ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG
INSTITUT FÜR INFORMATIK
Lehrstuhl für Mustererkennung und Bildverarbeitung

Feature Space Interpretation of SVMs with non Positive Definite Kernels

Internal Report 1/03

Bernard Haasdonk

October 2003

Feature Space Interpretation of SVMs with non Positive Definite Kernels

Bernard Haasdonk
Computer Science Department
Albert-Ludwigs-University Freiburg
79110 Freiburg, Germany
haasdonk@informatik.uni-freiburg.de

October 6, 2003

Abstract

The widespread habit of “plugging” arbitrary symmetric functions as kernels in support vector machines (SVMs) often yields good empirical classification results. However, in case of non conditionally positive definite (non-cpd) functions they are hard to interpret due to missing geometrical and theoretical understanding. In this paper we provide a step towards comprehension of SVM classifiers in these situations. We give a geometric interpretation of SVMs with non-cpd kernel functions. We show that such SVMs are optimal hyperplane classifiers not by margin maximization but by minimization of distances between convex hulls in pseudo-Euclidean spaces. This interpretation is basis for further analysis, e.g. investigating uniqueness or characterizing situations where SVMs with non-cpd kernels are suitable or not.

Keywords: support vector machine, indefinite kernel, pseudo-Euclidean space, separation of convex hulls, pattern recognition

1 Introduction

In recent years SVMs have been established as methods of first choice on various learning problems like classification or regression in many fields of applications, cf. [4]. There are several reasons for their success ranging from theoretical foundation in statistical learning theory to availability of easily applicable and fast implementations. A very important reason is the clear intuitive geometric interpretation as maximizing the margin of a hyperplane classifier in an (implicitly defined) Euclidean feature space. This is the basis for general understanding, adequate practical application, improvements and new algorithms. However, this geometric interpretation is only available in case of conditionally positive definite (cpd) kernel functions (cf. Section 2 for definitions).

In practice, the requirement of a kernel function to be cpd turns out to be a very strict assumption. It is not satisfied by many ad-hoc or sophisticated similarity or dissimilarity measures which one would like to incorporate in learning. So non-cpd kernels often are available, but it is not clear what is the best way to use them in the SVM framework.

A practical "heuristic" approach is using the kernels in SVMs as usual, e.g. [1, 5, 8, 15]. Problems like nonconvexity of the optimization problem can be handled e.g. by the widespread SVM implementation *libsvm*, as convergence to stationary points is guaranteed [10]. The empirical classification results of such non-cpd kernels often are very good, but theoretical foundation is missing.

The motivation for this work now stems from these two facts: Good empirical results demand for theoretical understanding, and geometry is a fundamental step towards such understanding. We therefore concentrate on providing a geometric interpretation of training and classification of SVMs with non-cpd kernels.

The structure of the paper is as follows: In the next section we introduce the necessary notations concerning SVM and the pseudo-Euclidean spaces which is the framework where the SVM will be interpreted. Section 3 then illustrates (independent of SVMs) linear classification in these spaces by minimizing the distance of convex hulls and giving the corresponding primal optimization problem. In Section 4 we then illustrate that SVM-classification exactly coincides with the pseudo-Euclidean convex hull classification, and we present examples of the correspondences. Section 5 comments on uniqueness of solutions, in Section 6 we set up criteria for deciding when a non-cpd SVM is suitable or not and conclude with final remarks in Section 7.

2 Notation

We will use the following general notations and terminology. x, b, f, etc. denote general variables or unstructured objects. $\mathbf{w}^T, \mathbf{M}^T$ stand for the transpose of a vector $\mathbf{w}$ or a matrix $\mathbf{M}$. $\mathbf{I}_n$ is the $n \times n$ identity matrix. $\mathbf{1}_p, \mathbf{0}_p, \mathbf{e}_i \in \mathbb{R}^p$ denote the vector of ones, zeros and the i-th unit-vector. diag$(\mathbf{v}_1, \ldots, \mathbf{v}_m)$ is the diagonal matrix with entries given by the concatenation of the vectors $\mathbf{v}_i$. Sums will be abbreviated by $\sum_i := \sum_{i=1}^n$, $\sum_{i,j} := \sum_{i,j=1}^n$.

Having a maximization problem $\max_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha})$ under some constraints, we will use the notion *feasible point* $\boldsymbol{\alpha}$ for a point satisfying the constraints of the optimization problem. A *feasible direction* $\mathbf{v}$ in $\boldsymbol{\alpha}$ will be a direction such that $\boldsymbol{\alpha} + \lambda\mathbf{v}$ is a feasible point for some range $\lambda \in [0, \epsilon]$ with some $\epsilon > 0$. A *stationary point* $\boldsymbol{\alpha}$ of the maximization problem is a point where the derivatives of the optimization function in direction of all feasible directions are nonpositive. A *local optimum* $\boldsymbol{\alpha}$ of a maximization problem is a stationary point where additionally the curvature of the optimization function in feasible directions with vanishing directional derivative is nonpositive. A stationary point includes possible saddle-points, which is the reason for considering the second order condition for local optima.

2.1 SVM

A fundamental ingredient in SVMs is the notion of a *kernel* k, which is usually a symmetric function k taking two arguments of an arbitrary set $\mathcal{X}$ where the data stems from, i.e. k :

$\mathcal{X}^2 \to \mathbb{R}$. For given data points $(x_i)_{i=1}^n \in \mathcal{X}^n$ the *kernel matrix* $\mathbf{K} := (k(x_i, x_j))_{i,j=1}^n$ can be defined. If for all n, all sets of data points and all vectors $\mathbf{v} \in \mathbb{R}^n$ holds $\mathbf{v}^T\mathbf{K}\mathbf{v} \geq 0$, then k is called *positive definite*. If this only is satisfied for all $\mathbf{v}$ with $\mathbf{1}_n^T\mathbf{v} = 0$, the kernel k is called *conditionally positive definite*.

Assuming training data $(x_i, y_i) \in \mathcal{X} \times \{\pm 1\}$ for $i = 1, \ldots, n$, the usual SVM-classification approach solves the dual optimization problem which we will refer to as (SVM-DU)

$$\begin{array}{cl} \max\limits_{\alpha_1,\ldots,\alpha_n} & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j k(x_i, x_j) \\ \text{s.t.} & 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_i \alpha_i y_i = 0. \end{array}$$

Here $C > 0$ is a factor penalizing data fitting errors. The classification of new patterns x is then based on the sign of

$$f(x) = \sum_i \alpha_i y_i k(x_i, x) + b \tag{1}$$

where b is determined such that f has identical absolute values on unbounded support vectors (SVs), i.e. x_i with $0 < \alpha_i < C$. In case of a cpd kernel k this procedure is well established and can be understood as optimal hyperplane classification in a Euclidean space, e.g. after the kernel PCA-map [14].

In the following we give up the cpd assumption and let k be an arbitrary symmetric function. In this case the optimization problem (SVM-DU) is no longer convex, but still quadratic. Starting with k, a corresponding *squared distance* can be defined by

$$d^2(x, x') := k(x, x) - 2k(x, x') + k(x', x'). \tag{2}$$

In case of a cpd kernel this corresponds to the induced distance in Euclidean feature space. For general symmetric kernels this definition will not define the square of a metric, as d^2 might be negative. But at least it yields a symmetric function d^2 with zero diagonal. This squared distance function will allow a representation of the data in certain vector spaces.

2.2 Pseudo-Euclidean Spaces

As the geometry of our main argumentation is taking place in pseudo-Euclidean spaces, we shortly recall some basic notions and corresponding illustrations, for details see [6, 12]. The relevance of these spaces is that they provide a unifying framework for both structural and vectorial data after appropriate embeddings, cf. [6]. In particular they can represent quite arbitrary, unstructured data, only assuming a squared distance function on the data, cf. Proposition 1.

With $\mathbb{R}^{(p,q)}$ we denote the pseudo-Euclidean space of signature (p, q), where $p, q \in \mathbb{N}_0$. This space can be seen as a product of a “real” and “imaginary” Euclidean vector space $\mathbb{R}^p \times i\mathbb{R}^q$, where $i = \sqrt{-1}$. Its elements are denoted with $\mathbf{z} = (\mathbf{z}_p^T, \mathbf{z}_q^T)^T$, the vector of real coordinates with respect to the basis $\{\mathbf{e}_k\}_{k=1}^p$ and $\{i\mathbf{e}_l\}_{l=1}^q$ of $\mathbb{R}^p$ resp. $i\mathbb{R}^q$. The *inner product* then is $\langle \mathbf{z}, \mathbf{z}' \rangle := \mathbf{z}_p^T\mathbf{z}_p' - \mathbf{z}_q^T\mathbf{z}_q' = \mathbf{z}^T\mathbf{M}\mathbf{z}'$ with $\mathbf{M} := \operatorname{diag}(\mathbf{1}_p, -\mathbf{1}_q)$. So this inner product is the difference of two standard Euclidean inner products. As the spaces are linear spaces, we can define *reduced convex hulls* of sets of points similar to [2, 3]

$$\operatorname{conv}_\mu(\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}) := \left\{ \sum_i \alpha_i \mathbf{z}_i | \sum_i \alpha_i = 1 \quad \text{and} \quad 0 \leq \alpha_i \leq \mu \right\}.$$

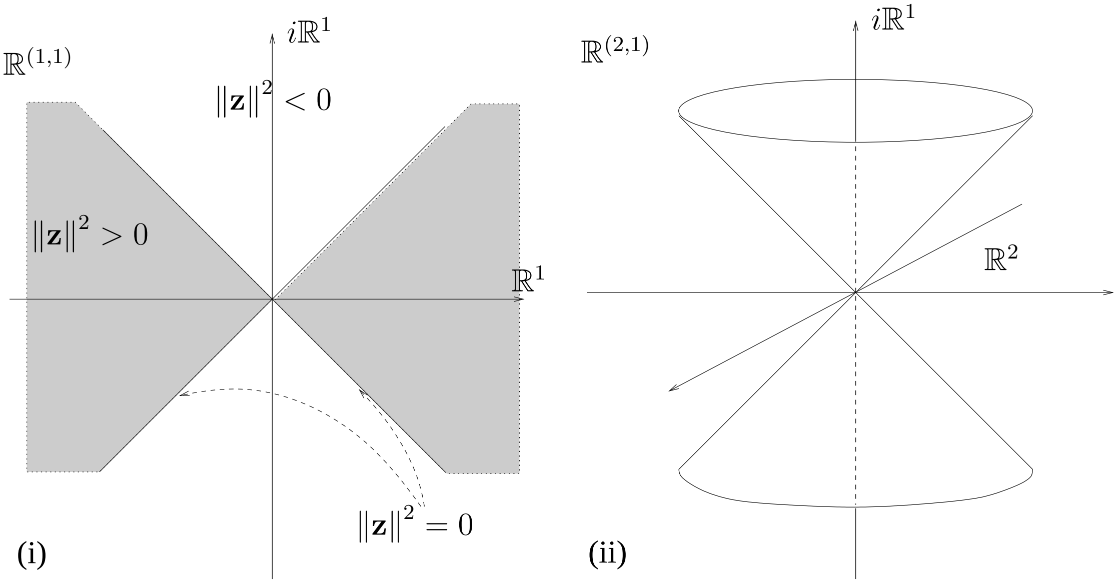

Figure 1: Illustration of pseudo-Euclidean spaces and partition of the spaces by the corresponding isotropic cones.

