \\ &= e^{-\frac{\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{2\sigma^2}} \\ &= e^{-\frac{c^2}{\sigma^2}} e^{\frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\sigma^2}}\,, \quad i, j = 1, \ldots, n, \end{aligned}$$

where the first term is a constant and the second one is just the exponential of the linear kernel. First, the exponential e^x is a monotone function of x, and second, it can be approximated for small $x < 1$ by $1 + x$ according to its Taylor series

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + x^2 \sum_{n=0}^{\infty} \frac{x^n}{(n+2)!}.$$

Thus, if σ is large which means that the exponent is small, the linear approximation is close to the exponential. This implies that the alignment with the Gaussian kernel is related to the alignment with the linear kernel

$$K_{linear}(\mathbf{x}, \mathbf{y}) := \langle \mathbf{x}, \mathbf{y} \rangle\,,\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

with feature map $\boldsymbol{\phi}_{linear} := \mathrm{id}$ or to the quantity

$$\langle \mathbf{D}, \mathbf{y}\mathbf{y}^T \rangle_F.$$

with $\mathbf{D} := (\langle \mathbf{x}_i, \mathbf{x}_j \rangle)_{i,j=1}^n$ the kernel matrix with respect to the linear kernel, the analogue of $\mathbf{K} = (\langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle_{\mathcal{F}_K})_{i,j=1}^n$.

There is also a result which bounds the generalisation accuracy of the expected Parzen window estimator by a function of the alignment: By [8, Theorem 4]

$$\mathrm{err}(f) \leq 1 - \hat{A}(\mathbf{K}, \mathbf{y}\mathbf{y}^T) + \hat{\epsilon} + \frac{1}{\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F}}$$

with probability greater than $1 - \delta$, where $\hat{\epsilon}$ is a function of the sample and the level of significance δ. The parameter δ and the term $\hat{\epsilon}$ are only needed because the sample alignment is used instead of its expected value. According to [8], when using the true alignment $A(k_1, k_2) := \frac{\langle k_1, k_2 \rangle_P}{\sqrt{\langle k_1, k_1 \rangle_P \langle k_2, k_2 \rangle_P}}$ where the inner product is defined as $\langle f, g \rangle_P := \int_{\mathcal{X}^2} f(\mathbf{x}, \mathbf{z}) g(\mathbf{x}, \mathbf{z}) dP(\mathbf{x}) dP(\mathbf{z})$, the bound even simplifies to $\mathrm{err}(f) \leq 1 - A(k, t(\cdot)t(\cdot))$ (where t is the target function as defined in Section 3). This shows that the alignment is directly related to the expected Parzen window estimator. Cristianini et al. ([8]) claim that, as the empirical Parzen estimator is concentrated, its generalisation is described by the empirical alignment $\hat{A}$ as well. The Parzen window estimator is related to a SVM. It is equivalent to a soft margin SVM with minimal outlier penalisation parameter $C = \frac{1}{n}$. This establishes the choice of the alignment as an adaptation criterion especially for soft margin SVMs, but as we will show in the next section, the alignment reliably predicts the margin for our SVM problems without outliers as well.

4.4 Class Centre Distance

Motivated by the alignments relation to linear quantities, we want to look at criteria in the original data space $\mathbb{R}^d$. Strauß and Steidl proposed in [41] to maximise the distance of the two class centres in the Euclidean metric in the original space $\mathbb{R}^d$. Denoting by n_i the number of samples and by $\boldsymbol{\mu}_i$ the mean of class $i = 1, -1$, i.e., $n_i := |\{j : y_j = i\}|$ and $\boldsymbol{\mu}_i := 1/n_i \sum_{y_j = i} \mathbf{x}_j$, then the criterion is given by

$$D := \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}\|.$$

For an equal number of training vectors for both classes, the criterion is equivalent to

$$\begin{aligned}
D^2 &= \| \sum_{\{i : y_i = 1\}} \mathbf{x}_i - \sum_{\{i : y_i = -1\}} \mathbf{x}_i \|^2 \\
&= \| \sum_{i=1}^n y_i \mathbf{x}_i \|^2 \\
&= \langle \sum_{i=1}^n y_i \mathbf{x}_i, \sum_{i=1}^n y_i \mathbf{x}_i \rangle \\
&= \sum_{i,j=1}^n y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
&= \langle \mathbf{D}, \mathbf{y}\mathbf{y}^T \rangle_F
\end{aligned}$$

with $\mathbf{D}$ again the matrix of inner products in data space. The quantity D is the replacement of the alignments numerator in data space. In effect, it approximates thus a substitute for the alignment in data space and even the alignment itself, as discussed in the previous section.

As argued in [41], for a normalised isotropic kernel that is monotonically decreasing in the arguments' Euclidean distance, the distance between two points in feature space is maximised if their distance in data space is maximised. For this class of kernels (including, e.g., the Gaussian kernel), this property hints why the criteria in feature space are related to their substitutes in data space. Especially, the true alignment may be close to the class centre distance.

Analogous to the alignment, the class centre distance also has an upper bound. According to (3), the class centre distance is then bounded by two times the norm of the original signals

$$D \leq 2 \max_{i \in \{1, \dots, n\}} \|\mathbf{x}_i\| \leq 2 \max_{i \in \{1, \dots, n\}} \|\mathbf{s}_i\|.$$

Apart from the simple criterion evaluation that comes from the plain form of D, the criterion is also easily differentiable. This may be a crucial point from the perspective of optimisation.

4.5 Scatter Measures

The class centre distance only takes into account the mean values of the classes. Obviously, we look for classes that are distant from each other and at the same concentrated around their means. A generalisation are measures using scatter matrices as described in [42, Section 5.5.3]:
The *within-class scatter matrix* describes the average feature variance in the classes and is defined as

$$\mathbf{S}_w := \frac{1}{n} \sum_{i\in\{-1,1\}} \sum_{y_j=i} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T.$$

The scattering of the whole classes is described by the *between-class scatter matrix*. Denoting by $\boldsymbol{\mu} := \sum_{i\in\{-1,1\}} \frac{n_i}{n}\boldsymbol{\mu}_i$ the common mean vector, the matrix is defined as

$$\mathbf{S}_b := \sum_{i\in\{-1,1\}} \frac{n_i}{n}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T.$$

Combining both scattering dimensions, the *mixture scatter matrix*

$$\mathbf{S}_m := \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^T$$

describes the common covariance. As a consequence, $\mathbf{S}_m = \mathbf{S}_w + \mathbf{S}_b$ holds.

Using these matrices, separability measures are defined that use the relation of $\mathbf{S}_m$ resp. $\mathbf{S}_b$ to $\mathbf{S}_w$. This can be done by maximising the quotient of either their traces or their determinants.

The special case of the measures $\frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}$ and $\frac{|\mathbf{S}_b|}{|\mathbf{S}_w|}$ for equiprobable classes in one dimension yields the Fisher discriminant ([14])

$$\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

for two classes $i = 1, -1$ with mean μ_i and scatter σ_i^2.

In the multi-dimensional case, using the determinant poses more computational requirements than the trace representing only the variances. Moreover, the single feature vector components are connected by the norm constraint anyway, so it plausible to ignore their correlation. Hence, for our two-class problem, for example the criterion $\frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}$ leads to

$$\begin{aligned} S :=& \frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)} \\ =& (\frac{n_1}{n}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}\|^2 + \frac{n_{-1}}{n}\|\boldsymbol{\mu}_{-1} - \boldsymbol{\mu}\|^2)/(\frac{n_1}{n}\sum_{k=1}^{d}\sigma_{1k}^2 + \frac{n_{-1}}{n}\sum_{k=1}^{d}\sigma_{-1k}^2), \end{aligned}$$

where σ_{ik}^2 denotes the marginal variance of class i along dimension k. For equiprobable classes, this simplifies to a multiple of

$$\begin{aligned} S &\propto \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}\|^2 / \sum_{k=1}^{d}(\sigma_{1k}^2 + \sigma_{-1k}^2) \\ &= D^2 / \sum_{k=1}^{d}(\sigma_{1k}^2 + \sigma_{-1k}^2) \end{aligned}$$

which is the class centre distance divided by a variance term. The variances serve to make the measure independent of the scaling as well as to include the classes' scattering. This criterion $\frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}$

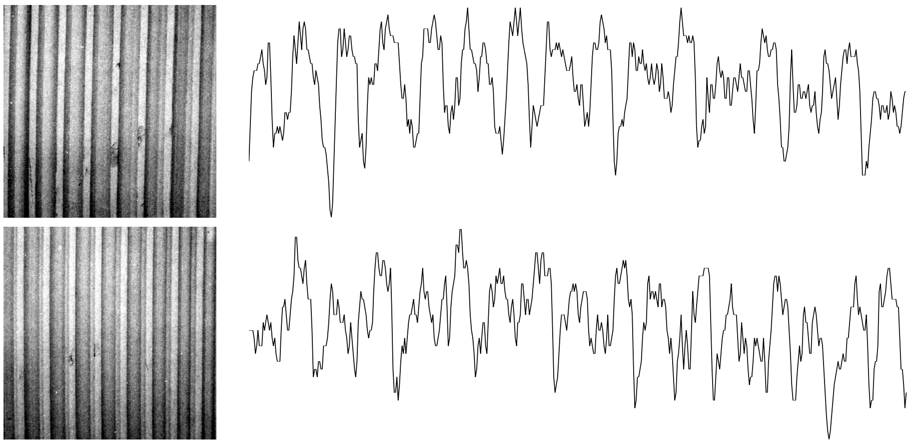


Figure 3: texture sample: linearly rescaled images and exemplary rows

also determines the choice of the projection for feature extraction by Linear Discriminant Analysis, whereas in the multi-dimensional case, the Multiple Discriminant Analysis, one uses the determinants of these matrices $\frac{|\mathbf{S}_b|}{|\mathbf{S}_w|}$ ([12, Chapter 3.8.3]). Note that in theory this criterion is unbounded in case of zero variance which, however, rarely happens in practise.

This section proposed many different criteria in feature and in data space, some of them directly related to generalisation error bounds. The usefulness of the bounds for feature adaptation still remains to be shown. To this end, the next section will show results for the evaluation of the criteria for several real-world problems.

5 Empirical Criteria Comparison

In the previous section we have proposed several criteria for judging the discrimination ability of a set of feature vectors. Some connections between the criteria have already been identified. It is now interesting to see how these links show up when analysing the bounds for real data, especially how close the bounds are together and which ones approximate the true generalisation ability best.

We want to evaluate the proposed criteria for the application described in Section 2: One-dimensional signals are to be classified according to the norm of their wavelet coefficients at each level. More precisely, for the feature extraction out of the signals, we make a full wavelet decomposition (i.e., nine decomposition steps for signals of length 512) with sub-sampling. This generates as many features from the data as possible. As already described, we thereby omit the low-pass component. By appropriate signal preprocessing, as argued in Section 2, we can still guarantee the ℓ_2-norm of the feature vectors to stay constant. We will use the ℓ_2-norm as well as the weighted ℓ_2-norm for feature extraction. For the classification, a hard-margin SVM with Gaussian kernel with kernel width $\sigma = 100$ in eqn. (5) is used. (For the rationale of this parameter choice see section 5.3.) We use the values of the margin and the radius - margin bound for the hard margin SVM to evaluate the quality of the proposed bounds to reduce the number of parameters (namely, to fix C to a simple value).