For $\mu = 1$ this is the usual convex hull. For smaller values μ the set reduces until, for $\mu = 1/n$, it consists of a single point, the mean of the $\mathbf{z}_i$. The shaded regions in Figures 3, 4 and 5 are examples of (reduced) convex hulls. In all illustrations the real space is plotted horizontally, the imaginary part vertically.

Pseudo-Euclidean spaces in particular generalize Euclidean spaces by $q = 0$. We now focus on the case $q > 0$ and the corresponding differences in the resulting geometry. The inner product is a symmetric bilinear form as usual, but no longer positive definite. Nevertheless, similar to the Euclidean case we obtain the *squared norm* as $\|\mathbf{z}\|^2 := \langle \mathbf{z}, \mathbf{z} \rangle$. Note that this can be negative in contrast to the Euclidean case, so it is not a norm in the strict sense. This notion immediately implies the *squared distance* of two points by $\|\mathbf{z} - \mathbf{z}'\|^2 :=< \mathbf{z} - \mathbf{z}', \mathbf{z} - \mathbf{z}' >= \left\|\mathbf{z}_p - \mathbf{z}'_p\right\|^2 - \left\|\mathbf{z}_q - \mathbf{z}'_q\right\|^2$. *Orthogonality* is defined consequently by the inner product of two vectors being zero.

These definitions give rise to some interesting geometric phenomena. The first observation is that there are nonzero points which are orthogonal to themselves: $< \mathbf{z}, \mathbf{z} >= 0$. These *isotropic* points form the *isotropic cone*, which separates the regions of points with positive and negative squared norm, cf. Figure 1.

The squared distance between points is the difference of the corresponding squared distances in real and imaginary directions. This can be negative, so the real square root can not necessarily be defined. Even if it could be defined, the resulting distance would not necessarily be a proper metric, as the triangle inequality does not have to be valid. This is exactly the reason why arbitrary nonmetric distances can be represented in these spaces. Examples of squared distances are given in Figure 2 (i). The shaded region demonstrates a violation of the triangle inequality by $\sqrt{4} > \sqrt{0} + \sqrt{0}$.

The mapping $\mathbf{M}\mathbf{z}$ defines the reflexion of a vector with respect to the real space. With this

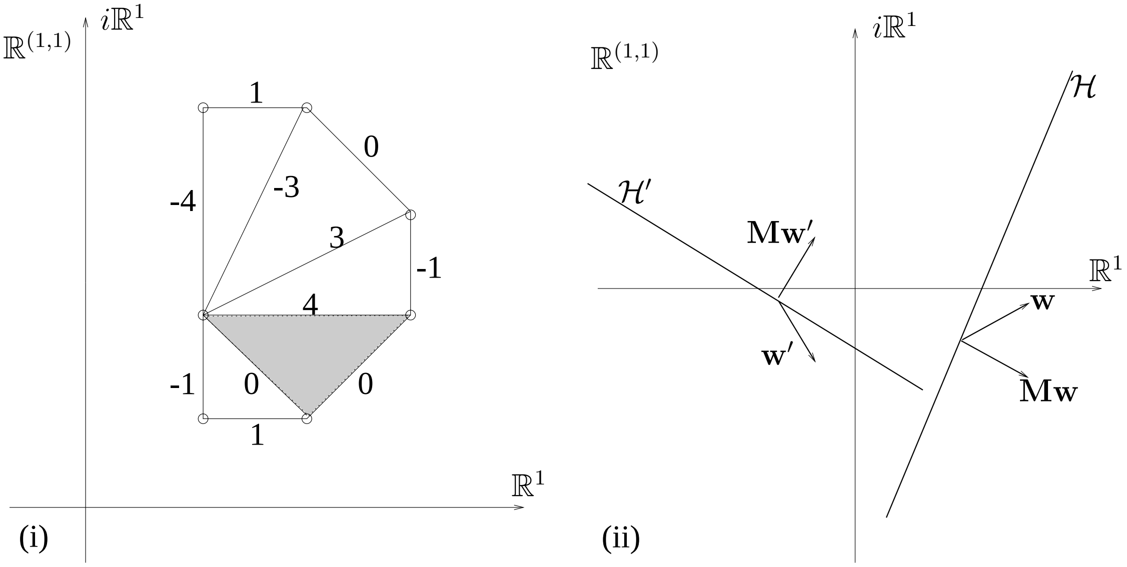

Figure 2: (i) Examples of squared distances between points and a violation of the triangle inequality. (ii) Hyperplanes $\mathcal{H}, \mathcal{H}'$ and corresponding normals $\mathbf{w}, \mathbf{w}'$, the flipped versions of which are ”orthogonal” to their planes in the common Euclidean sense.

in mind, it is obvious that two vectors $\mathbf{z}, \mathbf{z}'$ are orthogonal, if the reflexion $\mathbf{Mz}$ is orthogonal in the Euclidean sense to $\mathbf{z}'$.

Hyperplanes can be defined as usual: $\mathcal{H} : \langle \mathbf{w}, \mathbf{z} \rangle + b = 0$. The normal vector $\mathbf{w}$ is orthogonal to the plane $\mathcal{H}$ in the pseudo-Euclidean sense. Linear classification can easily be performed by taking the sign of any such linear function. Figure 2 (ii) illustrates these aspects.

An embedding of data in a pseudo-Euclidean space $\mathbb{R}^{(p,q)}$, which allows to maintain the distance information, was given by [6] and also used by [7, 12]. For our purpose we mainly need the existence of such a representation. Additionally, a concrete construction following [12] is given in the Appendix. The most important point for our purpose is that p, q are the number of positive resp. negative eigenvalues of a so called *centered* kernel matrix constructed from the distance information.

Proposition 1 (Isometric Embedding). *Let $\{x_i\}_{i=1}^n \in \mathcal{X}^n$ be data points, $d^2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function with zero diagonal. Then there exists a pseudo-Euclidean space $\mathbb{R}^{(p,q)}$ with $p + q < n$ and an embedding $\Phi : \{x_i\}_{i=1}^n \to \mathbb{R}^{(p,q)}$ such that for all i, j holds*

$$d^2(x_i, x_j) = \|\Phi(x_i) - \Phi(x_j)\|^2 . \tag{3}$$

The embedding can be performed such that it is centered (mean $\mathbf{0}$) and the coordinates are uncorrelated. For many embeddings of real data, the variance in the imaginary directions is empirically much lower than the variance in the real directions [12]. Note that the embedding in Proposition 1 does not assume nonnegativity of d^2. So in case of negative squared distances between points, such an isometric representation still exists.

A slightly more abstract framework called Krein- or Pontryagin-spaces [11] for embedding the whole original space $\mathcal{X}$ would also have been an appropriate framework for our interpretation. But as will be demonstrated in the next section, only the space resulting from embedding the (finite) training data is required for understanding the SVM. So we restrict to the more easily accessible class of finite dimensional Krein-spaces, which exactly are the pseudo-Euclidean spaces.

These indefinite spaces provide a representation of data which only depends on a given arbitrary symmetric kernel function. It is independent of the specific algorithm to be applied to this data. We therefore expect that these spaces provide the suitable geometry for investigating other kernel-methods which involve non-cpd functions.

3 Optimal Separation of Convex Hulls in $\mathbb{R}^{(p,q)}$

Different methods of linear classification in pseudo-Euclidean spaces have been proposed, e.g. the Fisher linear discriminant [6] or a generalized nearest mean classifier etc. [12]. Also methods for using SVM in these spaces have been proposed [7, 12]. They have in common that the non-Euclidean geometry of the space is removed and replaced by a Euclidean. This can either be done by interpreting $\mathbb{R}^{(p,q)}$ as the standard $\mathbb{R}^{p+q}$ or projecting the space on its real part with corresponding norm and positive definite scalar product. One obtains nice convex optimization problems by this, but the original geometry which reflects the a-priori-knowledge of the problem is changed by these procedures. Moreover, these methods require explicit operating in the feature space. In contrast, methods that accept the non-Euclidean geometry may result in more difficult optimization problems due to the nonconvexity, but they perfectly maintain the a-priori knowledge given by the kernel/distance. Additionally, they may avoid explicit operations in the feature space.

In this section we present a classification procedure, which has these characteristics. It turns out to be an optimal hyperplane classification method and exactly the operation that is performed by a non-cpd SVM. However, care has to be taken, as in certain situations the scheme is not suitable. We will find criteria for these situations. The classification method is based on convex hulls, therefore we denote it as CH-classification.

It has been shown that maximization of the (soft) margin in Hilbert-spaces can equivalently be formulated as minimization of distances of (reduced) convex hulls [2, 3]. This separation of convex hulls can be applied in pseudo-Euclidean spaces, as we also have the notions of distance and convex hulls.

We assume to have training data $(x_i, y_i) \in \mathcal{X} \times \{\pm 1\}$ for $i = 1, \ldots, n$ which is embedded isometrically in some $\mathbb{R}^{(p,q)}$ according to Proposition 1. The formalization of minimizing the distance in $\mathbb{R}^{(p,q)}$ between the reduced convex hulls of the positive and of the negative training examples is

$$\begin{aligned} \min_{\mathbf{z}^-, \mathbf{z}^+} \quad & \|\mathbf{z}^- - \mathbf{z}^+\|^2 \\ \text{s.t.} \quad & \mathbf{z}^- \in \operatorname{conv}_\mu\{\Phi(x_i) | i : y_i = -1\} \quad \text{and} \\ & \mathbf{z}^+ \in \operatorname{conv}_\mu\{\Phi(x_i) | i : y_i = +1\}. \end{aligned} \tag{4}$$

The optimization only requires distance calculations in the original space between training patterns. This can be seen analogously to the Euclidean case [9] by decomposing the squared norm

(4) due to its bilinearity and replacing $\mathbf{z}^-, \mathbf{z}^+$ by their representations as convex combinations

$$\mathbf{z}^\pm = \sum_{i:y_i=\pm 1} \bar{\alpha}_i \Phi(x_i). \tag{5}$$

We end up with the dual optimization problem minimizing the distance between the convex hulls which we will refer to as (CH-DU)

$$\begin{array}{ll} \max\limits_{\bar{\alpha}_1,\ldots,\bar{\alpha}_j} & \frac{1}{2}\sum_{i,j} \bar{\alpha}_i \bar{\alpha}_j y_i y_j d^2(x_i, x_j) \\ \text{s.t.} & 0 \leq \bar{\alpha}_i \leq \mu, \quad \sum_i \bar{\alpha}_i y_i = 0 \quad \text{and} \quad \sum_i \bar{\alpha}_i = 2. \end{array}$$

Note that this optimization problem is quadratic, but not necessarily convex, as the quadratic form can be indefinite. This can cause phenomena like multiple local optima as will be illustrated in Section 5.

The natural classifier in $\mathbb{R}^{(p,q)}$ based on a feasible point $\bar{\boldsymbol{\alpha}}$ from (CH-DU) is the minimum distance classifier with respect to the two points $\mathbf{z}^+$ and $\mathbf{z}^-$, i.e. the sign of

$$g(\mathbf{z}) = \left\|\mathbf{z} - \mathbf{z}^-\right\|^2 - \left\|\mathbf{z} - \mathbf{z}^+\right\|^2. \tag{6}$$

Similar to the Euclidean case this is a hyperplane classifier. Interestingly, it can be evaluated using only squared distances to (images of) training points. This can be obtained by inserting the representations (5), using bilinearity and the constraints of (CH-DU) which yields

$$g(\mathbf{z}) = -\sum_i \bar{\alpha}_i y_i \left\|\mathbf{z} - \Phi(x_i)\right\|^2 + \frac{1}{2}\sum_{i,j} \bar{\alpha}_i \bar{\alpha}_j y_i \left\|\Phi(x_i) - \Phi(x_j)\right\|^2. \tag{7}$$

In particular this implies that if a point $\mathbf{z}$ has a pre-image $x \in \mathcal{X}$ (i.e. $\mathbf{z} = \Phi(x)$), we can perform the classification in the original space without explicit mapping by taking the sign of

$$f(x) = -\sum_i \bar{\alpha}_i y_i d^2(x_i, x) + b \tag{8}$$

$$\text{with} \quad b = \frac{1}{2}\sum_{i,j} \bar{\alpha}_i \bar{\alpha}_j y_i d^2(x_i, x_j). \tag{9}$$

We now argue why this classification rule can be applied to the whole (possibly infinite) space $\mathcal{X}$. If x has to be classified and x has no image in $\mathbb{R}^{(p,q)}$, we imagine another isometric embedding Φ' in a pseudo-Euclidean space $\mathbb{R}^{(p',q')}$ where x is simultaneously embedded with the training data. As the training procedure (CH-DU) and the classification rule (8) are independent of the specific embedding, training and classification of $\Phi'(x)$ will exactly result in the decision rule (8). At this point we see that it is no limitation that the embedding is data dependent, as we do not explicitly make use of it.

Figure 3 gives an illustration of the classification behaviour of the classifier by plotting simple data embeddings $\{\Phi(x_i)\}_{i=1}^n$ in the low dimensional pseudo-Euclidean space $\mathbb{R}^{(1,1)}$ and the corresponding decision boundaries. The classification boundary is the line passing through the midpoint of $\mathbf{z}^+\mathbf{z}^-$ which is orthogonal (in pseudo-Euclidean sense) to the line connecting

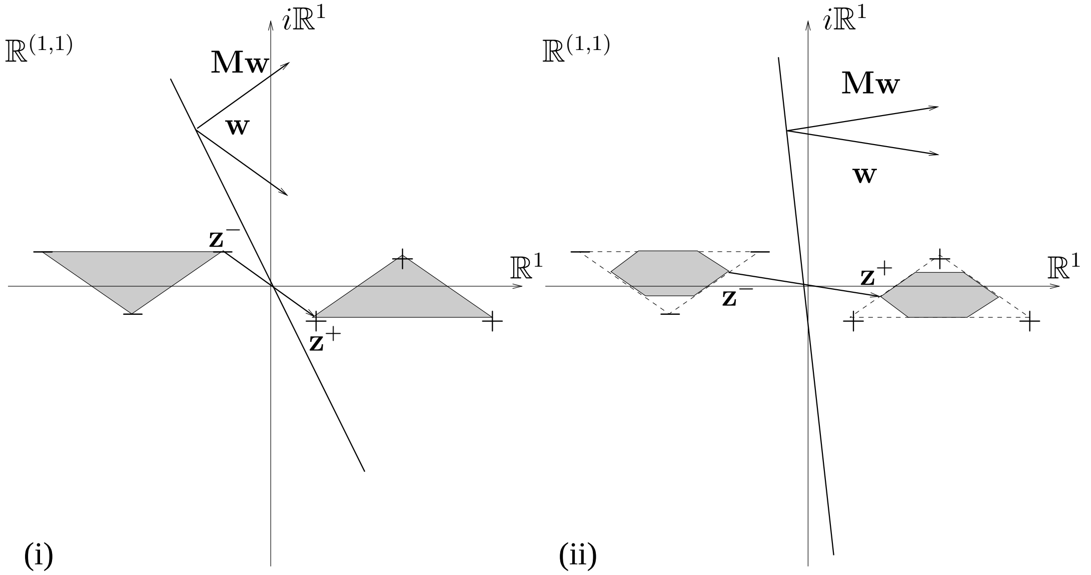

Figure 3: Illustration of pseudo-Euclidean CH-classification with ordinary convex hulls (i) and reduced convex hulls (ii).

$\mathbf{z}^+, \mathbf{z}^-$. So we obtain a classification method which is an optimal hyperplane classifier in the sense that it is the minimum distance classifier with respect to closest points of convex hulls.

(CH-DU) is similar to the ν-SVM dual [13] and we similarly can set up the corresponding convex hull primal optimization problem. We emphasize however, that there is no strict "duality" between primal and dual solutions in nonconvex optimization problems. But we keep the notions to emphasize the relation to the cpd case. Note that the choice of b differs from (9), but we again wanted to keep the standard notation. This result is an extension of the Euclidean case in [3]. The proof is given in the appendix.

Proposition 2 (CH Primal in $\mathbb{R}^{(p,q)}$). *Let $\bar{\boldsymbol{\alpha}}$ be a stationary point of (CH-DU), such that there exist two non-bounded coefficients of different classes, i.e. $0 < \bar{\alpha}_k, \bar{\alpha}_l < \mu$ with $y_k = +1, y_l = -1$, which induce a positive ρ as defined below. Let $\Phi : \{x_i\}_{i=1}^n \to \mathbb{R}^{(p,q)}$ be an isometric embedding according to Proposition 1.*

Then we obtain a stationary point $\mathbf{w} \in \mathbb{R}^{(p,q)}, b \in \mathbb{R}, \rho \in \mathbb{R}_+, \boldsymbol{\xi} \in \mathbb{R}_+^n$ of the convex hull primal optimization problem

$$\min_{\mathbf{w},b,\rho,\boldsymbol{\xi}} \quad \tfrac{1}{2}\mathbf{w}^T\mathbf{M}\mathbf{w} - 2\rho + \mu \textstyle\sum_i \xi_i$$
$$\textit{s.t.} \quad y_i(\mathbf{w}^T\mathbf{M}\Phi(x_i) + b) \geq \rho - \xi_i, \quad \xi_i \geq 0 \quad \textit{and} \quad \rho \geq 0$$

by setting $\mathbf{w} = \sum_i \bar{\alpha}_i y_i \Phi(x_i)$, $b := -\frac{1}{2}\left(\mathbf{w}^T\mathbf{M}\Phi(x_k) + \mathbf{w}^T\mathbf{M}\Phi(x_l)\right)$,

$$\rho := \frac{1}{2}\mathbf{w}^T\mathbf{M}(\Phi(x_k) - \Phi(x_l)) \tag{10}$$

and

$$\xi_i := \begin{cases} \rho - y_i(\mathbf{w}^T\mathbf{M}\Phi(x_i) + b) & \textit{if} \quad \bar{\alpha}_i = \max_j \bar{\alpha}_j \\ 0 & \textit{otherwise.} \end{cases}$$

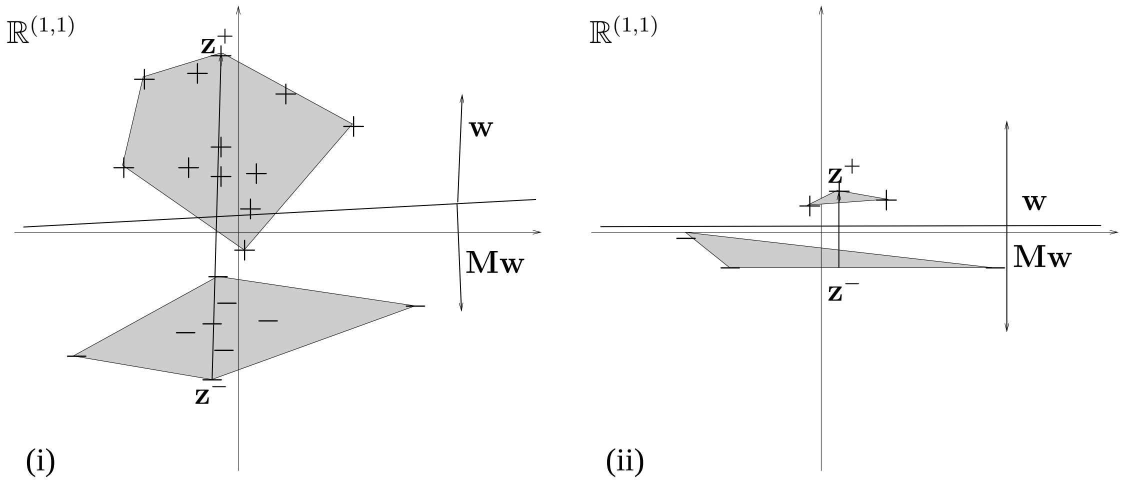

Figure 4: Illustration of unsuitable pseudo-Euclidean CH-classification due to $\rho < 0$.