We use different data sets to evaluate the criteria: The first classification problem is the detection of ventricular tachycardia from electro-physiological data. The samples used here were obtained by inducing ventricular tachycardia during examinations at the University Hospital of Homburg, Germany. Data segments of $10\ \mathrm{sec}$ duration were recorded, equally for periods of normal cardiac activity. The episodes have been filtered and single beats have been cut out within a time–frame of $256\,\mathrm{ms}$ resulting in waveforms $\mathbf{s} \in \mathbb{R}^{512}$. For each patient, eight beats from a single episode are used for classifier training. Some exemplary beats for the sample patient picked out here to illustrate the criteria behaviour are shown in the introduction in Figure 1.

The second group of data are real world texture images from the MeasTex collection [38]. We use

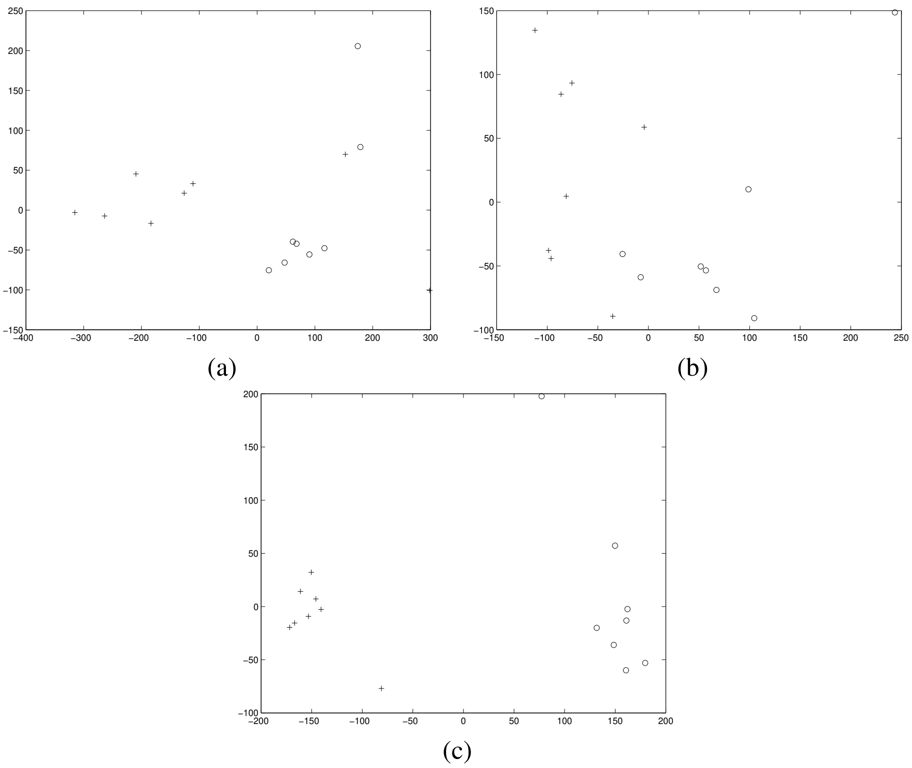


Figure 4: principal components of training vectors for sample heart patient and ℓ_2 norm: (a) for the Haar wavelet, (b) for the Daubechies wavelet with three vanishing moments, (c) for the optimally aligned wavelet

single rows of the texture images to have one-dimensional data structurally different from the cardiac data. The example that is shown in the paper indeed exhibits a one-dimensional structure. We use the two images of corrugated iron 'Misc.0002' and 'Misc.0003' here. Both images with normalised contrast as well as two exemplary rows are shown in Figure 3. The task is to classify which of two given textures the rows belong to. Here, the first 32 rows of each texture are used for classifier training.

Accounting for the properties of the feature extraction operator $T_{\boldsymbol{\theta},\|\,\|}$ described in Section 2, the original sample signals $\mathbf{s}_i$ $(i = 1, \ldots, n)$, cardiac data as well as texture image rows, have been ℓ_2-normalised according to $\|\mathbf{s}_i\| = 1000$ $(i = 1, \ldots, n)$ and their average signal value has been set to zero. Besides, the normalised margin bound for the SVM implies that all feature vectors ought to be normalised to obtain a high generalisation ability of the classifier.

5.1 Wavelet Adaptation Problem

The fundamental observation that has to be ensured at first is that the wavelet adaptation makes a significant difference. We illustrate below that wavelet feature adaptation may lead to a considerable increase of discriminatory power for real-world signal classification.

Therefore, we visualise the training data for the sample heart patient by extracting the principal two components of the nine-dimensional training vectors and visualising them. The Principal Components Analysis (PCA) projects the data from $\mathbb{R}^9$ to $\mathbb{R}^2$ by the projection that retains most of the total variance of the data.

The results for the Haar wavelet (parameters $(\theta_0, \theta_1) = (0, 0)$), the Daubechies wavelet with three vanishing moments ([9], parameters $(\theta_0, \theta_1) \approx (1.47, 0.50)$) and the wavelet that produces the optimal alignment with the kernel $\mathbf{y}\mathbf{y}^T$ (parameters $(\theta_0, \theta_1) = (2.04, 0.56)$) for the ℓ_2 norm are shown in Figure 4. The variance still contained in the plots is approximately 90%, 75% and 92% of the total variance, respectively.

This single example with few training data already shows that the wavelet choice heavily influences the classification performance: Neither the Haar wavelet, nor the Daubechies wavelet with three vanishing moments appear to make the training data linearly separable. The wavelet that achieves the maximal alignment, on the other hand, well separates the data (Figure 4 (c)). Moreover, the classes are nicely clustered now.

Indeed, for example for this patient with two further test episodes, the error for the weighted norm varies from 0 to 56% for different wavelets. Also, the optimum does not always lie in the same region. Even for different patients (but still the same problem class), the optimal wavelets differ heavily. As a consequence, utilising standard wavelets such as Haar or the Daubechies wavelet with three vanishing moments does not guarantee well-discriminating features and a small generalisation error.

5.2 Criteria Comparison

Motivated by the results of the previous section, next we evaluate and compare the criteria discussed in section 4. For this purpose, we will generate plots that show the criterion values subject to the two-dimensional wavelet parameter space. We will analyse the distance of the class centres D, the generalised Fisher criterion $S = \frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}$, the alignment with the kernel $\mathbf{y}\mathbf{y}^T$, the margin and the radius - margin classification error bound $\frac{1}{n}\frac{R^2}{\rho^2}$.

For the case of two filter bank parameters that corresponds to filters of length six, the adaptation criteria can be directly visualised over the parameter space. The resulting images for the criteria and the two problem samples heart data and texture rows are shown in Figures 5 and 6, respectively. Here, we use the weighted ℓ_2-norm for energy computation as we favour it over the ℓ_2-norm. The parameter space was discretised with 128 angles per dimension and for all parameter combinations, the feature vectors were computed by wavelet decomposition. Next, the criteria were be evaluated. The resulting values are plotted using a linear grey scale except for the radius - margin bound $R^2/n\rho^2$ which is plotted on a logarithmic scale due to its large variation. Additionally, the larger values are clipped to the trivial error bound 1 to enhance the contrast. To assess the effect of the clipping, the distribution of the logarithm of the bound is indicated by a histogram in Figures 5 and 6 (f). Light spots represent favourable criterion values in all criteria plots.

We want to examine the criteria when using the original ℓ_2-norm for energy computation as well. As for the particular heart patient, the plots for all criteria much resemble the ones for the weighted norm in this example, the ℓ_2-norm plots are not included. The corresponding ℓ_2-norm plots for the texture classification are given in Figure 7.

General Problem Some general properties are visible in the plots: The parameter space is apparently periodic in both parameters as argued in Section 2. Additionally, the parameter space seems to be structured since some characteristic lines appear in all criteria plots. Another parameter of the feature extraction is the filter length. All orthonormal filters of length four can be generated by a single parameter. Equivalently, all parameter combinations $(0, \theta)$ with $\theta \in [0, \pi)$ correspond to these filters. Regarding the first row of the plots, one can compare the difference between the values achieved there and on the whole parameter space. Only for the texture row classification with weighted ℓ_2-norm depicted in Figure 6, the optimal value on the whole parameter space differs significantly from the optimal value on the first row. For the other classification problems, there is already no systematic gain in augmenting the filter length from four to six.

Criteria Concerning the criteria, for all three figures, the first overall impression is that all shown criteria are alike. Moreover, all criteria show a detailed structure for the wavelet parameter space. This indicates that effectively finding the optimal wavelet according to the chosen criterion is not easy even for the simple criteria.

The class centre distance and particularly the alignment resemble the margin. That is, the wavelets that generate a high class centre distance or alignment also guarantee a large margin.

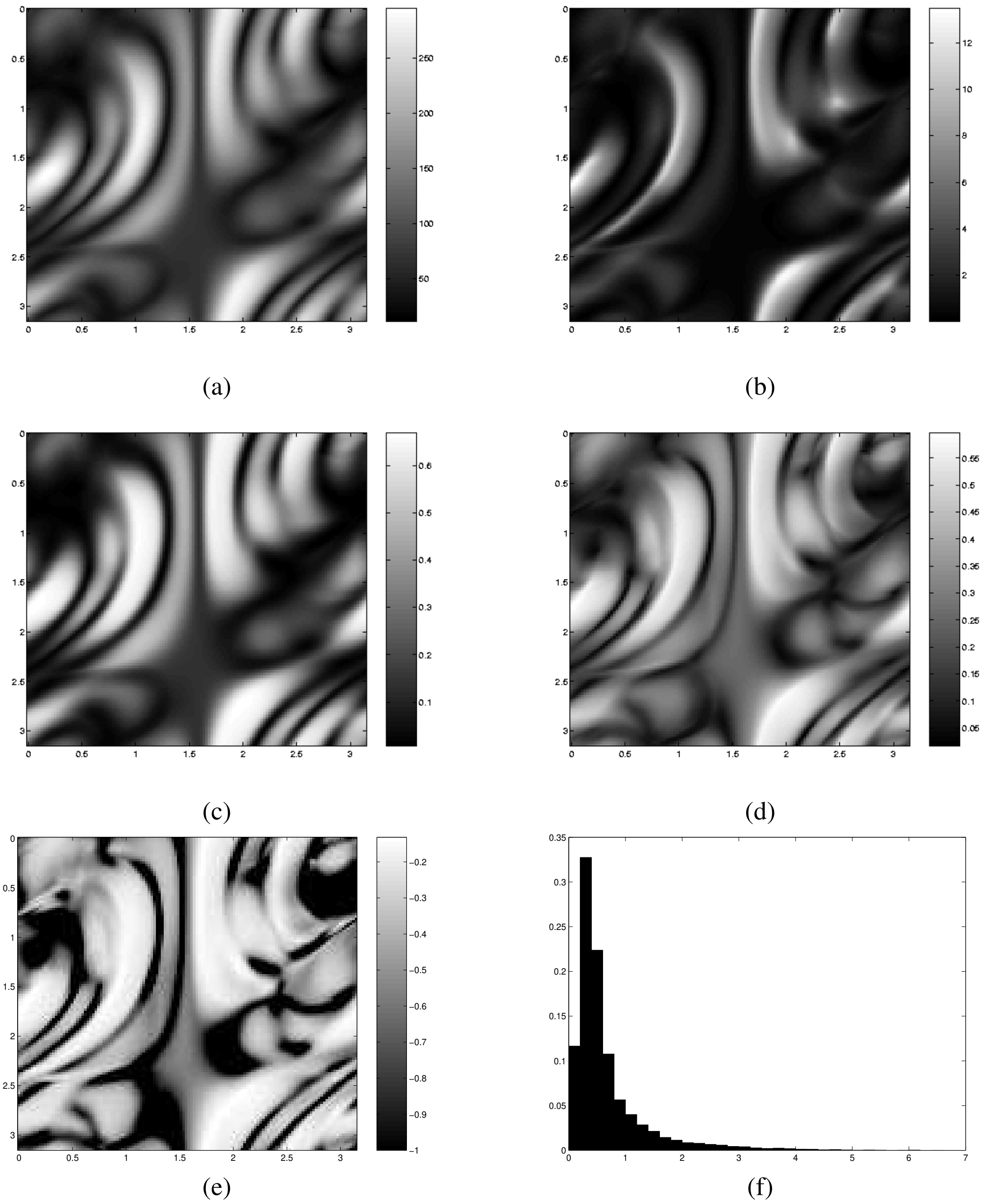


Figure 5: criteria for heartbeat classification with weighted ℓ_2-norm: (a) class centre distance, (b) scatter criterion, (c) alignment, (d) margin, (e) logarithm of radius - margin bound $(-\log_2(1+\frac{1}{n}\frac{R^2}{\rho^2}))$ clipped at 1, (f) histogram of logarithm of radius - margin bound