The well-definedness of b and ρ can be proven. An exact correspondence of local optima can also be achieved under some more restrictive conditions, namely requiring $\Delta\mathbf{w}^T\mathbf{M}\Delta\mathbf{w} \geq 0$ for all feasible directions $\mathbf{v} := (\Delta\mathbf{w}^T, \Delta b, \Delta\rho, \Delta\boldsymbol{\xi}^T)^T$ in which the derivative of the primal optimization function vanishes.

This proposition indicates that a crucial property of a good solution is $\rho > 0$. This means geometrically that according to (10) the vector $\mathbf{w}$ must have positive (pseudo-Euclidean) inner product with all vectors connecting pairs of (unbounded) support vectors $\Phi(x_k) - \Phi(x_l)$. In case of no bounded SV, this is equivalent to the requirement $\mathbf{w}^T\mathbf{M}(\mathbf{z}^+ - \mathbf{z}^-) = \mathbf{w}^T\mathbf{M}\mathbf{w} > 0$, as $\mathbf{z}^\pm$ are convex combinations of SVs. So $\mathbf{M}\mathbf{w}$ must point to the decision region where $\mathbf{z}^+$ is located.

If this is not the case, the resulting classifier will not be suitable for the given problem. Figure 4 demonstrates such situations. We see that CH-classification performs (almost) completely wrong already on the training data. This can be seen as $\mathbf{M}\mathbf{w}$, which is used for classification, is directing towards the negative examples. As the convex hulls of the positive and negative training points can lie above each other, the points of minimum distance ($\mathbf{z}^+$ and $\mathbf{z}^-$) are not longer located on the Euclidean closest boundaries, but on the farthest. A crucial assumption of distance based classification is violated which requires “lower (squared) distance” meaning “higher similarity”: In the example points have negative squared distance to points from the convex hull of the other class, but have 0 distance to themselves.

In the left sketch we see that having many negative d^2 between the training points is a problematic starting point for maximizing the distance between the convex hulls in pseudo-Euclidean spaces. However, this situation can even arise if the original squared distances between the training samples are nonnegative, as occurs in the right illustration.

4 Interpretation of non-cpd SVM in $\mathbb{R}^{(p,q)}$

In this section we establish the geometric interpretation of non-cpd SVMs in pseudo-Euclidean space. We first state the primal optimization problem corresponding to (SVM-DU). The statement is that these SVMs in both separable and nonseparable cases have a similar primal target as ordinary SVMs. The proof is skipped in this presentation as it is similar to the previous proposition.

Proposition 3 (SVM Primal in $\mathbb{R}^{(p,q)}$). *Let $\boldsymbol{\alpha}$ be a stationary point of (SVM-DU) for arbitrary symmetric k, such that there exist two non-bounded coefficients of different classes, i.e. $0 < \alpha_k, \alpha_l < C$ with $y_k = +1, y_l = -1$. Let $\Phi : \{x_i\}_{i=1}^n \to \mathbb{R}^{(p,q)}$ be an isometric embedding according to Proposition 1 corresponding to the squared distance measure induced by k in (2). Then we obtain a stationary point $\mathbf{w} \in \mathbb{R}^{(p,q)}, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n_+$ of the primal optimization problem*

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \tfrac{1}{2}\mathbf{w}^T\mathbf{M}\mathbf{w} + C\sum_i \xi_i \tag{11}$$

$$\textit{s.t.} \quad y_i(\mathbf{w}^T\mathbf{M}\Phi(x_i) + b) \geq 1 - \xi_i \quad \textit{and} \quad 0 \leq \xi_i$$

by setting $\mathbf{w} = \sum_i \alpha_i y_i \Phi(x_i)$, $b := -\frac{1}{2}\mathbf{w}^T\mathbf{M}(\Phi(x_k) + \Phi(x_l))$ and

$$\xi_i := \begin{cases} 1 - y_i(\mathbf{w}^T\mathbf{M}\Phi(x_i) + b) & \textit{if} \quad \alpha_i = C \\ 0 & \textit{otherwise.} \end{cases}$$

The relation between this primal and the common SVM primal is that the common squared norm and inner products are replaced by the corresponding pseudo-Euclidean notions. If $\mathbf{M} = \mathbf{I}_n$ we perfectly recover the common SVM primal.

The relevance of this result is twofold. First we recover a finding from [10], which states that a stationary point of (SVM-DU) satisfies certain separability constraints, which turn out to be equivalent to our constraints of the SVM primal. They imply that a non-cpd SVM is a reasonable classifier in the sense that similar to an ordinary SVM it correctly classifies the training data with $\alpha_i < C$ and $\alpha_i = C$, $\xi_i < 1$. Examples with $\alpha_i = C$, $\xi_i > 1$ are wrongly classified.

The second important conclusion from this proposition is that we found the regularizer $\mathbf{w}^T\mathbf{M}\mathbf{w}$. A geometrical margin can be defined analogous to the cpd case as $2/\sqrt{\mathbf{w}^T\mathbf{M}\mathbf{w}}$, if the regularizer is positive. This might be a starting point for learning theoretic investigations, which could give further insights in usability and provide a further theoretic underpinning of non-cpd SVMs.

An important comment is appropriate at this point. We have a notion of *margin* for a non-cpd SVM and Proposition 3 indicates that training is *related* to maximizing this quantity in the sense that we obtain a stationary point. However, in general, *training of a non-cpd SVM is not identical to margin maximization.* The reason is that margin maximization by (11) is in general not well defined. We give a separable two point example: Take $y_1 = +1, y_2 = -1$ and $\mathbf{x}_i = (y_i, 0)^T \in \mathbb{R}^{(1,1)}$. One can easily check that $\mathbf{w} := (1, \lambda)^T, b = 0, \boldsymbol{\xi} = (0,0)^T$ satisfies the constraints with equality, however the optimization value $\mathbf{w}^T\mathbf{M}\mathbf{w} = 1 - \lambda^2$ diverges to $-\infty$ as λ increases. This is a fundamental difference to the cpd case.

So margin maximization is not the right interpretation of non-cpd SVM, instead optimal separation of convex hulls is adequate, which will be formulated and proven in the following. The main consequence is a justification for using non-cpd SVM. We additionally obtain a constructive interpretation. This allows easy understanding of SVM classification as the operation of separating convex hulls is geometrically easily accessible.

We now present the main result, which settles the relation between solutions of (CH-DU) and (SVM-DU) and corresponding decision boundaries. The proposition states that a local optimum of (SVM-DU) implies a local optimum of (CH-DU) for certain μ. The implication in the other direction however is not valid in general.

For the cpd case the statements follow from [3, 13]. Their proofs however explicitly use the Euclidean primal and positive definiteness, e.g. for optimality only first order derivative conditions have to be checked. We present a detailed proof for the case of arbitrary symmetric k. It emphasizes that the correspondences only rely on properties of the quadratic problems and do not require primal solutions. We use $\mathbf{Q}$ for the matrix with entries $Q_{ij} = y_i y_j k(x_i, x_j)$.

Proposition 4 (Equivalence of SVM and CH). *Let k be an arbitrary symmetric function and d^2 be the induced squared distance as given in (2).*

i) A nonzero stationary point $\boldsymbol{\alpha}$ of (SVM-DU) induces a stationary point $\bar{\boldsymbol{\alpha}} := 2\boldsymbol{\alpha}/\sum_i \alpha_i$ of (CH-DU) with $\mu = 2C/\sum_i \alpha_i$.
If additionally $\boldsymbol{\alpha}$ is a local optimum, then $\bar{\boldsymbol{\alpha}}$ is a local optimum.

ii) A stationary point $\bar{\boldsymbol{\alpha}}$ of (CH-DU) induces a stationary point $\boldsymbol{\alpha} := \bar{\boldsymbol{\alpha}}/\rho$ of (SVM-DU) with $C = \mu/\rho$, if there are two unbounded coefficients of opposite classes, i.e. $0 < \bar{\alpha}_k, \bar{\alpha}_l < \mu, y_k = +1, y_l = -1$, such that $\rho := \frac{1}{2}\bar{\boldsymbol{\alpha}}^T \mathbf{Q}(\mathbf{e}_k + \mathbf{e}_l)$ is positive.
If additionally $\bar{\boldsymbol{\alpha}}$ is a local optimum then $\boldsymbol{\alpha}$ is a local optimum in the case of $\mathbf{Q}$ being positive semidefinite in all feasible directions $\mathbf{v}$ of (SVM-DU) with $\langle \mathbf{1}_n - \mathbf{Q}\boldsymbol{\alpha}, \mathbf{v} \rangle = 0$.

iii) In both cases i) and ii) the corresponding decision planes defined by (8) and (1) are parallel, even identical if $\boldsymbol{\alpha}$ resp. $\bar{\boldsymbol{\alpha}}$ are not upper bounded.