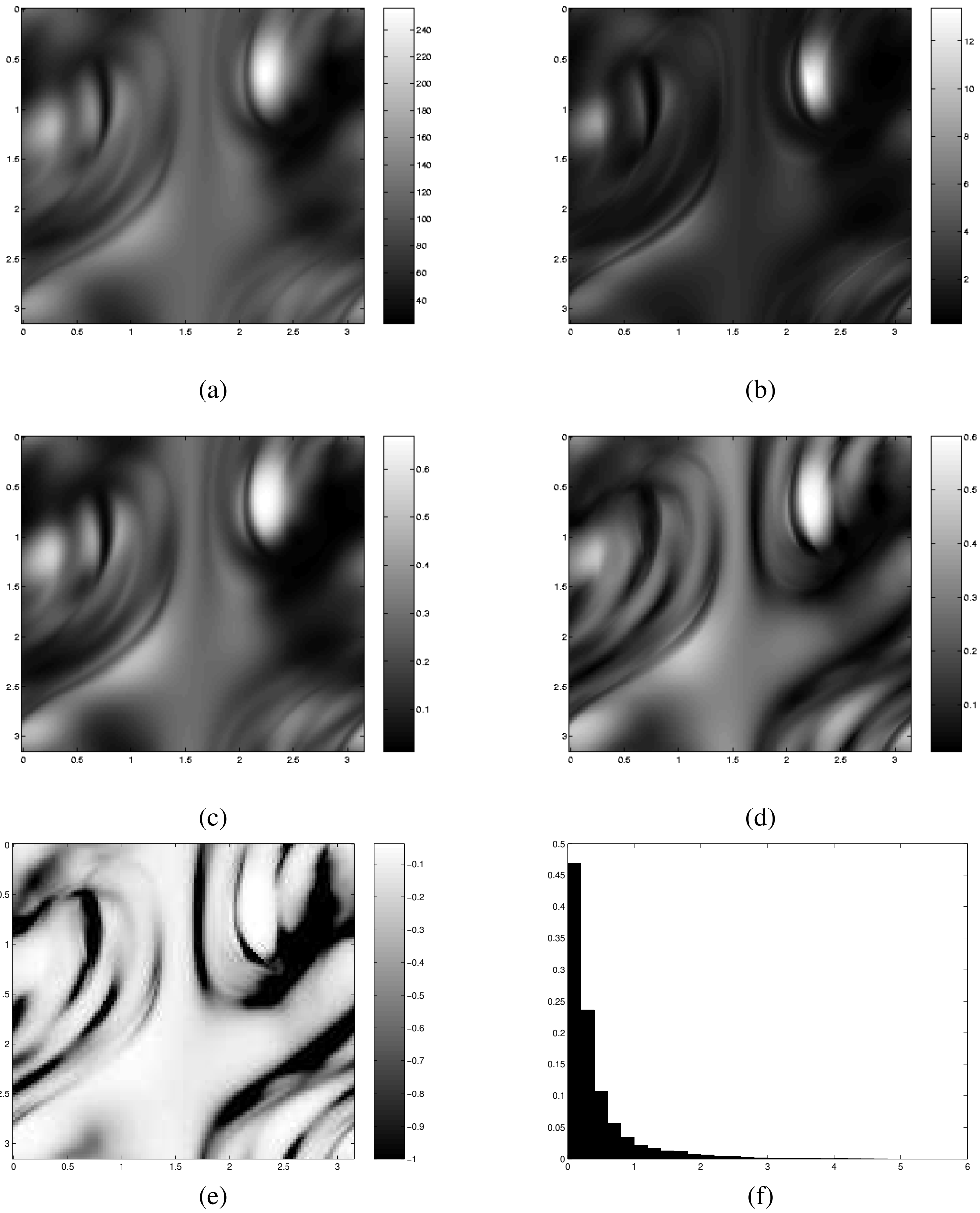
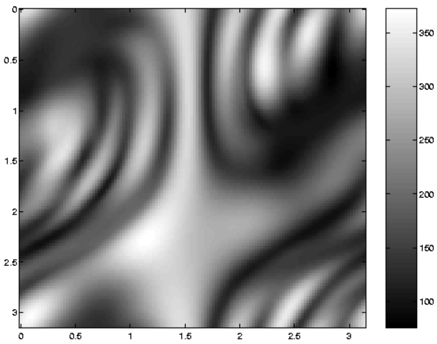
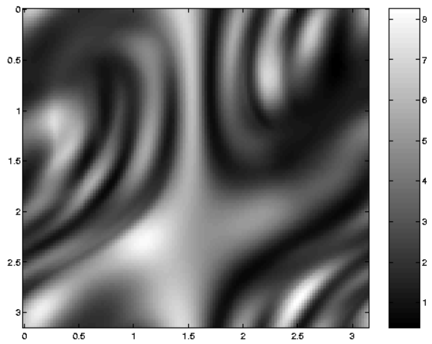


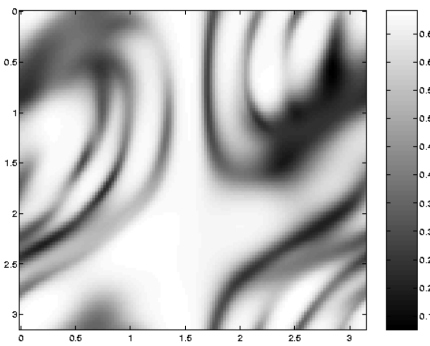
Figure 6: criteria for texture row classification with weighted ℓ_2-norm: (a) class centre distance, (b) scatter criterion, (c) alignment, (e) logarithm of radius - margin bound $(-\log_2(1 + \frac{1}{n}\frac{R^2}{\rho^2}))$ clipped at 1, (f) histogram of logarithm of radius - margin bound



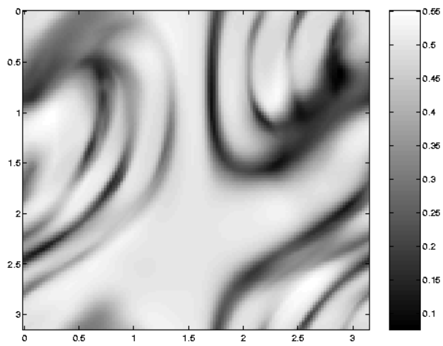
(a)



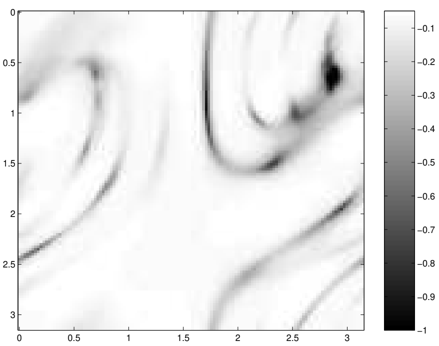
(b)



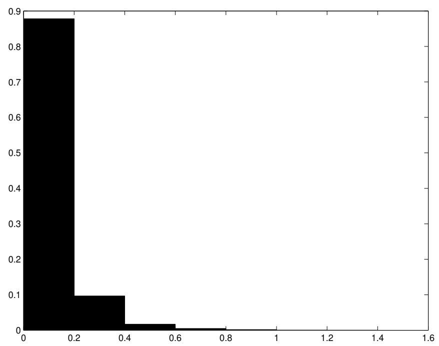
(c)



(d)

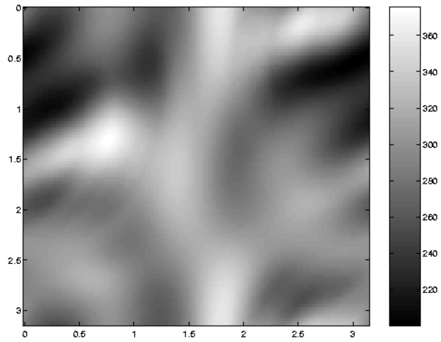


(e)

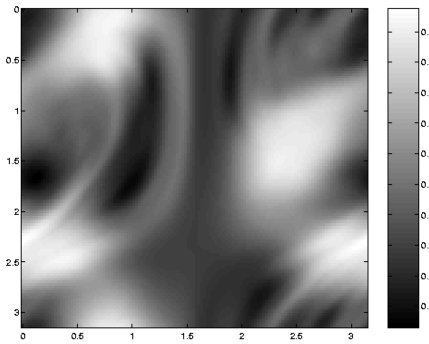


(f)

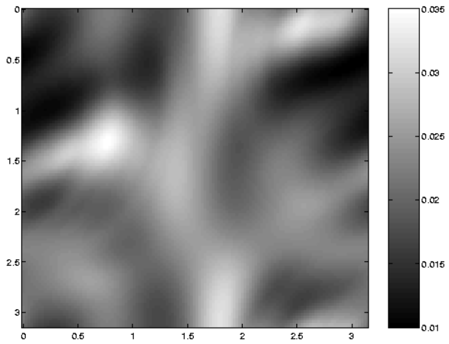
Figure 7: criteria for texture row classification with ℓ_2-norm: (a) class centre distance, (b) scatter criterion, (c) alignment, (d) margin, (e) logarithm of radius - margin bound ($-\log_2(1+\frac{1}{n}\frac{R^2}{\rho^2})$) clipped at 1, (f) histogram of logarithm of radius - margin bound



(a)



(b)



(c)

Figure 8: relationship between class centre distance and alignment for the ℓ_2 norm in another texture example: (a) class centre distance, (b) alignment for kernel width $\sigma = 100$, (c) alignment for kernel width $\sigma = 1000$

Although the scatter criterion $\frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}$ reveals more detailed structures, it doesn't seem to be superior to the simplest criterion, the class centre distance.

The radius margin bound $\frac{1}{n}\frac{R^2}{\rho^2}$ covers a large range of values. As its maximum goes up to 84 in the examples here, we plotted its logarithm $\log_2(1+\frac{1}{n}\frac{R^2}{\rho^2})$. Apart from the different distribution of the values, it rates the features mostly like the margin. Confirming the arguments regarding the wavelet adaptation problem, the range of values for the radius - margin bound from 10 resp. 3% to 100% (the maximum meaningful error bound) indicates the significance of the wavelet choice.

Norm Although the plots for the sample patient did not differ, there may be an important difference between using the ℓ_2 and the weighted ℓ_2 norm as exhibited by Figure 6 and Figure 7, even though the features are only weighted differently. Moreover, for the original ℓ_2 norm as depicted in Figure 7, the class centre distance differs slightly more from the kernel based criteria, namely alignment and margin.

Alignment - Class Centre Distance As reasoned in Section 4, the alignment should be linked to the class centre distance. The larger the kernel width σ is, the closer they are to each other. Motivated by this connection, the alignment for different kernel widths σ for another texture example (images 'Asphalt.0000' and 'Misc.0000' again from the MeasTex collection [38]) where the class centre distance and the alignment differed heavily is visualised in Figure 8.

Even though the distribution of the alignment for the smaller kernel with $\sigma = 100$ (with exponent $\frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\sigma^2} \approx 10^2$) and of the class centre distance are almost inverse, for the larger kernel with $\sigma = 1000$ (with exponent $\frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\sigma^2} \approx 1$) they again look very similar. Concerning the choice of the kernel parameter, the highest alignment is achieved for a kernel width of about $\sigma = 200$ for the original ℓ_2 norm and $\sigma = 80$ for the weighted norm.

5.3 Distances in Feature Space

To confirm the assumption that the distances in feature space resemble the original distances, we try to visualise the feature vectors. To visualise the original feature vectors $\mathbf{x}_i$, we again use PCA as in Section 5.1. To retain most of the total variance of the data, PCA projects the points on the eigendirections of the sample covariance matrix or mixture scatter matrix $\mathbf{S}_m = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^T$ corresponding to its largest eigenvalues. For the points $\boldsymbol{\phi}(\mathbf{x}_i)$ in feature space, we only know the matrix K of inner products as the mapping $\boldsymbol{\phi}$ isn't given explicitly and the feature space may even be infinite-dimensional. The PCA in feature space is called *kernel PCA* ([35]) and is carried out by projecting the potentially infinite-dimensional feature vectors onto the eigendirections of the centred kernel matrix

$$\tilde{\mathbf{K}} := \mathbf{CKC}$$

with centring matrix

$$\mathbf{C} := \mathbf{I} - \frac{1}{n}\mathbf{e}\mathbf{e}^T.$$

As $\tilde{\mathbf{K}}$ is symmetric and positive definite, it is unitary diagonalisable with positive eigenvalues which reads

$$\begin{aligned} \tilde{\mathbf{K}} &= \mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{V}^T \\ &=: \hat{\mathbf{X}}\hat{\mathbf{X}}^T. \end{aligned}$$

The rows of $\hat{\mathbf{X}}$ provide vectors in $\mathbb{R}^n$ that have the same distances as the vectors $\boldsymbol{\phi}(\mathbf{x}_i)$ in feature space. We get our approximations in $\mathbb{R}^2$ by only taking into account the first two components of these vectors (corresponding to the largest eigenvalues).

The resulting feature vectors for one wavelet in the example of Figure 8 (corresponding to a single point in each of the plots in Figure 8) are given in Figure 9. For the visualisation, we chose the optimal wavelet according to the alignment with the smaller Gaussian kernel (with $\sigma = 100$). Its filter bank angles are $(\varphi_1, \varphi_2) \approx (2.3, 3.1)$ marking the lightest spot in Figure 8 (b). One has to be careful with the interpretation of the resulting vector plots. If two scatter plots look different, there are two effects influencing this: Naturally, if the points are differently distributed, their projection in $\mathbb{R}^2$ will likely be different. But if this is the case, there may also be chosen other principal components. The best variance preserving linear projection isn't unique anyway, but we restricted the projection directions to be the eigendirections and took care with the signs and scales as well. Nevertheless, the quality of the projection has to be examined. In the given example, for the original feature vectors (a), the feature vectors for $\sigma = 100$ (b) and the feature vectors for $\sigma = 1000$ (c) approximately 88%, 46% and 87% of the total variance is retained, respectively. These numbers seem sufficient to draw some conclusions, and the scattering of the feature vectors for the larger kernel closely matches the scattering of the original feature vectors, as predicted in Section 4.3. For our feature vectors $\mathbf{x}$ with $\|\mathbf{x}\| = 1000$, we observed that from a kernel width σ of 500-750 on, the relative point positions resemble the original ones (depicted in Figure 9 in the example) so that one can still easily identify the single points in the feature space with the input points $\mathbf{x}_i$.

Besides, Figure 9 (b) shows how the optimal wavelet combined with the nonlinear feature map succeed in making the points easily separable.

6 Conclusion

For the number of parameters of an algorithm two opposite trends exist. Having more parameters increases the flexibility and adaptability of an algorithm. But, on the other hand, which is often

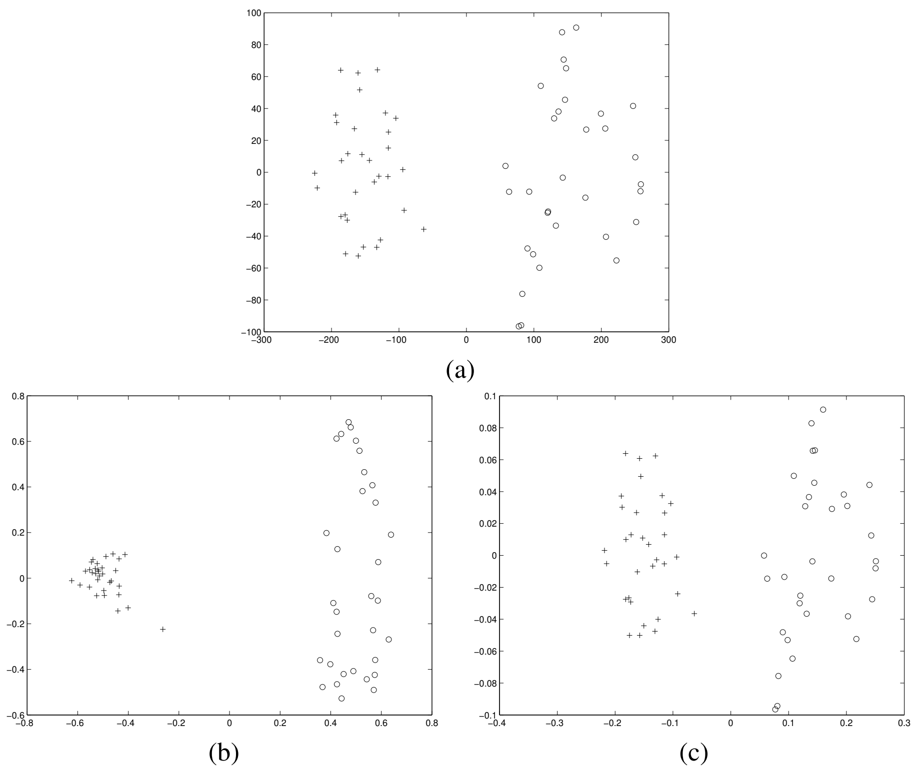


Figure 9: principal components of feature vectors for optimally aligned wavelet in example of Figure 8: (a) in original space $\mathbb{R}^d$, (b) in feature space for Gaussian kernel with $\sigma = 100$, (c) in feature space for Gaussian kernel with $\sigma = 1000$

considered more important, the lack of parameters disburdens users of choosing the right parameter values.

The same problem appears for the wavelet choice in the decomposition based signal classification. An adaptation of the wavelet to the specific problem reduces the classification error, but raises the problem of picking out the 'best' wavelets.

For that purpose, we assembled and proposed several easy to evaluate criteria for the rating of features for support vector machine classification. According to our evaluation, the simple criteria well resemble the margin and the radius - margin classification error bound and thus are sufficient for parameter selection. Particularly, it could be seen by derivations and examples that the class centre distance of the feature vectors in the data space on certain conditions well resembles the alignment.

But, unfortunately, the criteria plots made evident that even the simple criteria as the class centre distance, although differentiable and easily evaluated, possess many local minima. Hence, for effectively choosing the optimal wavelet in the proposed problem setting, simple optimisation techniques won't suffice.

Additionally to the criteria comparison, for a class of kernels this paper gives an easier way to compute the radius of a set of feature vectors by reducing the problem on the single class support vector classification problem.

A Filter Banks and Discrete Orthonormal Wavelets

We will give some background about filter banks and discrete wavelets here and introduce our notation about them. The concept of 'filter banks' comes from engineering sciences and is widely used in all signal processing areas as pattern recognition. Work on wavelets was pioneered by S. Mallat ([26]) and I. Daubechies ([10]). The usual wavelet is a continuous function. However, we want to apply the term 'wavelet' in the discrete setting. The derivation of the analogies will be given in the following.

A.1 Filter Banks

A *filter bank* is a system of filters, linked by operations as up- and down-sampling to analyse a signal or synthesise it again. The essential information is extracted from the resulting subband signals of an analysis filter bank. Our notion of filter banks is mainly based upon the book [40]. We will only use two–channel filter banks whose analysis filters normally consist of a low–pass and a high–pass filter.

Let $H_0(z) := \sum_{k\in\mathbb{Z}} h_0[k]z^{-k}$ resp. $H_1(z) := \sum_{k\in\mathbb{Z}} h_1[k]z^{-k}$ be the z–transform of these two filters. For the signal decomposition, we are interested in the filter coefficient sequences $(h_0[k])_{k\in\mathbb{Z}}, (h_1[k])_{k\in\mathbb{Z}} \in \ell_2$.

A filter bank with analysis filters H_0 and H_1 is called *paraunitary* (also referred to as *orthogonal*) if

$$\begin{pmatrix} H_0(z^{-1}) & H_1(z^{-1}) \\ H_0(-z^{-1}) & H_1(-z^{-1}) \end{pmatrix} \begin{pmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{pmatrix} =$$

$$\begin{pmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{pmatrix} \begin{pmatrix} H_0(z^{-1}) & H_1(z^{-1}) \\ H_0(-z^{-1}) & H_1(-z^{-1}) \end{pmatrix} = 2\mathbf{I}. \tag{10}$$

The corresponding synthesis filters are given by

$$G_0(z) = H_0(z^{-1}),\ G_1(z) = H_1(z^{-1}). \tag{11}$$

The *polyphase matrix* of a paraunitary filter bank is defined as

$$\mathbf{H}_{\mathrm{pol}}(z) := \begin{pmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{pmatrix}$$

with entries from the polyphase decomposition

$$H_i(z) := H_{i0}(z^2) + z^{-1}H_{i1}(z^2) \quad i = 0, 1. \tag{12}$$

To split the signals into different frequency bands, often high-pass filters with at least one vanishing moment are considered. To this end, the filter bank has to satisfy the low-pass condition

$$H_0(1) = \sqrt{2} \tag{13}$$

or, equivalently,

$$H_1(1) = 0. \tag{14}$$

For our practical purposes, we are interested in finite impulse response (FIR) filters. The analysis filters of order $2L + 2$ then read

$$H_i(z) = \sum_{k=0}^{2L+1} h_i[k]z^{-k} \quad i = 0, 1.$$

A.2 Discrete Orthonormal Wavelets

To justify the term 'wavelet decomposition' for our feature extraction process described in Section 2, we note that filter banks are connected to wavelets. Every orthogonal continuous wavelet corresponds to a paraunitary filter bank in that the discrete wavelet transform yields the same as filtering with the corresponding filter bank. But not all orthonormal filter banks covered by the parameter space here are related to continuous wavelets. To cope with this mismatch, in the style of the books [47, Section 3.3.2] and [44, Section 11.4], we want to introduce 'discrete–time scaling sequences' and 'discrete–time wavelets'.