We present some examples which depict the classification behaviour of non-cpd SVMs in illustrative low dimensional feature space $\mathbb{R}^{(1,1)}$. An easy method of obtaining such feature space is by directly assuming $\mathcal{X} = \mathbb{R}^2$ with kernel function $k(\mathbf{x}, \mathbf{x}') := x_1 x_1' - x_2 x_2'$. This kernel function is non-cpd, which can easily be checked. An isometric embedding then is interpreting $\mathbf{x}$ as element of $\mathbb{R}^{(1,1)}$. Note that the illustrated $\mathbf{w}$ compared to the definitions in Prop. 2 and 3 are scaled.

We start with examples, where the corresponding non-cpd SVM is a useful classifier. The situation presented in Figure 3 (i) already illustrated the “hard margin” case: The figure not only demonstrates CH-classification, but this coincides with the corresponding SVM according to Prop. 4 i) resp. ii). This situation is not only restricted to positive d^2, as Figure 5 (i) illustrates the same situation with several negative squared distances within both classes of training points.

Even in the “soft margin” formulation, where some α_i are bounded by C the result of a non-cpd SVM is reasonable. The decision plane is due to Prop. 4 iii) not exact CH-classification, but due to different choice of offset values a decision plane that is slightly shifted. The line

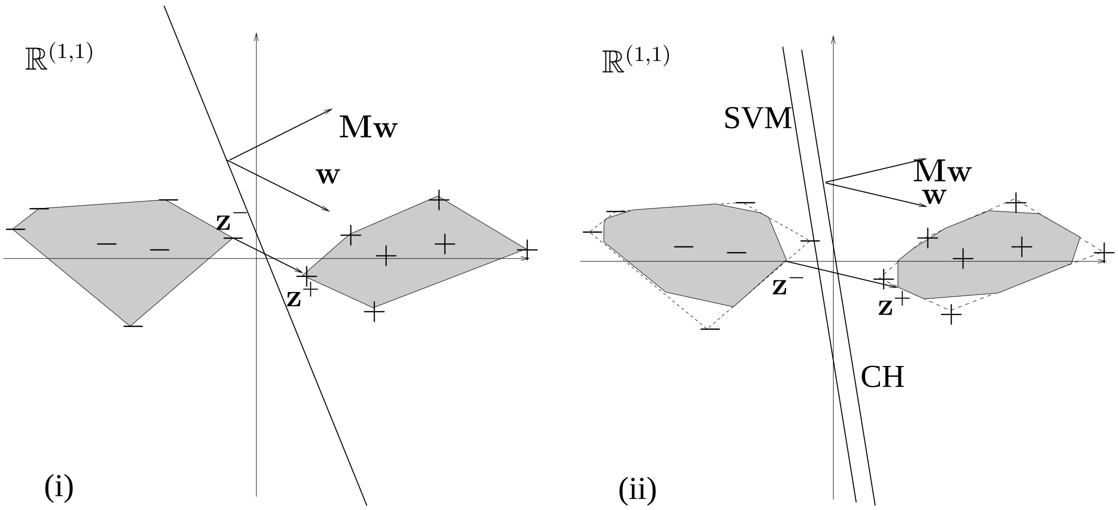

Figure 5: Illustration of non-cpd SVM-classification in case of occasional $d^2 < 0$. (i) identical CH and SVM decision-line, (ii) SVM being parallel to CH.

in Figure 3 (ii) therefore is parallel to the SVM decision boundary. Again, this is not only restricted to $d^2 \geq 0$, Figure 5 (ii) gives a demonstration for the case of some negative squared training distances.

In the previous section we explained that CH-classification is not useful in situations where $\mathbf{w}^T\mathbf{M}\mathbf{w} \leq 0$. On the other hand every SVM decision plane corresponds to some CH-classification plane by simple positive scaling. So the requirement of $\mathbf{w}^T\mathbf{M}\mathbf{w} > 0$ transfers identically to SVM classifiers.

This can be interpreted as follows: For ordinary SVM the requirement for successful application is that the data is reasonably separable by a hyperplane in some implicit Hilbert-space. For non-cpd SVM linear separability is not enough, separability with a hyperplane that has *positive norm* is required.

As an example where the SVM solution violates this condition, we recall Figure 4, which shows unsuitable CH-classification. These two examples do *not* have a correspondence to any SVM-solution, so they are a confirmation that the conditions in statement ii) of Proposition 4 are not superfluous as for the other direction i).

The question arises what is the result of training a non-cpd SVM on the data points given in Figure 4. Due to the equivalence statement Proposition 4 i) there has to be an interpretation as a (parallel of) a CH-classifier. It turns out that during SVM-training most of the α_i get upper bounded by C and the μ according to Proposition 4 i) gets very close to its lowest value. So the SVM-solution is working as the minimum distance classifier based on highly reduced convex hulls, cf. Fig. 6. Again, many of the training points get wrongly classified, which makes the SVM an unsuitable classifier in this situation.

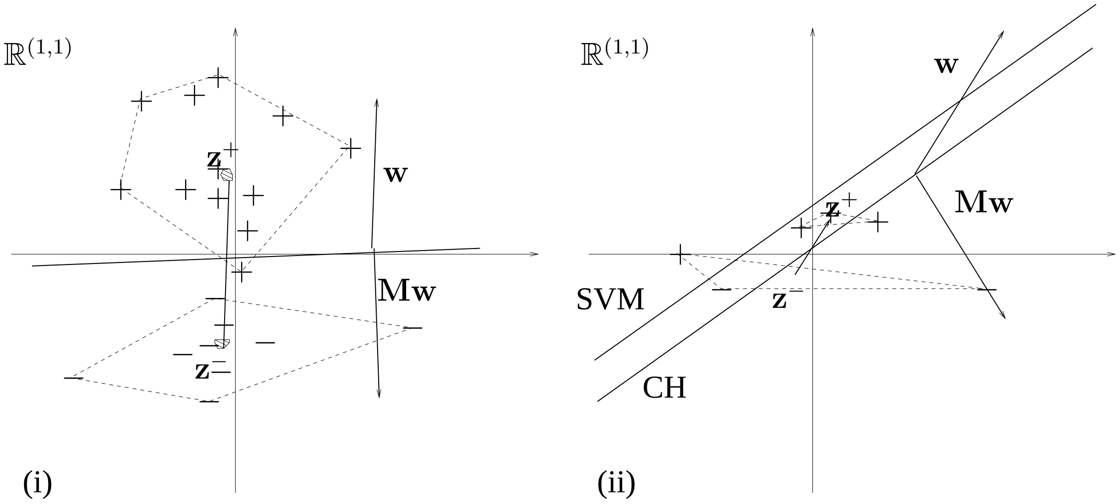

Figure 6: Illustration of unsuitable SVM-classification

5 Uniqueness of Solutions

In the following two sections we want to exemplify how the presented interpretation can serve as basis for further analysis. We start with addressing the question of uniqueness of non-cpd SVM and derive some simple statements.

An easy example in Figure 7 illustrates that in general uniqueness can not be expected. The SVM/CH-classifier has two local optima $\mathbf{w}, \mathbf{w}'$ (with identical value of the optimization function). The resulting classifiers are different, but both reasonable. One can easily see that by shifting $\mathbf{z}^{-\prime}$ down one can obtain arbitrary different norms of $\mathbf{w}'$, so different local solutions can have arbitrary differing values of the optimization function.

In this example the points $\mathbf{z}^-$ and $\mathbf{z}^{-\prime}$ lie on the vertices of a line. By adding further imaginary dimensions, i.e. increasing q, one can easily obtain examples where the multiple solutions $\mathbf{z}^-$ are located on the vertices of a square for $q = 2$, of a cube for $q = 3$ etc. This is the basic idea for the proof of

Lemma 5 (Exponential Number of Local Optima). *For every $q \geq 1$ there exist kernels and points with suitable labeling, such that the data can be embedded to $\mathbb{R}^{(1,q)}$ according to Prop. 1 and the corresponding optimization problem (SVM-DU) has 2^q local optima, which all perform correct separation.*

The other extreme situation is uniqueness of solutions. As simple results we refer to Figures 3 (i) and 5 (i), which demonstrate uniqueness of local optima and stationary points in case of non-cpd SVMs. This uniqueness does not only hold in case of this low dimensional example, but can be extended to $\mathbb{R}^{(1,q)}$, under similar conditions. The proof can again be found in the appendix.

Lemma 6 (Uniqueness of Stationary Points). *If given training data with a non-cpd kernel induces an isometric embedding to $\mathbb{R}^{(1,q)}$ and (SVM-DU) has a stationary point $\boldsymbol{\alpha}$ such that the corresponding $\mathbf{z}^+, \mathbf{z}^-$ have positive squared distance and positive squared distance to all points of their corresponding convex hulls, then $\mathbf{z}^+, \mathbf{z}^-$ are unique.*

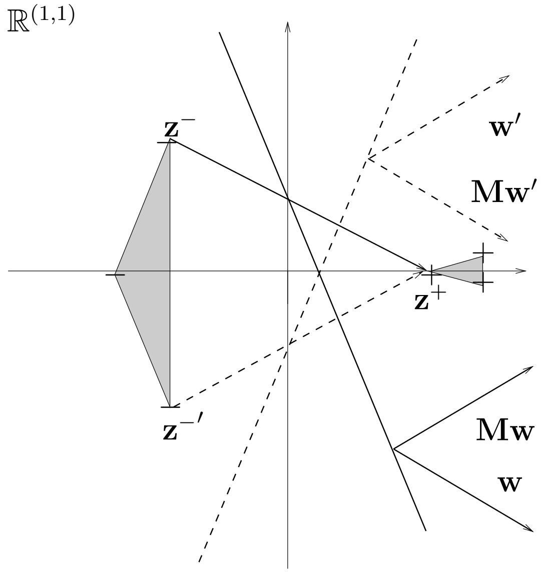

Figure 7: Illustration of multiple local optima of non-cpd SVM-classification.