Given a possibly infinite signal $\mathbf{s} = (s_i)_{i\in\mathbb{Z}} \in \ell_2$ and a paraunitary filter bank with analysis filter coefficients $(h_0[k])_{k\in\mathbb{Z}}, (h_1[k])_{k\in\mathbb{Z}} \in \ell_2$ and synthesis filters $(g_0[k])_{k\in\mathbb{Z}} \overset{(11)}{=} (h_0[-k])_{k\in\mathbb{Z}}, (g_1[k])_{k\in\mathbb{Z}} \overset{(11)}{=} (h_1[-k])_{k\in\mathbb{Z}} \in \ell_2$, we want to analyse the signal with the corresponding filter bank. With $\mathbf{g}_{jk} := (g_j[i-2k])_{i\in\mathbb{Z}} \in \ell_2 (j = 0, 1; k \in \mathbb{Z})$, the orthogonality conditions for the z-transforms (10) and (11) imply the orthogonality conditions

$$\begin{aligned} \langle \mathbf{g}_{ji}, \mathbf{g}_{jk} \rangle_{\ell_2} &= \delta(i-k)\,,\ j = 0, 1\,, i, k \in \mathbb{Z}, \\ \langle \mathbf{g}_{0i}, \mathbf{g}_{1k} \rangle_{\ell_2} &= 0\,, \qquad\qquad\qquad\quad i, k \in \mathbb{Z} \end{aligned}$$

for the filter coefficients, where

$$\delta(x) := \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Due to the perfect reconstruction property of paraunitary filter banks, the set

$$\{\mathbf{g}_{jk} : j = 0, 1; k \in \mathbb{Z}\}$$

forms an orthonormal basis of ℓ_2. Hence, we can represent the signal as

$$\mathbf{s} = \sum_{k\in\mathbb{Z}} c_k^1 \mathbf{g}_{0k} + \sum_{k\in\mathbb{Z}} d_k^1 \mathbf{g}_{1k} \tag{15}$$

with

$$\begin{aligned} c_k^1 &= \langle \mathbf{s}, \mathbf{g}_{0k} \rangle_{\ell_2} = \langle \mathbf{s}, (h_0[2k-i])_{i\in\mathbb{Z}} \rangle_{\ell_2} = \sum_{i\in\mathbb{Z}} s_i h_0[2k-i], \\ d_k^1 &= \langle \mathbf{s}, \mathbf{g}_{1k} \rangle_{\ell_2} = \langle \mathbf{s}, (h_1[2k-i])_{i\in\mathbb{Z}} \rangle_{\ell_2} = \sum_{i\in\mathbb{Z}} s_i h_1[2k-i] \end{aligned}$$

for $k \in \mathbb{Z}$. Equivalently, in the z-domain, this reads

$$S(z) = C^1(z^2) G_0(z) + D^1(z^2) G_1(z) \tag{16}$$

with

$$C^1(z^2) = \frac{1}{2}(H_0(z)S(z) + H_0(-z)S(-z)), \tag{17}$$

$$D^1(z^2) = \frac{1}{2}(H_1(z)S(z) + H_1(-z)S(-z)). \tag{18}$$

If we want to perform several decomposition steps, we refine the signal representations (15) or (16) further and obtain

$$\mathbf{c}^1 = (c_k^1)_{k\in\mathbb{Z}} = \sum_{k\in\mathbb{Z}} c_k^2 \mathbf{g}_{0k} + \sum_{k\in\mathbb{Z}} d_k^2 \mathbf{g}_{1k}$$

with

$$\begin{aligned} c_k^2 &= \sum_{j\in\mathbb{Z}} c_j^1 h_0[2k-j] = \langle \mathbf{c}^1, \mathbf{g}_{0k} \rangle_{\ell_2} = \langle \langle \mathbf{s}, \mathbf{g}_{0j} \rangle_{\ell_2}, \mathbf{g}_{0k} \rangle_{\ell_2}, \\ d_k^2 &= \sum_{j\in\mathbb{Z}} c_j^1 h_1[2k-j] = \langle \mathbf{c}^1, \mathbf{g}_{1k} \rangle_{\ell_2} = \langle \langle \mathbf{s}, \mathbf{g}_{0j} \rangle_{\ell_2}, \mathbf{g}_{1k} \rangle_{\ell_2} \end{aligned}$$

for $k \in \mathbb{Z}$ or, equivalently, in the z-domain

$$
\begin{aligned}
S(z) &= C^1(z^2)G_0(z) + D^1(z^2)G_1(z) \\
&= (C^2(z^4)G_0(z^2) + D^2(z^4)G_1(z^2))G_0(z) + D^1(z^2)G_1(z) \\
&= C^2(z^4)G_0(z^2)G_0(z) + D^2(z^4)G_1(z^2)G_0(z) + D^1(z^2)G_1(z)
\end{aligned}
$$

with

$$
\begin{aligned}
C^2(z^4) &= \frac{1}{2}(H_0(z^2)C^1(z^2) + H_0(-z^2)C^1(-z^2)), \qquad (19)\\
D^2(z^4) &= \frac{1}{2}(H_1(z^2)C^1(z^2) + H_1(-z^2)C^1(-z^2)).
\end{aligned}
$$

We are looking for the filter coefficients corresponding to these iterated filters that produce the full decomposition. We therefore define the coefficient sequences $\mathbf{v_0^j}, \mathbf{w_0^j}$ of the z-transforms

$$V^j(z) := \prod_{m=0}^{j-1} G_0(z^{2^m}) \ , \quad j \in \mathbb{N}_0 \qquad (20)$$

$$W^j(z) := G_1(z^{2^{j-1}}) \prod_{m=0}^{j-2} G_0(z^{2^m}) \ , \quad j \in \mathbb{N} \qquad (21)$$

as *discrete–time scaling sequences* and *discrete–time wavelets*, respectively. Let $\mathbf{v_k^j} := v_0^j[\cdot - 2^j k]$ and $\mathbf{w_k^j} := w_0^j[\cdot - 2^j k]$ $(k \in \mathbb{Z})$ denote the translates of the sequences by multiples of their sample length 2^j. The following orthogonality relations then hold for all $i, k \in \mathbb{Z}; j, m \in \mathbb{N}$:

$$\langle \mathbf{v_i^j}, \mathbf{v_k^j} \rangle_{\ell_2} = \delta(k-i), \qquad (22)$$

$$\langle \mathbf{w_i^j}, \mathbf{w_k^m} \rangle_{\ell_2} = \delta(k-i)\delta(m-j), \qquad (23)$$

$$\langle \mathbf{v_i^j}, \mathbf{w_k^j} \rangle_{\ell_2} = 0. \qquad (24)$$

As in the case of continuous wavelets, the sequences $\mathbf{v_k^j}$ and $\mathbf{w_k^j}$ $(j \in \mathbb{N}, k \in \mathbb{Z})$ have widths scaled by two and lie in different resolution subspaces j, and the wavelets and scaling sequences on each level j form a basis of the space spanned by the scaling sequences on the above level $j-1$. Hence, in analogy to the continuous case the discrete–time scaling sequences span a multiresolution analysis of ℓ_2, with the major difference that they may not be scaled smaller which would require a negative j. To see this, we define the sequence of spaces

$$
\begin{aligned}
V^j &:= \overline{\operatorname{span}\{\mathbf{v_k^j} : k \in \mathbb{Z}\}} \subset \ell_2 \ , \ j \in \mathbb{N}_0 \, , \\
W^j &:= \overline{\operatorname{span}\{\mathbf{w_k^j} : k \in \mathbb{Z}\}} \subset \ell_2 \ , \ j \in \mathbb{N}.
\end{aligned}
$$

The multiresolution properties

$$
\begin{aligned}
&V^0 \supset V^1 \supset V^2 \supset \dots \, , \\
&\bigcup_{j \in \mathbb{N}_0} V^j = V^0 = \ell_2, \\
&\bigcap_{j \in \mathbb{N}_0} V^j = \{\mathbf{0}\}
\end{aligned}
$$

then hold due to the definition of the scaling sequence filters (20). And due to (22), the set $\{\mathbf{v_k^j} : k \in \mathbb{Z}\}$ forms an orthonormal basis of V^j. Further, (21) and (24) imply that the detail spaces W^j form the orthogonal complement to the approximation spaces V^j in the next bigger spaces V^{j-1}

$$V^{j-1} = V^j \oplus W^j \ , \ j \in \mathbb{N}.$$

As a consequence, the whole space of sequences can be decomposed into $\ell_2 = \bigoplus_{j\in\mathbb{N}} W^j = V^J \oplus \bigoplus_{j=1}^{J} W^j$ with orthonormal basis

$$\{\mathbf{v_k^J}, \mathbf{w_k^j} : j = 1, \ldots, J; k \in \mathbb{Z}\}.$$

The representation of a signal $\mathbf{s} \in \ell_2$ in terms of this basis corresponds to a wavelet decomposition of J steps:

$$\mathbf{s} = \sum_{k\in\mathbb{Z}} c_k^J \mathbf{v_k^J} + \sum_{j=1}^{J} \sum_{k\in\mathbb{Z}} d_k^j \mathbf{w_k^j}$$

with the *wavelet coefficients*

$$\mathbf{c^j} := (c_k^j)_{k\in\mathbb{Z}} = (\langle \mathbf{s}, \mathbf{v_k^j} \rangle_{\ell_2})_{k\in\mathbb{Z}}\ ,\ j \in \mathbb{N},$$
$$\mathbf{d^j} := (d_k^j)_{k\in\mathbb{Z}} = (\langle \mathbf{s}, \mathbf{w_k^j} \rangle_{\ell_2})_{k\in\mathbb{Z}}\ ,\ j \in \mathbb{N}.$$

Beside the orthonormality, another property that is important for our feature extraction holds for the filter banks, and therewith the discrete wavelets, generated by the lattice structure (1). The average signal value is always preserved in the low-pass channel.

Lemma 1 (average signal value). *Given a paraunitary filter bank that satisfies the low-pass condition (13), the low-pass coefficients satisfy*

$$C^j(1) = \left(\frac{1}{\sqrt{2}}\right)^j S(1)\ ,\ j \in \mathbb{N}$$

for all signals S(z).

Proof. Consider the decomposition equation (16) with low-pass coefficients (17). The orthogonality condition (10) equivalently reads

$$H_0(z^{-1})H_0(z) + H_1(z^{-1})H_1(z) = 2,$$
$$H_0(z^{-1})H_0(-z) + H_1(z^{-1})H_1(-z) = 0.$$

¿From the low-pass condition $H_0(1) = \sqrt{2}$ (13), it follows by the first equation for $z = 1$ that

$$H_1(1) = 0,$$

and further, by the second equation,

$$H_0(-1) = 0.$$

With these relations, (17) reads

$$C^1(1) = \frac{1}{2}(H_0(1)S(1) + H_0(-1)S(-1)) = \frac{1}{\sqrt{2}}S(1).$$

The property for higher levels j follows by induction by iterating the decomposition on the low-pass coefficients as indicated by (19). □

In terms of the coefficients s_k and c_k^j ($k \in \mathbb{Z}$), $S(1)$ and $C^j(1)$ ($j \in \mathbb{N}$) are the sums of the appropriate coefficients, so the lemma states that the average signal value $\frac{1}{l}S(1)$ is directly related to the (finite) sum of the low-pass coefficients $C^j(1)$. Especially for the choice $S(1) = 0$, this implies that the coefficients will also have mean zero by $C^j(1) = 0$.

Note that for non-subsampled decomposition, by $C^1(z) = S(z)H_0(z)$, the analogous property

$$C^j(1) = \left(\sqrt{2}\right)^j S(1)\ ,\ j \in \mathbb{N}$$

holds.

Given the finite analysis filters $(h_0[k])_{k=0,\ldots,2L+1}$, $(h_1[k])_{k=0,\ldots,2L+1}$, the decomposition of a signal $\mathbf{s}$ with length $l = n2^d$ $(n \in \mathbb{N})$ in $J = d$ steps should be done easily. But since we are not able to calculate infinite coefficient sequences, we restrict the wavelets and scaling sequences to the finite-dimensional space $\mathbb{R}^l$. Due to this restriction, the question what to do at the boundary is coming up. We propose to continue the wavelets and scaling sequences l-periodically to preserve the orthogonality of the transform with the finite orthonormal basis

$$\{\tilde{\mathbf{v}}_{\mathbf{k}}^{\mathbf{J}} = (\mathbf{v}_{\mathbf{k}}^{\mathbf{J}}[\mathbf{i} \bmod \mathbf{l}])_{i=0,\ldots,l-1}, \tilde{\mathbf{w}}_{\mathbf{k}}^{\mathbf{j}} = (\mathbf{w}_{\mathbf{k}}^{\mathbf{j}}[\mathbf{i} \bmod \mathbf{l}])_{i=0,\ldots,l-1} : j = 1, \ldots, J; k = 1, \ldots, l/2^j\}.$$

B Support Vector Machines

Here we provide the tools concerning the support vector machine classification. Our approach is based on the pioneering work of Vapnik [45] and the book of Cristianini and Shawe–Taylor [7], where the reader can find a detailed introduction in terms of statistical learning theory.

B.1 Mathematical Background: Reproducing Kernel Hilbert Spaces

The main innovation of the Support Vector Machine was the use of a kernel function to state a nonlinear classifier in terms of a linear classifier. Thereto, one has to be able to evaluate inner products between nonlinearly mapped feature vectors with a kernel function. The mathematics involved in this are 'reproducing kernel Hilbert spaces' which we will introduce now.

By $L_2(\mathcal{X})$ we denote the Hilbert space of real valued square integrable functions on $\mathcal{X}$ with inner product $\langle f, g\rangle_{L_2} = \int_{\mathcal{X}} f(x)g(x)\,\mathrm{d}x$. A *kernel* is a positive definite symmetric function $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ in $L_2(\mathcal{X} \times \mathcal{X})$. Following [30], we call a function $K \in L_2(\mathcal{X} \times \mathcal{X})$ positive definite iff 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. The kernel K resp. the matrix of inner products $(K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ now embodies the mapping into feature space. This definition does not apply to the linear mapping $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y}\rangle$. One could also generalise the definition to conditionally positive definite functions. But, in this paper 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

$$K(\mathbf{x}, \mathbf{y}) = k(||\mathbf{x} - \mathbf{y}||). \tag{25}$$

For a given kernel 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

$$\langle K(\tilde{\mathbf{x}}, \cdot), K(\bar{\mathbf{x}}, \cdot)\rangle_{\mathcal{H}_K} := K(\tilde{\mathbf{x}}, \bar{\mathbf{x}}) \tag{26}$$

which has reproducing kernel, 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 *Mercers Theorem*, K can be expanded in a uniformly convergent series on $\mathcal{X} \times \mathcal{X}$

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(\mathbf{x})\varphi_j(\mathbf{y}), \tag{27}$$

where $\lambda_j \geq 0$ are the eigenvalues of the integral operator $T_K : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ with $T_K f(\cdot) := \int_{\mathcal{X}} K(\mathbf{x}, \cdot)f(\mathbf{x})\,\mathrm{d}\mathbf{x}$ and where $\{\varphi_j\}_{j\in\mathbb{N}}$ are the corresponding $L_2(\mathcal{X})$–orthonormalised eigenfunctions.

The feature map then reads

$$\boldsymbol{\phi}(\cdot) := \left(\sqrt{\lambda_j}\varphi_j(\cdot)\right)_{j\in\mathbb{N}}.$$

By (27), we have that $\phi(\mathbf{x})$ ($\mathbf{x} \in \mathcal{X}$) is an element in ℓ_2 with

$$||\phi(\mathbf{x})||^2 = \sum_{j=1}^{\infty} \lambda_j \varphi_j^2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) = k(0).$$

We define the *feature space* $\mathcal{F}_K \subset \ell_2$ by the ℓ_2–closure of all finite linear combinations of elements $\phi(\mathbf{x})$ ($\mathbf{x} \in \mathcal{X}$)

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

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 isomorphic with isometry $\iota : \mathcal{F}_K \to \mathcal{H}_K$ defined by

$$\iota(\mathbf{w}) := f_{\mathbf{w}}(\cdot) = \langle \mathbf{w}, \phi(\cdot) \rangle_{\mathcal{F}_K} = \sum_{j=1}^{\infty} w_j \sqrt{\lambda_j} \varphi_j(\cdot). \tag{28}$$

In particular, we have that

$$\|f_{\mathbf{w}}\|_{\mathcal{H}_K} = \|\mathbf{w}\|_{\mathcal{F}_K}. \tag{29}$$

Note that from another point of view $\mathcal{F}_K$ is the space of sequences of the Fourier coefficients of the functions of $\mathcal{H}_K$ with respect to the orthonormal basis $\{\sqrt{\lambda_j}\varphi_j : j = 1, \dots\}$ of $\mathcal{H}_K$.

B.2 Solution

Now we consider the classification problem introduced in section 3. Again assume the training set (4) is given. We just introduced the function space $\mathcal{H}_K$ corresponding to the chosen kernel K. To build a classifier, we reformulate the unconstrained optimisation problem (6) already set up in section 3 to the following equivalent constrained optimisation problem:

$$\begin{aligned} \min_{f \in \mathcal{H}_K, \xi_i \in \mathbb{R} (i=1,\dots,n)} \quad & C \left(\sum_{i=1}^{n} \xi_i \right) + \frac{1}{2} ||f||^2_{\mathcal{H}_K} \\ \text{subject to} \quad y_i f(\mathbf{x}_i) \geq & \;\; 1 - \xi_i \,, \quad i = 1, \dots, n, \\ \xi_i \geq & \;\; 0 \,, \quad i = 1, \dots, n. \end{aligned} \tag{30}$$

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 is only needed for inner product functions that are only positive semidefinite ([17]). With our definition of a *kernel*, the bias is always included implicitly as the set of eigenfunctions $\{\varphi\}_{j \in \mathbb{N}}$ always contains a constant function, implying $1 \in \mathcal{H}_K$. As a consequence, $f = h + b \in \mathcal{H}_K$ for $h \in \mathcal{H}_K$ and $b \in \mathbb{R}$, so the minimisation already takes into account all functions of this form.

Every function $f \in \mathcal{H}_K$ corresponds uniquely to a sequence $\mathbf{w} \in \mathcal{F}_K$. Thus, by (28) and (29), the optimisation problem (30) can be rewritten as follows:

$$\min_{\mathbf{w} \in \mathcal{F}_K, \xi_i \in \mathbb{R} (i=1,\dots,n)} \quad C\left(\textstyle\sum_{i=1}^{n} \xi_i\right) + \frac{1}{2}||\mathbf{w}||^2_{\mathcal{F}_K} \tag{31}$$

$$\begin{aligned} \text{subject to} \quad & y_i \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle_{\mathcal{F}_K} && \geq 1 - \xi_i \,, \quad i = 1, \dots, n, \\ & \xi_i && \geq 0 \,, \quad i = 1, \dots, n. \end{aligned} \tag{32}$$

In general the feature space $\mathcal{F}_K \subset \ell_2$ is infinite-dimensional. For a better illustration of (31) we assume for a moment that $\mathcal{F}_K \subset \mathbb{R}^n$. Then the function $\tilde{f}_{\mathbf{w}}(\cdot) := \langle \mathbf{w}, \cdot \rangle_{\mathcal{F}_K}$ 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 $\tilde{\mathbf{v}} \in \mathcal{F}_K$ has the distance $|\langle \mathbf{w}, \tilde{\mathbf{v}} \rangle_{\mathcal{F}_K}|/||\mathbf{w}||_{\mathcal{F}_K}$ from $H_{\mathbf{w}}$. Note that $\tilde{f}_{\mathbf{w}}(\phi(\mathbf{x})) = f_{\mathbf{w}}(\mathbf{x})$. Thus, the constraints $y_i\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle_{\mathcal{F}_K}/||\mathbf{w}||_{\mathcal{F}_K} \geq 1/||\mathbf{w}||_{\mathcal{F}_K} - \xi_i/||\mathbf{w}||_{\mathcal{F}_K}$ ($i = 1, \dots, n$) in (32) require that every $\phi(\mathbf{x}_i)$ must at least have the distance $1/||\mathbf{w}||_{\mathcal{F}_K} - \xi_i/||\mathbf{w}||_{\mathcal{F}_K}$ from $H_{\mathbf{w}}$.

If there exists $\mathbf{w} \in \mathcal{F}_K$ so that (32) can be fulfilled with $\xi_i = 0$ ($i = 1, \dots, n$), then we say that our training set is linearly separable in $\mathcal{F}_K$. Of course, for Gaussian kernels like all positive kernels every finite training set is linearly separable in $\mathcal{F}_K$, see, e.g., [39]. Then the optimisation problem (31) can be further simplified to

$$\begin{aligned} \min_{\mathbf{w}\in\mathcal{F}_K} \quad & \frac{1}{2}||\mathbf{w}||^2_{\mathcal{F}_K} \\ \text{subject to} \quad & y_i\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{F}_K} \geq 1\ , \quad i = 1,\ldots,n. \end{aligned} \tag{33}$$

Given $\mathcal{H}_K$ and $\mathcal{Z}$, the optimisation problem above has a unique solution $f_{\mathbf{w}^*}$. In our hyperplane context $H_{\mathbf{w}^*}$ is exactly the hyperplane which has maximal distance ρ from the training data, where

$$\rho := \frac{1}{||\mathbf{w}^*||_{\mathcal{F}_K}} = \frac{1}{||f_{\mathbf{w}^*}||_{\mathcal{H}_K}} = \max_{\mathbf{w}\in\mathcal{F}_K} \min_{i=1,\ldots,n} \left\{ \frac{|\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{F}_K}|}{||\mathbf{w}||_{\mathcal{F}_K}} \right\}. \tag{34}$$

The value ρ is called the *margin* of $f_{\mathbf{w}^*}$ with respect to the training set $\mathcal{Z}$. In this context, the solutions of the optimisations problems (31) and (33) are called *soft margin* and *hard margin SVM classifiers*, respectively. Due to the regularity of the kernel matrix, there always exists a solution even for hard margin SVMs. Note that this restriction is equivalent to the choice $C = \infty$ in (30) or (31).

Next we consider the solution of the SVM problem (31), where we follow mainly the notation of [48]. Here the notion 'support vector' comes into play.

By the *Representer Theorem* ([24, 48]), the minimiser of (31) has the form

$$f(\mathbf{x}) = \sum_{j=1}^{n} c_j K(\mathbf{x}, \mathbf{x}_j). \tag{35}$$

Setting $\mathbf{f} := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$ and $\mathbf{K} := (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ we obtain that

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

Note that $\mathbf{K}$ is positive definite. Further, define $\mathbf{Y} := \operatorname{diag}(y_1, \ldots, y_n)$. Then the optimisation problem (31) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{\xi}\in\mathbb{R}^n, \mathbf{c}\in\mathbb{R}^n} \quad & C\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\mathbf{c}^T\mathbf{K}\mathbf{c} \\ \text{subject to} \quad & \mathbf{Y}\mathbf{K}\mathbf{c} \geq \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \tag{36}$$

The dual problem with Lagrange multipliers $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ reads

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}^n} \mathcal{L}(\mathbf{c}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

where

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

subject to

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

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

$$\mathbf{c} = \mathbf{Y}\boldsymbol{\alpha}. \tag{37}$$

Further we have by $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\xi}} = \mathbf{0}$ that $\boldsymbol{\beta} = C\mathbf{e} - \boldsymbol{\alpha}$. Thus, our optimisation problem becomes

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & -\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Y}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha} + \mathbf{e}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{e}. \end{aligned} \tag{38}$$

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., $|\mathcal{Z}| > 4000$, specifically designed large scale algorithms should be applied, e.g., SVM*light* [22].