This particularly implies that the stationary point is the global optimum. This is an improbable situation but a quite remarkable result as it states uniqueness for kernel functions which are extremely non-cpd in the sense that their (centered) kernel matrices have only one positive eigenvalue.

This indicates that the number of local optima may be very well behaved under certain assumptions. More detailed investigations concerning number, quality and relation of local optima or conditions for uniqueness might be promising.

For application of non-cpd SVM, random initializations could be performed to deal with the problem of multiple local solutions. However, it has been found that different local solutions tend to perform very similar [10], which indicates minor importance of randomization.

6 Suitability Criteria

Continuing with practically relevant aspects, we focus on criteria for problem specific suitability of non-cpd SVM. The feature space interpretation provides us with different nontrivial criteria for checking the suitability or unsuitability of a trained SVM or kernel.

Bound on the training error: As we have seen in the interpretation of the SVM primal problem following Proposition 3, the common interpretation of the coefficients and slack variables also holds in the non-cpd case. In particular the ratio of upper bounded coefficients is an upper bound on the training error. So low value of this quantity is an indicator of a suitable (but possibly overfitted) SVM solution.

Sign of $\mathbf{w}^T\mathbf{M}\mathbf{w}$: We have seen that a crucial property of a suitable solution is that $\mathbf{M}\mathbf{w}$ and $\mathbf{w}$ point to identical sides of the decision line. Formally this requires $\mathbf{w}^T\mathbf{M}\mathbf{w} > 0$ or equivalently

$\mathbf{w}$ must have positive squared norm. This quantity can be calculated without explicit mapping to $\mathbb{R}^{(p,q)}$ by using the representation of $\mathbf{w}$ as linear combination of embedded training points.

$$\mathbf{w}^T\mathbf{M}\mathbf{w} = \sum_{i,j} \alpha_i\alpha_j y_i y_j \Phi(x_i)^T\mathbf{M}\Phi(x_j) = \sum_{i,j} \alpha_i\alpha_j y_i y_j k(x_i, x_j). \tag{12}$$

The last reformulation follows from (2) and using $\sum_i \alpha_i y_i = 0$. If this value is nonpositive, it is a criterion for dismissing a given SVM solution.

Number of negative eigenvalues: In some situations it is even possible to predict the unsuitability of a non-cpd SVM before training, or even data-independently. Eq. (12) indicates that a kernel matrix which is the negative of a cpd matrix clearly produces $\mathbf{w}^T\mathbf{M}\mathbf{w} \leq 0$ independent of the y_i or α_i. This is particularly satisfied, if k is the negative of a cpd function. So kernels like $k(\mathbf{x}, \mathbf{x}') = -\langle \mathbf{x}, \mathbf{x}' \rangle$ or $k(\mathbf{x}, \mathbf{x}') = -e^{-\gamma||\mathbf{x}-\mathbf{x}'||^2}$ are not suitable for SVMs, independent of the given training data.

In general, an increasing number of negative eigenvalues of the kernel matrix makes $\mathbf{w}^T\mathbf{M}\mathbf{w} \leq 0$ more likely, so useful separation of the training data is getting more difficult. Therefore the number of negative eigenvalues of the kernel matrix is a rough criterion of how difficult it is to obtain a suitable solution by training the corresponding SVM.

7 Conclusion

We have shown that using SVMs with arbitrary symmetric kernels, in particular non-cpd kernels, is not only a heuristic procedure, but has a reasonable interpretation as optimal hyperplane classifiers in pseudo-Euclidean spaces. They are minimum distance classifiers with respect to certain points from the convex hulls of embedded training points. This interpretation already existed in the Euclidean case, we extended it to the pseudo-Euclidean case. We explained that non-cpd SVMs in general *cannot* be seen as margin maximizers, although a notion of margin can be defined.

The interpretation results in a constructive method to illustrate the classification behaviour of an SVM in the corresponding pseudo-Euclidean space. We demonstrated how the geometric understanding can serve as basis for further analysis. To exemplify this we commented on uniqueness of the local solutions in certain situations. We further gave practically relevant criteria for checking whether a given non-cpd SVM classifier is promising or can clearly be dismissed. An important requirement is a positive squared norm of the resulting normal vector $\mathbf{w}$. For some kernels their unsuitablility can be decided without SVM-training or even data-independently.

From a practical point of view, we conclude that it is “safe” to use non-cpd kernels for SVM. If they work, the result has a reasonable interpretation. Requirements are to use implementations with known convergence behaviour, e.g. libsvm [10], and to be aware that obtained solutions might be only local optima or stationary points.

Acknowledgements

The author wants to thank B. Schölkopf for the kind hospitality during the visit at the MPI Tübingen and U. von Luxburg and M. Hein for valuable comments.

A Proofs

Proof of Proposition 1 (Isometric Embedding). We use the construction from [12]. The symmetric and zero-diagonal function d^2 allows to define the matrix of squared distances by $\mathbf{D}^{(2)} := (d^2(x_i, x_j))_{i,j=1}^n$. With the centering matrix $\mathbf{J} := \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ we can construct the *centered* kernel matrix $\mathbf{K} := -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}$. This matrix is symmetric and singular, as $\mathbf{J}$ has eigenvalue 0. The eigendecomposition can be performed as $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ with $\mathbf{U}$ being orthogonal and $\boldsymbol{\Lambda}$ being diagonal starting with the p positive eigenvalues, followed by the q negative ones and $n - p - q \geq 1$ eigenvalues equal 0. Therefore $p + q < n$ and

$$\mathbf{K} = \mathbf{U}|\boldsymbol{\Lambda}|^{1/2}\text{diag}(\mathbf{1}_p, -\mathbf{1}_q, \mathbf{0}_{n-p-q})|\boldsymbol{\Lambda}|^{1/2}\mathbf{U}^T.$$

This can be interpreted as an inner product matrix in the pseudo-Euclidean space $\mathbb{R}^{(p,q)}$: Using the corresponding matrix $\mathbf{M} := \text{diag}(\mathbf{1}_p, -\mathbf{1}_q)$ and defining the mapping $\Phi : \{x_i\}_{i=1}^n \to \mathbb{R}^{(p,q)}$ by $\Phi(x_i)$ being the vector consisting of the first $p{+}q$ components of the i-th column of $|\boldsymbol{\Lambda}|^{1/2}\mathbf{U}^T$ we obtain the matrix element $K_{ij} = \Phi(x_i)^T\mathbf{M}\Phi(x_j) = \langle\Phi(x_i), \Phi(x_j)\rangle$.

As $\|\Phi(x_i) - \Phi(x_j)\|^2 = K_{ii} - 2K_{ij} + K_{jj}$ it suffices for the isometry to show that the right hand is equal to $d^2(x_i, x_j)$. This follows by simple calculations from the fact that the elements of $\mathbf{K}$ are defined as

$$K_{kl} = -\frac{1}{2}\left(d^2(x_k, x_l) - \frac{1}{n}\sum_i d^2(x_k, x_i) - \frac{1}{n}\sum_i d^2(x_i, x_l) + \frac{1}{n^2}\sum_{i,j} d^2(x_i, x_j)\right).$$

□

Proof of Proposition 2 (CH Primal in $\mathbb{R}^{(p,q)}$). We start with an equivalence that will be used: For all i, j holds

$$\mathbf{w}^T\mathbf{M}(\Phi(x_i) - \Phi(x_j)) = \bar{\boldsymbol{\alpha}}^T\mathbf{Q}(y_i\mathbf{e}_i - y_j\mathbf{e}_j). \tag{13}$$

This can be seen by expressing both sides in terms of kernel evaluations. For the right hand we obtain by definition of $\mathbf{Q}$

$$\bar{\boldsymbol{\alpha}}^T\mathbf{Q}(y_i\mathbf{e}_i - y_j\mathbf{e}_j) = \sum_m \bar{\alpha}_m y_m (k(x_i, x_m) - k(x_j, x_m)).$$

For the left hand we obtain with $h(x) := \Phi(x)^T\mathbf{M}\Phi(x) - k(x, x)$ and inserting (2) and (3)

$$\mathbf{w}^T\mathbf{M}(\Phi(x_i) - \Phi(x_j)) = \sum_m \bar{\alpha}_m y_m \left(k(x_i, x_m) - k(x_j, x_m) + \frac{1}{2}(h(x_i) - h(x_j))\right).$$

The last terms cancel out due to $\bar{\boldsymbol{\alpha}}^T\mathbf{y} = 0$, and we get the claimed identity (13).

We start with the proof of $(\mathbf{w}^T, b, \rho, \boldsymbol{\xi}^T)^T$ being a feasible point. Obviously $\mathbf{w}, b$ are not explicitly constrained and $\rho > 0$ by assumption.

We now argue why $\xi_i \geq 0$. If $\bar{\alpha}_i < \max_j \bar{\alpha}_j$ then $\xi_i = 0$ by definition. If $\alpha_i = \max_j \bar{\alpha}_j$ we have to show that ξ_i is nonnegative.

We assume $y_i = 1$, the other case follows similarly. By using the definitions of ρ and b and applying (13) we get

$$\begin{aligned} \xi_i &= \rho - \frac{1}{2}\mathbf{w}^T\mathbf{M}(\Phi(x_i) - \Phi(x_k)) - \frac{1}{2}\mathbf{w}^T\mathbf{M}(\Phi(x_i) - \Phi(x_l)) \\ &= \rho - \mathbf{w}^T\mathbf{M}(\Phi(x_i) - \Phi(x_k)) - \frac{1}{2}\mathbf{w}^T\mathbf{M}(\Phi(x_k) - \Phi(x_l)) \\ &= \mathbf{w}^T\mathbf{M}(\Phi(x_k) - \Phi(x_i)) = \bar{\boldsymbol{\alpha}}^T\mathbf{Q}(y_k\mathbf{e}_k - y_i\mathbf{e}_i) = \left\langle \bar{\boldsymbol{\alpha}}^T\mathbf{Q}, \mathbf{e}_k - \mathbf{e}_i \right\rangle \geq 0. \end{aligned}$$

The last step follows from $\mathbf{v} := \mathbf{e}_k - \mathbf{e}_i$ being a feasible direction of (CH-DU) and the derivative of (CH-DU) $\left\langle -\bar{\boldsymbol{\alpha}}^T\mathbf{Q}, \mathbf{v} \right\rangle \leq 0$ due to stationarity of $\bar{\boldsymbol{\alpha}}$. Feasibility of $\mathbf{v}$ can be seen as feasible directions of (CH-DU) are characterized by

$$\mathbf{y}^T\mathbf{v} = 0, \quad \mathbf{1}^T\mathbf{v} = 0, \quad v_i \geq 0 \quad \text{if} \quad \bar{\alpha}_i = 0 \quad \text{and} \quad v_i \geq 0 \quad \text{if} \quad \bar{\alpha}_i = \mu. \tag{14}$$

We now show that the last constraint is satisfied, i.e. $y_i(\mathbf{w}^T\mathbf{M}\Phi(x_i) + b) \geq \rho - \xi_i$ holds for all i. In case of $\bar{\alpha}_i = \max_j \bar{\alpha}_j$ this is valid by equality due to definition of the ξ_i. In case of $\bar{\alpha} < \max_j \bar{\alpha}_j$ we know $\xi_i = 0$ and it suffices to show that

$$y_i(\mathbf{w}^T\mathbf{M}\Phi(x_i) + b) - \rho \geq 0. \tag{15}$$

Similar as above, by using the definitions of ρ and b, equality (13) and the stationarity of $\bar{\boldsymbol{\alpha}}$ the statement follows. More precisely, (15) holds with equality if $0 < \bar{\alpha}_i < \max_j \bar{\alpha}_j$ and with ”$\geq$” in case of $\bar{\alpha}_i = 0$. So $(\mathbf{w}^T, b, \rho, \boldsymbol{\xi}^T)^T$ is a feasible point.

We continue with arguing that it is a stationary point. It is sufficient to show that for every feasible direction $\mathbf{v} := (\Delta\mathbf{w}^T, \Delta b, \Delta\rho, \Delta\boldsymbol{\xi}^T)^T$ of the primal optimization problem $J(\mathbf{w}, \Delta b, \Delta\rho, \boldsymbol{\xi})$ satisfies $\langle \nabla J, \mathbf{v} \rangle \geq 0$. By computing and inserting the partial derivatives, this is equivalent to showing

$$\mathbf{w}^T\mathbf{M}\Delta\mathbf{w} + \mu\mathbf{1}^T\Delta\boldsymbol{\xi} - 2\Delta\rho \geq 0. \tag{16}$$

If $\alpha_i > 0$ then the slack constraint is satisfied with equality, so a feasible direction $\mathbf{v}$ satisfies

$$y_i(\Delta\mathbf{w}^T\mathbf{M}\Phi(x_i)) \geq \Delta\rho - \Delta\xi_i - y_i\Delta b.$$

Multiplying with $\bar{\alpha}_i$, summing over all i, using $\sum_i \bar{\alpha}_i = 2$ and the definition of $\mathbf{w}$ we obtain

$$\mathbf{w}^T\mathbf{M}\Delta\mathbf{w} + \bar{\boldsymbol{\alpha}}^T\Delta\boldsymbol{\xi} - 2\Delta\rho \geq 0.$$

So (16) is particularly satisfied as $\mu \geq \bar{\alpha}_i$ for all i. We conclude that $(\mathbf{w}, \Delta b, \Delta\rho, \boldsymbol{\xi})$ is a stationary point of the primal optimization problem. □

Proof of Proposition 4 (Equivalence of CH and SVM). (SVM-DU) is equivalent to

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathbf{1}_n^T\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{0}_n \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \quad \text{and} \quad \mathbf{y}^T\boldsymbol{\alpha} = 0. \end{aligned} \tag{17}$$

Similarly (CH-DU) is equivalent to

$$\begin{aligned} \max_{\bar{\boldsymbol{\alpha}}} \quad & -\tfrac{1}{2}\bar{\boldsymbol{\alpha}}^T\mathbf{Q}\bar{\boldsymbol{\alpha}} \\ \text{s.t.} \quad & \mathbf{0}_n \leq \bar{\boldsymbol{\alpha}} \leq \mu\mathbf{1}_n, \quad \mathbf{y}^T\bar{\boldsymbol{\alpha}} = 0 \quad \text{and} \quad \mathbf{1}_n^T\bar{\boldsymbol{\alpha}} = 2. \end{aligned} \tag{18}$$

This can be seen by replacing d^2 by (2), decomposing the sum and using $\sum_i \bar{\alpha}_i y_i = 0$. So it is sufficient to show the correspondences of local solutions of (17) and (18).

A feasible direction $\mathbf{v}$ of (SVM-DU) in the point $\boldsymbol{\alpha}$ is characterized by

$$\mathbf{y}^T\mathbf{v} = 0, \quad v_i \geq 0 \quad \text{if} \quad \alpha_i = 0 \quad \text{and} \quad v_i \geq 0 \quad \text{if} \quad \alpha_i = C. \tag{19}$$

Similar but with an additional constraint we already characterized the feasible directions of (CH-DU) in (14).

i) A nonzero local optimum $\boldsymbol{\alpha}$ of (17) clearly induces a feasible point $\bar{\boldsymbol{\alpha}}$ of (18), as the scaling exactly results in $\mathbf{1}^T\boldsymbol{\alpha} = 2$ and corresponding bounding constraints. It remains to check that $\bar{\boldsymbol{\alpha}}$ is a stationary point and even a local optimum. Let $\mathbf{v}$ be a feasible direction of (CH-DU). As $\mathbf{v}$ simultaneously is a feasible direction of (SVM-DU) and $\boldsymbol{\alpha}$ a stationary point we know $\langle \mathbf{1}_n - \mathbf{Q}\boldsymbol{\alpha}, \mathbf{v}\rangle \leq 0$. As $\mathbf{1}_n^T\mathbf{v} = 0$ we get nonpositive derivatives of (18): $\langle -\mathbf{Q}\bar{\boldsymbol{\alpha}}, \mathbf{v}\rangle \leq 0$. So $\bar{\boldsymbol{\alpha}}$ is a stationary point of (CH-DU). It similarly is even a local optimum: Let $\mathbf{v}$ be a feasible direction of (CH-DU) with $\langle -\mathbf{Q}\bar{\boldsymbol{\alpha}}, \mathbf{v}\rangle = 0$ Then we similarly get $\langle \mathbf{1}_n - \mathbf{Q}\boldsymbol{\alpha}, \mathbf{v}\rangle = 0$. As $\mathbf{v}$ is a feasible direction of (SVM-DU), the second derivative is nonpositive $-\mathbf{v}^T\mathbf{Q}\mathbf{v} \leq 0$. The second derivatives of (SVM-DU) and (CH-DU) coincide, so we have shown that the curvature of (CH-DU) in direction $\mathbf{v}$ is nonpositive, $\bar{\boldsymbol{\alpha}}$ is a local optimum.

ii) Clearly a local solution $\bar{\boldsymbol{\alpha}}$ of (18) induces a feasible point $\boldsymbol{\alpha}$ of (17) for any choice of $\rho > 0$. It remains to show that it is a stationary point and a local optimum.

For being a stationary point we show that for every feasible direction $\mathbf{v}$ satisfying (19) holds

$$\langle \mathbf{1}_n - \mathbf{Q}\boldsymbol{\alpha}, \mathbf{v}\rangle \leq 0. \tag{20}$$

We have two unbounded coefficients α_k, α_l. Then we decompose $\mathbf{v}$:

$$\begin{aligned} \mathbf{v} &= \sum_i v_i\mathbf{e}_i \\ &= \sum_{i:y_i=1} y_i v_i(y_i\mathbf{e}_i - y_k\mathbf{e}_k) + \sum_{i:y_i=-1} y_i v_i(y_i\mathbf{e}_i - y_l\mathbf{e}_l) + \sum_{i:y_i=1} v_i(y_k\mathbf{e}_k - y_l\mathbf{e}_l) \\ &=: T_1 + T_2 + T_3. \end{aligned}$$

So it is sufficient to show (20) for contributions $\mathbf{v}'$ contained in T_1, T_2 or T_3. We start with T_1. For $\mathbf{v}' = y_i v_i(y_i\mathbf{e}_i - y_k\mathbf{e}_k)$ we obtain

$$\langle \mathbf{1}_n - \mathbf{Q}\boldsymbol{\alpha}, \mathbf{v}'\rangle = y_i v_i \langle \mathbf{1}_n, \mathbf{e}_i - \mathbf{e}_k\rangle - y_i v_i \frac{1}{\rho} \langle \mathbf{Q}\bar{\boldsymbol{\alpha}}, y_i\mathbf{e}_i - y_k\mathbf{e}_k\rangle . \tag{21}$$

The first term vanishes, and it remains to argue why the last term is nonpositive. In case of $0 < \alpha_i < C$, we know that $\bar{\alpha}_i$ due to scaling also is not bounded. So $\pm(y_i\mathbf{e}_i - y_k\mathbf{e}_k)$ are feasible directions of (CH-DU), and the corresponding derivatives $\pm\langle \mathbf{Q}\bar{\boldsymbol{\alpha}}, y_i\mathbf{e}_i - y_k\mathbf{e}_k\rangle \leq 0$, therefore the last term in (21) is 0. In case of $\alpha_i = 0$, the corresponding $v_i \geq 0$ as $\mathbf{v}$ is a feasible direction of (SVM-DU). So $\mathbf{e}_i - y_iy_k\mathbf{e}_k$ is a feasible direction of (CH-DU), which implies that the last inner product in (21) has the same sign as y_i. We conclude similarly in case of $\alpha_i = C$ that $v_i \leq 0$ and $-(\mathbf{e}_i - y_iy_k\mathbf{e}_k)$ is feasible for (CH-DU) which again yields the desired nonpositivity of the last term of (21). For directions $\mathbf{v}'$ in T_2 the argumentation is analogous, so we continue with T_3. The vector $y_k\mathbf{e}_k - y_l\mathbf{e}_l$ is not a feasible direction of (CH-DU), so the argumentation is different. Similarly as before we have to show the nonpositivity of

$$\langle \mathbf{1}_n - \mathbf{Q}\boldsymbol{\alpha}, v_i(\mathbf{e}_k + \mathbf{e}_l)\rangle = v_i\left(\langle \mathbf{1}_n, \mathbf{e}_k + \mathbf{e}_l\rangle - \frac{1}{\rho}\langle \mathbf{Q}\bar{\boldsymbol{\alpha}}, \mathbf{e}_k + \mathbf{e}_l\rangle\right).$$

This is particularly satisfied, if the term in brackets vanishes, i.e. if $\rho := \frac{1}{2}\langle \mathbf{Q}\bar{\boldsymbol{\alpha}},\ \mathbf{e}_k + \mathbf{e}_l\rangle$ and if $\rho > 0$. So we conclude that $\boldsymbol{\alpha}$ is a stationary point of (SVM-DU).