If we again denote by I the index set of the support vectors $I := \{i \in \{1, \dots, n\} : \alpha_i \neq 0\}$ then by (35) and (37), 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 (34) and (26) that

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

Due to the Kuhn–Tucker conditions [15] the solution f of the QP problem (36) has to fulfil

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

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

$$\rho = \left(\sum_{i \in I} \alpha_i \right)^{-\frac{1}{2}} . \tag{39}$$

C An SVM Formulation for Radius Computation

This section describes how to efficiently compute the radius of the smallest sphere enclosing a set of points. This radius was involved in the radius - margin criterion (9) and its efficient computation was used in the error bounds visualised in Figures 5 to 7 (e).

A direct approach to determine the radius R for the points $\mathbf{x}_j$ ($j = 1, \dots, n$) in feature space is to solve the optimisation problem

$$\begin{aligned} &\min_{\mathbf{a} \in \mathcal{F}_K, R \in \mathbb{R}} R^2 \\ &\text{subject to} \quad ||\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}||^2 \leq R^2 , \quad j = 1, \dots, n. \end{aligned} \tag{40}$$

In fact, this quadratic problem is a special case of the problem

$$\begin{aligned} &\min_{\mathbf{a} \in \mathcal{F}_K, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n} R^2 + C \sum_{j=1}^{n} \xi_j \\ &\text{subject to} \quad ||\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}||^2 \leq R^2 + \xi_j , \quad j = 1, \dots, n \\ &\qquad\qquad\qquad\qquad\qquad \xi_j \geq 0 , \quad j = 1, \dots, n \end{aligned} \tag{41}$$

considered in [2] for clustering. Therefore we will refer to (41) as *SV clustering problem*. We will show that (41) can be solved by a single-class SVM, i.e., an SVM classification problem with all points belonging to the same class. Then the matrix $\mathbf{Y}$ in (38) is the identity matrix so that (38) simplifies to

$$\begin{aligned} &\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \\ &\text{subject to} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \hat{C} \mathbf{e}. \end{aligned} \tag{42}$$

Although this is still a quadratic program we think that it is profitable to use this connection, since, for standard SVMs, sophisticated algorithms are included into many software implementations.

We will prove the following theorem:

Theorem 2. *Let K be a kernel with corresponding feature map $\boldsymbol{\phi}$ and with the property that $K(\mathbf{x},\mathbf{x}) = \kappa$ for all $\mathbf{x} \in \mathbb{R}^d$. Then there exists $\hat{C} > 0$ such that the optimal radius R in (41) can be obtained by solving the dual problem (42) of a single-class SVM. More precisely, if $\boldsymbol{\alpha}$ is the solution of (42), then*

$$R^2 = \kappa + \boldsymbol{\beta}^T \mathbf{K}\boldsymbol{\beta} - 2(\mathbf{K}\boldsymbol{\beta})_j, \tag{43}$$

where $\boldsymbol{\beta} := \frac{\boldsymbol{\alpha}}{\mathbf{e}^T\boldsymbol{\alpha}}$ and $j \in \{1,\dots,n\}$ denotes some index with $0 < \beta_j < C$.

Note that $C = \hat{C} = \infty$ for our original problem (40).

Our proof proceeds in two steps: first we show that the SV clustering problem (41) is equivalent to a single-class SVM with additional bias term also included in the objective function. This SVM was used for novelty detection in [37] and is therefore called *SV novelty detection problem* in the following. Then we prove that the SV novelty detection problem is equivalent to an ordinary single-class SVM (42) without bias term.

C.1 Equivalence of the SV Clustering Problem and the SV Novelty Detection Problem

The equivalence is best shown considering the dual problems. For solving (41), we introduce the Lagrangian

$$\mathcal{L}(R,\mathbf{a},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\mu}) := R^2 + C\mathbf{e}^T\boldsymbol{\xi} - \sum_{j=1}^{n} \beta_j (R^2 + \xi_j - \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|^2) - \boldsymbol{\mu}^T\boldsymbol{\xi}$$

with Lagrange multipliers $\boldsymbol{\beta},\boldsymbol{\mu} \geq \mathbf{0}$. Setting the derivative of $\mathcal{L}$ with respect to R, $\mathbf{a}$ and $\boldsymbol{\xi}$ to zero, it follows

$$\mathbf{e}^T\boldsymbol{\beta} = 1,$$

$$\mathbf{a} = \sum_{j=1}^{n} \beta_j \boldsymbol{\phi}(\mathbf{x}_j), \tag{44}$$

$$\boldsymbol{\beta} = C\mathbf{e} - \boldsymbol{\mu}. \tag{45}$$

Using these equations, the Lagrangian yields the dual problem

$$\begin{aligned} &\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \left(W(\boldsymbol{\beta}) := \sum_{j=1}^{n} \beta_j \|\boldsymbol{\phi}(\mathbf{x}_j)\|^2 - \sum_{j,k=1}^{n} \beta_j\beta_k \langle \boldsymbol{\phi}(\mathbf{x}_j), \boldsymbol{\phi}(\mathbf{x}_k)\rangle \right) \\ &\text{subject to} \quad \mathbf{e}^T\boldsymbol{\beta} = 1 \\ &\qquad\qquad\quad\; \mathbf{0} \leq \boldsymbol{\beta} \leq C\mathbf{e}. \end{aligned}$$

By (27) and the definition of the feature map, the function $W(\boldsymbol{\beta})$ can be rewritten as

$$W(\boldsymbol{\beta}) = \sum_{j=1}^{n} \beta_j K(\mathbf{x}_j,\mathbf{x}_j) - \sum_{j,k=1}^{n} \beta_j\beta_k K(\mathbf{x}_j,\mathbf{x}_k).$$

In our applications we are mainly interested in isotropic kernels $K(\mathbf{x},\mathbf{y}) = k(\|\mathbf{x}-\mathbf{y}\|)$, e.g. in the Gaussian kernel $K(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$. These kernels $K(\mathbf{x},\mathbf{x}) = \kappa$ for some $\kappa > 0$ and all $\mathbf{x} \in \mathbb{R}^d$. Then $W(\boldsymbol{\beta})$ can be further simplified to

$$W(\boldsymbol{\beta}) = \kappa - \boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta}$$

so that we finally have to solve the dual optimisation problem

$$\begin{aligned} &\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} \\ &\text{subject to} \quad \mathbf{e}^T\boldsymbol{\beta} = 1 \\ &\qquad\qquad\quad\; \mathbf{0} \leq \boldsymbol{\beta} \leq C\mathbf{e}. \end{aligned} \tag{46}$$

Note that this problem coincides with our optimisation problem (42) except for the first constraint $\mathbf{e}^T\boldsymbol{\beta} = 1$. The Kuhn-Tucker conditions for problem (41) are

$$\begin{aligned} \mu_j \xi_j &= 0\,, \quad j = 1, \ldots, n, &(47)\\ \beta_j(R^2 + \xi_j - \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|^2) &= 0\,, \quad j = 1, \ldots, n. &(48) \end{aligned}$$

For $0 < \beta_j < C$, equations (45) and (47) imply that $\mu_j > 0$ and thereby $\xi_j = 0$. Now it follows by (48) that R^2 can be computed as

$$\begin{aligned} R^2 &= \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|^2 \\ &\overset{(44)}{=} K(\mathbf{x}_j, \mathbf{x}_j) + \sum_{i,k=1}^{n} \beta_i \beta_k K(\mathbf{x}_i, \mathbf{x}_k) - 2 \sum_{k=1}^{n} \beta_k K(\mathbf{x}_j, \mathbf{x}_k) \\ &= \kappa + \boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta} - 2(\mathbf{K}\boldsymbol{\beta})_j. \end{aligned}$$

Let us turn to the SV novelty detection problem investigated by Schölkopf et al. in [37]. We are looking for a decision function

$$f(\mathbf{x}) = a(\mathbf{x}) + b := \sum_{j=1}^{n} \beta_j K(\mathbf{x}, \mathbf{x}_j) + b. \tag{49}$$

with bias term b which solves the modified single-class SVM problem

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^n, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^n} \; & C\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\|a(\mathbf{x})\|^2_{\mathcal{H}_K} + b = C\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} + b \\ \text{subject to} \quad & a(\mathbf{x}_j) + b \geq 1 - \xi_j\,, \quad j = 1, \ldots, n \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \tag{50}$$

Analogous to the SV clustering problem, we build the Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) := C\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} + b - \sum_{j=1}^{n} \alpha_j (a(\mathbf{x}_j) + b - 1 + \xi_j) - \boldsymbol{\mu}^T\boldsymbol{\xi}$$

with Lagrange multipliers $\boldsymbol{\alpha}, \boldsymbol{\mu} \geq \mathbf{0}$. Setting the derivative of $\mathcal{L}$ with respect to b, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ to zero, it follows

$$\mathbf{e}^T\boldsymbol{\alpha} = 1, \tag{51}$$

$$\boldsymbol{\alpha} = \boldsymbol{\beta} \overset{(51)}{\Rightarrow} \mathbf{e}^T\boldsymbol{\beta} = 1, \tag{52}$$

$$\boldsymbol{\alpha} = C\mathbf{e} - \boldsymbol{\mu} \overset{(52)}{\Rightarrow} \boldsymbol{\beta} = C\mathbf{e} - \boldsymbol{\mu}. \tag{53}$$

Using these equations, the Lagrangian yields the dual problem

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^n} \; & \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} - 1 \\ \text{subject to} \quad & \mathbf{e}^T\boldsymbol{\beta} = 1 \\ & \mathbf{0} \leq \boldsymbol{\beta} \leq C\mathbf{e}. \end{aligned} \tag{54}$$

This problem is obviously equivalent to the dual SV clustering problem (46). We summarise:

Lemma 3. *Let K be a kernel with corresponding feature map $\boldsymbol{\phi}$ and with the property that $K(\mathbf{x}, \mathbf{x}) = \kappa$ for all $\mathbf{x} \in \mathbb{R}^d$. Then the optimisation problems (41) and (50) are equivalent in that they lead to the same dual problem (46).*

From the dual solution $\boldsymbol{\beta}$ of (46), the primal solution $\mathbf{a}, R, \boldsymbol{\xi}$ of (41) may be obtained by (44) and (43) and the Kuhn-Tucker conditions (47) and (48). The optimal parameter values $b, \boldsymbol{\xi}$ for problem (50) may be obtained by the Kuhn-Tucker complementary conditions as well.

This lemma was also proved in [37]. Further, Vapnik ([46, Section 10.7]) already showed that R^2 can be computed as described by (43) with problem (46) for hard margin ($C = \infty$).

At first sight, it is astonishing that although the quadratic optimisation problems for SV clustering and SV novelty detection are deviated from quite different initial problems (41) and (50), they are equivalent. The paper [33] provides a nice geometrical interpretation for that. The above condition on the kernel implies that all feature vectors lie on a sphere centred at the origin. The hyperplane that separates the data from the origin with maximal margin will then be spanned by the smallest enclosing spheres centre, that is

$$a(\mathbf{x}) = f_{\mathbf{a}}(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle$$

and, in the hard margin case, the distance of the spheres centre from the origin is $\sqrt{1-b}$ so that the relation between R and b reads

$$R^2 + (1-b) = \kappa - C\mathbf{e}^T\boldsymbol{\xi},$$

where $\boldsymbol{\xi}$ are the residuals with respect to the novelty detection problem. (As a matter of fact, the residuals only differ by a factor of two because of the quadratic constraint terms in the clustering problem.) For the hard margin case ($C = \infty$) and the Gaussian kernel, this implies that

$$R^2 = b.$$

C.2 Equivalence of the SV Novelty Detection Problem and the Single-Class SVM without Bias Term

The previous subsection shows the equivalence of the SV clustering problem which can be used for radius computation to a modified SVM (50) with bias term. We will now show that this special problem is equivalent to a single-class SVM without bias term. With $a(\mathbf{x})$ defined as in (49), the common single-class SVM is described by the problem

$$\begin{aligned} &\min_{\boldsymbol{\beta}\in\mathbb{R}^n, \boldsymbol{\xi}\in\mathbb{R}^n} \hat{C}\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\|a(\mathbf{x})\|^2_{\mathcal{H}_K} = \hat{C}\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} \\ &\text{subject to} \quad a(\mathbf{x}_j) \geq 1 - \xi_j, \quad j = 1, \ldots, n, \quad \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \tag{55}$$

By setting up the Lagrangian as above, the traditional single-class SVM leads to the dual quadratic problem (42).

Lemma 4. *There exists $\hat{C} > 0$ such that the SV novelty detection problem (50) with parameter C is equivalent to the standard SVM problem (55) with parameter $\hat{C}$ in that the solutions are derivable from one another. The dual solutions $\boldsymbol{\alpha}$ of (42) and $\boldsymbol{\beta}$ of (54) are related by*

$$\boldsymbol{\beta} = \frac{\boldsymbol{\alpha}}{\mathbf{e}^T\boldsymbol{\alpha}} \tag{56}$$

or conversely by $\boldsymbol{\alpha} = \frac{1}{1-b}\boldsymbol{\beta}$ with the primal variable b.

Proof. The proof consists of two parts. Firstly, the dual solution of the biased SVM (54) will be derived from the dual solution of the SVM without bias (42). Secondly, the primal solution of the unbiased SVM (55) will be derived from the primal solution of the biased SVM (50).

1. Suppose problem (42) is solved by $\boldsymbol{\alpha}$. With $a := \mathbf{e}^T\boldsymbol{\alpha} > 0$, set $\boldsymbol{\beta} := \frac{\boldsymbol{\alpha}}{a}$. Then $\boldsymbol{\beta}$ is valid in problem (54) if $C = \frac{\hat{C}}{a}$. Suppose that $\boldsymbol{\beta}$ is not the optimal solution of problem (54), i.e., there exists some $\tilde{\boldsymbol{\beta}}$ satisfying $\mathbf{e}^T\tilde{\boldsymbol{\beta}} = 1, \mathbf{0} \leq \tilde{\boldsymbol{\beta}} \leq C\mathbf{e}$ so that

$$\begin{aligned} &\frac{1}{2}\tilde{\boldsymbol{\beta}}^T\mathbf{K}\tilde{\boldsymbol{\beta}} < \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} \\ \Rightarrow\ &\frac{1}{2}(a\tilde{\boldsymbol{\beta}})^T\mathbf{K}(a\tilde{\boldsymbol{\beta}}) < \frac{1}{2}(a\boldsymbol{\beta})^T\mathbf{K}(a\boldsymbol{\beta}) \\ \Rightarrow\ &\frac{1}{2}\tilde{\boldsymbol{\alpha}}^T\mathbf{K}\tilde{\boldsymbol{\alpha}} - a < \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} - a, \end{aligned}$$

where $\tilde{\boldsymbol{\alpha}} := a\tilde{\boldsymbol{\beta}}$. Since $\tilde{\boldsymbol{\alpha}}$ fulfils $\mathbf{0} \leq \tilde{\boldsymbol{\alpha}} \leq \hat{C}\mathbf{e}$ and $a = \mathbf{e}^T\tilde{\boldsymbol{\alpha}}$ holds, this is a contradiction to the assumption that $\boldsymbol{\alpha}$ is the optimal solution of problem (42).

2. On the other hand, let $(\boldsymbol{\beta}, b, \boldsymbol{\xi}^{\boldsymbol{\beta}})$ be the optimal solution of the primal problem (50). Then $\boldsymbol{\alpha} := \frac{1}{1-b}\boldsymbol{\beta}$, $\boldsymbol{\xi}^{\boldsymbol{\alpha}} := \frac{1}{1-b}\boldsymbol{\xi}^{\boldsymbol{\beta}}$ is a valid solution for problem (55). Note that $b < 1$ due to the dual constraint $\mathbf{e}^T\boldsymbol{\beta} = 1$ and the Kuhn-Tucker conditions. Assume that $\tilde{\boldsymbol{\alpha}}$ is valid for (55) as well, then the vector $(\tilde{\boldsymbol{\beta}}, b, \boldsymbol{\xi}^{\tilde{\boldsymbol{\beta}}}) := ((1-b)\tilde{\boldsymbol{\alpha}}, b, (1-b)\boldsymbol{\xi}^{\tilde{\boldsymbol{\alpha}}})$ is valid for problem (50). Now we obtain for $\hat{C} = \frac{C}{1-b}$

$$
\begin{aligned}
&\frac{1}{2}\tilde{\boldsymbol{\alpha}}^T\mathbf{K}\tilde{\boldsymbol{\alpha}} + \hat{C}\mathbf{e}^T\boldsymbol{\xi}^{\tilde{\boldsymbol{\alpha}}} < \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \hat{C}\mathbf{e}^T\boldsymbol{\xi}^{\boldsymbol{\alpha}} \\
\Leftrightarrow\ &\frac{1}{2}((1-b)\tilde{\boldsymbol{\alpha}})^T\mathbf{K}((1-b)\tilde{\boldsymbol{\alpha}}) + (1-b)\hat{C}\mathbf{e}^T\left((1-b)\boldsymbol{\xi}^{\tilde{\boldsymbol{\alpha}}}\right) < \\
&\frac{1}{2}((1-b)\boldsymbol{\alpha})^T\mathbf{K}((1-b)\boldsymbol{\alpha}) + (1-b)\hat{C}\mathbf{e}^T\left((1-b)\boldsymbol{\xi}^{\boldsymbol{\alpha}}\right) \\
\Leftrightarrow\ &\frac{1}{2}\tilde{\boldsymbol{\beta}}^T\mathbf{K}\tilde{\boldsymbol{\beta}} + C\mathbf{e}^T\boldsymbol{\xi}^{\tilde{\boldsymbol{\beta}}} + b < \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} + C\mathbf{e}^T\boldsymbol{\xi}^{\boldsymbol{\beta}} + b.
\end{aligned}
$$

Consequently, since $(\boldsymbol{\beta}, b, \boldsymbol{\xi}^{\boldsymbol{\beta}})$ is the optimal solution for problem (50), $\boldsymbol{\alpha}$ is the optimal solution of (55). □

So far, we have shown that for special values of C depending on the solution of the problem, the biased and unbiased single-class SVMs are equivalent. Anyway, as C is a tuning parameter that cannot be determined analytically, this condition does not restrain the equivalence. Especially, for $C = \infty$, the hard margin case, no condition with respect to the weight factor has to be taken into account.

References

[1] R. Azencott, J.-P. Wang, and L. Younes. Texture classification using windowed fourier filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):148–153, 1997.

[2] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. A support vector method for clustering. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 367–373. MIT Press, 2001.

[3] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 196–202. MIT Press, 2001.

[4] P. Bradley. *Mathematical Programming Approaches to Machine Learning and Data Mining*. PhD thesis, University of Wisconsin, Computer Sciences Department, Madison, WI, USA, 1998. TR-98-11.

[5] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2001.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

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

[8] N. Cristianini, J. Shawe-Taylor, and A. Elisseeff. On kernel-target alignment. In *Advances in Neural Information Processing Systems, volume 14*, 2001.

[9] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Commun. on Pure and Appl. Math.*, 41:909–996, 1988.

[10] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[11] K. Duan, S. S. Keerthi, and A. N. Poo. Evaluation of simple performance measures for tuning SVM hyper parameters. Technical Report CD-01-11, Department of Mechanical Engineering, National University of Singapore, 10, Kent Ridge Crescent, 119260 Singapore, 2001.

[12] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, New York, second edition, 2000.

[13] D. Dunn, W. E. Higgins, and J. Wakeley. Texture segmentation using 2-d gabor elementary functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):130–149, 1994.

[14] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[15] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New York, second edition, 1987.

[16] S. Floyd and M. K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, December 1995.

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

[18] M. Heiler. Optimization criteria and learning algorithms for large margin classifiers. Master's thesis, Universität Mannheim, October 2001.

[19] R. Herbrich. *Learning kernel classifiers: theory and algorithms*. MIT Press, Cambridge, Massachusetts, 2002.

[20] R. Herbrich and T. Graepel. A PAC-bayesian margin bound for linear classifiers: Why SVMs work. In *Advances in Neural Information System Processing 13*, pages 224–230, 2001.

[21] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[22] T. Joachims. Making large–scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184. MIT Press, Cambridge, MA, 1999.

[23] E. Jones, P. Runkle, N. Dasgupta, L. Couchman, and L. Carin. Genetic algorithm wavelet design for signal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):890–895, 2001.

[24] G. S. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *J. Math. Anal. Applications*, 33(1):82–95, 1971. Kimeldorf and Wahba's Lemma 6.1 on variational problems in reproducing kernel spaces with linear inequality constraints is highly relevant to modern work on kernel methods in support vector machines.

[25] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. In *Proceedings of 19th International Conference on Machine Learning*, pages 323–330, 2002.

[26] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, London, 1999.

[27] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000.

[28] T. Randen and J. H. Husøy. Filtering for texture classification: A comparative study. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(4):291–310, April 1999.

[29] T. R. Reed and J. H. du Buf. A review of recent texture segmentation and feature extraction techniques. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 57:359–372, 1993.

[30] R. Schaback. Creating surfaces from scattered data using radial basis functions. In M. Daehlen, T. Lyche, and L. L. Schumaker, editors, *Mathematical Methods for Curves and Surfaces*, pages 477–496. Vanderbilt University Press, Nashville, 1995.

[31] P. Scheunders, S. Livens, G. Van de Wouwer, P. Vautrot, and D. Van Dyck. Wavelet-based texture analysis. *International Journal on Computer Science and Information Management*, 1(2):22–34, 1998.

[32] B. Schölkopf. *Support Vector Learning*. PhD thesis, Technische Universität Berlin, 1997.

[33] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. Technical Report 99-87, Microsoft Research, 1999. Short version appeared in *Neural Computation*, 2001.

[34] B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Kernel-dependent support vector error bounds. In *Proceedings of the Ninth International Conference on Artificial Neural Networks 470*, pages 304–309, London, 1999. IEE Conference Publications.

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

[36] 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 Transactions on Signal Processing*, 45(11):2758–2765, 1997.

[37] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12: Proceedings of the 1999 Conference*, pages 582–588. MIT Press, 2000.

[38] G. Smith. MeasTex image texture database and test suite. Available at `http://www.cssip.uq.edu.au/meastex/meastex.html`, May 1997. Version 1.1.

[39] I. Steinwart. On the influence of the kernel on the generalization ability of support vector machines. Technical Report 01-01, FSU Jena, 2001.

[40] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley, 1996.

[41] D. Strauß and G. Steidl. Hybrid wavelet-support vector classification of waveforms. *Journal of Computational and Applied Mathematics*, 148:375–400, 2002.

[42] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, London, UK, 1999.

[43] M. Unser. Texture classification and segmentation using wavelet frames. *IEEE Transactions on Image Processing*, 4(11):1549–1560, 1995.

[44] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, Englewood Cliffs, NJ, 1993.

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

[46] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[47] M. Vetterli and J. Kovačević. *Wavelets and Subband Coding*. Signal Processing. Prentice Hall, Englewood Cliffs, NJ, 1995.

[48] G. Wahba. Support vector machines, reproducing kernel hilbert spaces and the randomized GACV. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, chapter 6, pages 69–88. MIT Press, Cambridge, MA, 1999.

[49] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In *NIPS*, pages 668–674, 2000.