The positive definiteness of $\mathbf{Q}$ on the space of all feasible directions $\mathbf{v}$ of (SVM-DU) with $\langle \mathbf{1}_n - \mathbf{Q}\boldsymbol{\alpha}, \mathbf{v}\rangle = 0$ further implies that $\boldsymbol{\alpha}$ is a local optimum.

iii) By replacing d^2 in (8) by (2), decomposing the sum and using $\sum_i \bar{\alpha}_i y_i = 0$, we obtain exactly a (positively) scaled and shifted version of (1). So the resulting hyperplanes are parallel.

In case of no bounded coefficients $\bar{\boldsymbol{\alpha}}$ (resp. $\boldsymbol{\alpha}$) one can show that for two (trivially existing) unbounded $\bar{\alpha}_k, \bar{\alpha}_l$ of different classes $y_k = +1, y_l = -1$ holds $f(x_k) = -f(x_l)$ for f chosen as the SVM (1) or CH-classification (8) function. This implies the identity of the classification boundaries and regions.

We demonstrate it for the CH-classification function, for the SVM-function it is analogous.

For CH-classification we already know that $\mathbf{z}^+$ and $\mathbf{z}^-$ have identical value of g in (6). So it remains to show that $f(x_k) = g(\mathbf{z}^+)$ and $f(x_l) = g(\mathbf{z}^-)$. We only show the first equality, the second follows analogously.

Using (7) and (8) we get after replacing d^2 by (2)

$$g(\mathbf{z}^+) - f(x_k) = \sum_i \bar{\alpha}_i \Phi(x_i)^T \mathbf{M}\left(\sum_{j:y_j=+1} \bar{\alpha}_j \Phi(x_j) - \Phi(x_k)\right).$$

Using the definition of $\mathbf{w}$ from Proposition 2 and $\sum_{j:y_j=+1} \bar{\alpha}_j = 1$ this is

$$= \mathbf{w}^T\mathbf{M}\left(\sum_{j:y_j=+1} \bar{\alpha}_j(\Phi(x_j) - \Phi(x_k))\right).$$

For arguing that this is zero it is sufficient to show that all single terms in the sum vanish. By (13) we get

$$\mathbf{w}^T\mathbf{M}(\Phi(x_j) - \Phi(x_k)) = \bar{\boldsymbol{\alpha}}\mathbf{Q}(y_j\mathbf{e}_j - y_k\mathbf{e}_k).$$

This is exactly the derivative of (CH-DU) in direction $\mathbf{v} := -\mathbf{e}_j + \mathbf{e}_k$. As $\pm\mathbf{v}$ are feasible directions this must be 0 by stationarity of $\bar{\boldsymbol{\alpha}}$. □

Proof of Lemma 6 (Uniqueness of Stationary Points). Let $(\bar{\mathbf{z}}^-, \bar{\mathbf{z}}^+)$ be another stationary point of (4). We denote the optimization function as J. Then $(\Delta\mathbf{z}^-, \Delta\mathbf{z}^+) := (\mathbf{z}^- - \bar{\mathbf{z}}^-, \mathbf{z}^+ - \bar{\mathbf{z}}^+)$ is a feasible direction in $(\bar{\mathbf{z}}^-, \bar{\mathbf{z}}^+)$. Noting that $\nabla_{\mathbf{z}^-} J = -\nabla_{\mathbf{z}^+} J = 2\mathbf{M}(\mathbf{z}^- - \mathbf{z}^+)$ it suffices to obtain a contradiction to the stationarity of $(\bar{\mathbf{z}}^-, \bar{\mathbf{z}}^+)$ by

$$\left\langle \nabla J, (\Delta\mathbf{z}^-, \Delta\mathbf{z}^+) \right\rangle < 0. \tag{22}$$

Inserting the definitions and using $\bar{\mathbf{z}}^+ - \bar{\mathbf{z}}^- = \bar{\mathbf{z}}^+ - \mathbf{z}^+ + \mathbf{z}^+ - \mathbf{z}^- + \mathbf{z}^- - \bar{\mathbf{z}}^-$ we have to show the negativity of

$$\begin{aligned}
&(\bar{\mathbf{z}}^+ - \bar{\mathbf{z}}^-)^T \mathbf{M}(\mathbf{z}^+ - \bar{\mathbf{z}}^+ - \mathbf{z}^- + \bar{\mathbf{z}}^-) \\
&\quad = \quad (\bar{\mathbf{z}}^+ - \mathbf{z}^+)^T \mathbf{M}(-\bar{\mathbf{z}}^+ + \mathbf{z}^+) + (\bar{\mathbf{z}}^+ - \mathbf{z}^+)^T \mathbf{M}(-\mathbf{z}^- + \bar{\mathbf{z}}^-) \\
&\qquad + (\mathbf{z}^+ - \mathbf{z}^-)^T \mathbf{M}(\mathbf{z}^+ - \bar{\mathbf{z}}^+ - \mathbf{z}^- + \bar{\mathbf{z}}^-) \\
&\qquad + (\mathbf{z}^- - \bar{\mathbf{z}}^-)^T \mathbf{M}(-\mathbf{z}^- + \bar{\mathbf{z}}^-) + (\mathbf{z}^- - \bar{\mathbf{z}}^-)^T \mathbf{M}(-\bar{\mathbf{z}}^+ + \mathbf{z}^+) \\
&\quad =: T_1 + T_2 + T_3 + T_4 + T_5.
\end{aligned}$$

Due to the assumption of positive distance of $\mathbf{z}^+, \mathbf{z}^-$ to the remaining points of their corresponding convex hulls, we have $T_1 < 0, T_4 < 0$. Stationarity of J in $(\mathbf{z}^-, \mathbf{z}^+)$ implies $T_3 \leq 0$. So it remains to show $T_2 = T_5 \leq 0$, i.e.

$$(\bar{\mathbf{z}}^+ - \mathbf{z}^+)^T \mathbf{M}(\bar{\mathbf{z}}^- - \mathbf{z}^-) \leq 0. \tag{23}$$

This follows with the assumption of working in the space $\mathbb{R}^{(1,q)}$. In these spaces the cone of points with positive squared norm and positive first coordinate is closed under additions, positive scalings and reflection with $\mathbf{M}$. Additionally inner products between such points are always positive. So it remains to show that either both $\bar{\mathbf{z}}^+ - \mathbf{z}^+$ and $\mathbf{z}^- - \bar{\mathbf{z}}^-$ or their negatives lie within this cone. As their squared norm is positive by assumption, it remains to show that they have the same sign in their first component. This can be obtained as both signs must be equal to the sign of the first component of $\mathbf{z}^+ - \mathbf{z}^-$. If e.g. $\bar{\mathbf{z}}^+ - \mathbf{z}^+$ would have different sign in its first component, then the inner product with $\mathbf{z}^+ - \mathbf{z}^-$ would be negative, which would be a contradiction to the stationarity of $(\mathbf{z}^-, \mathbf{z}^+)$. □

References

[1] C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line Handwriting Recognition with Support Vector Machines—A Kernel Approach. In *Proc. of the 8th IWFHR*, pages 49–54, 2002.

[2] K. P. Bennett and E. J. Bredensteiner. Duality and geometry in SVM classifiers. In *Proc. 17th International Conf. on Machine Learning*, pages 57–64. Morgan Kaufmann, San Francisco, CA, 2000.

[3] D. J. Crisp and C.J.C. Burges. A geometric interpretation of nu-svm classifiers. In *Advances in Neural Information Processing Systems (NIPS) 12*. MIT Press, Cambridge, MA, 1999.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, 2000.

[5] D. DeCoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46(1):161–190, 2002.

[6] L. Goldfarb. A new approach to pattern recognition. In L.N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition 2*, pages 241–402. Elsevier Science Publishers B.V., 1985.

[7] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems (NIPS) 11*, pages 438–444. MIT Press, Cambridge, MA, 1999.

[8] B. Haasdonk and D. Keysers. Tangent distance kernels for support vector machines. In *Proc. of the 16th Int. Conf. on Pattern Recognition*, volume 2, pages 864–868, Quebec, Canada, September 2002.

[9] M. Hein and O. Bousquet. Maximal margin classification for metric spaces. In B. Schölkopf and M. Warmuth, editors, *Proceedings of the 16th Annual Conference on Computational Learning Theory*, pages 72–86. Springer, Berlin, 2003.

[10] H.-T. Lin and C.-J. Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. *submitted to Neural Computation*, 2003.

[11] X. Mary. *Hilbertian subspaces, subdualities and applications*. PhD thesis, INSA Rouen, 2003.

[12] E. Pekalska, P. Paclik, and R. Duin. A generalized kernel approach to dissimilarity based classification. *J. of Mach. Learn. Research*, (2):175–211, 2001.

[13] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. *Neural Computation*, (12):1083 – 1121, 2000.

[14] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

[15] H. Shimodaira, K. Noma, M. Nakai, and S. Sagayama. Dynamic time-alignment kernel in support vector machine. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS) 14*, pages 921–928. MIT Press, Cambridge, MA, 2002.