EPiC Series in Computing

Volume 104, 2024, Pages 126–150

Proceedings of 3rd International Workshop on
Mathematical Modeling and Scientific Computing

A Mathematical Introduction to SVMs with Self-Concordant Kernel

Florian Jarre

Faculty of Natural Sciences,
Heinrich Heine Universität Düsseldorf, Germany
jarre@hhu.de

Abstract

A derivation of so-called "soft-margin support vector machines with kernel" is presented along with elementary proofs that do not rely on concepts from functional analysis such as Mercer's theorem or reproducing kernel Hilbert spaces which are frequently cited in this context. The analysis leads to new continuity properties of the kernel functions, in particular a self-concordance condition for the kernel. Practical aspects concerning the implementation and the choice of the kernel are addressed and illustrated with some numerical examples. The derivations are intended for a general audience, requiring basic knowledge of calculus and linear algebra, while some more advanced results from optimization theory are being introduced in a self-contained form.

1 Introduction

For the problem of estimating yes-no answers – based on a given data set with known answers but with unknown structure – so-called "support vector machines" (SVMs) are an approach that can be applied under weak assumptions.

More precisely, SVMs are methods for the automated classification of new data into two classes based on a set of old data with corresponding classifications. For example, the data can be digitized images of handwritten characters, and the classification involves deciding whether the pixels of the images represent a given character or not. The old data are called training data whose classification has been made in some way beforehand, for example by a human who recognizes the images and manually enters the corresponding characters. If many different images of handwritten characters have been scanned and classified, the SVM is used to recognize new images of handwritten characters automatically without a programmer first having to enter specifications of the sort "the digit three has the following characteristics".

SVMs estimate yes-no decisions; for certain more complex answers, multiple different SVMs can be combined; however, in general, other approaches are more appropriate in these cases. Despite the limited form of the response, SVMs can indeed be a helpful approach for complicated decision problems.

V.L. Turova, A.E. Kovtanyuk and J. Zimmer (eds.), MMSC 2024 (EPiC Series in Computing, vol. 104),
pp. 126–150

1.1 Limitations of SVMs

Applications of SVMs include, for example, automatically classifying images or the task of determining from a database of patient data whether a new patient falls into a risk group for a specific disease – or, using another database, deciding whether a customer should be granted a loan. The last two examples illustrate a problem that has repeatedly occurred in applications of AI techniques and is described, for example, in [11]. When a SVM is used to identify risks in patients that might otherwise be overlooked, it serves human well-being; however, it does not do so when used to deny health insurance or a loan without further examination.

Another problem arises when the training data itself has been automatically classified. The classification errors made during this process typically persist in the SVM that is developed from this training data. The same is true, of course, when the training data has been incorrectly classified by humans.

Even if the given data has been classified correctly, the quality or quantity of the data often is not sufficient to derive a clear classification from it. Nevertheless, an SVM generally does provide some form of classification but without the information how reliable the output is.

Another problem arises when the training data is not uniformly distributed. If the handwritten digits from the above example were collected in the USA (where the digits 1 and 7 are written differently than in Germany), an increased error rate can be expected when an SVM developed in the USA is applied in Germany. (This is a rather harmless example!)

Finally, the limitations of SVMs also concern their applicability to large data sets. When the data consists of more than $m = 5000$ data points, the SVM approach presented in this paper gets increasingly expensive a since optimization problems with dense $m \times m$-matrices need to be solved. Modifications for solving large scale problems are proposed in [2, 4, 8, 7], for example.

The following will not further address such modeling and interpretation errors but explain the mechanism of SVMs. Complementing surveys are given, for example, in [1, 12, 19].

1.2 Outlook

The basic idea is to assign similar data to the same class. However, the term "similar data" is very imprecise. For example, the pixels of two scans of handwritten digits can be completely different even if the same digit is represented. One problem that SVMs ideally solve automatically is to derive and utilize an appropriate criterion for "similarity" from the data.

The basics of SVMs are well-researched and understood, see e.g. [13, 3, 10, 18] and the references therein. Below, an introductory mathematical summary will be provided. The chosen presentation is "minimalistic" in the sense that some concepts commonly used in the consideration of kernel functions are omitted, and also the results from statistical learning theory are not addressed. In particular, the completeness of the "feature" space and associated theorems such as Riesz's representation theorem or Mercer's theorem are not used. Nevertheless, a central property, the continuity of the kernel functions, will be established and analyzed. Finally, the effect of a normalization of the kernel on the continuity properties and on the conditioning of the kernel matrix is addressed.

1.3 Notation

$A \succeq 0$ denotes that A is a symmetric positive semidefinite matrix. Given two matrices $A, B \in \mathbb{R}^{m\times n}$, their Hadamard product $A \circ B$ is defined by component-wise multiplication,

$$(A \circ B)_{ij} = A_{ij}B_{ij} \quad \text{for } 1 \le i \le m,\ 1 \le j \le n.$$

For a matrix A, the matrix norm induced by the 2-norm is denoted by $\|A\|_2$. The derivative $Df(x)$ of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ is represented as an $m \times n$ matrix (the Jacobian), and in the case $m = 1$, the gradient $\nabla f(x) = Df(x)^T$ is a column vector. The vector $e := (1, 1, \ldots, 1)^T$ always denotes the all-one-vector with its dimension given by the context.

2 Basic form of support vector machines

The initial situation is as follows. There are given training data consisting of points $x^{(i)}$ from a compact convex set $\Omega \subset \mathbb{R}^n$ for $1 \le i \le m$ and associated classifications $\zeta_i \in \{-1, 1\}$.

The simplest case of defining an SVM is when there exists a hyperplane $\{x \in \mathbb{R}^n \mid a^T x = \beta\}$ with a fixed vector $a \in \mathbb{R}^n \backslash \{0\}$ and a constant $\beta \in \mathbb{R}$ such that

$$\begin{aligned} a^T x^{(i)} &> \beta \quad \forall\, i \text{ with } \zeta_i = 1 \text{ and} \\ a^T x^{(i)} &< \beta \quad \forall\, i \text{ with } \zeta_i = -1. \end{aligned} \tag{1}$$

In this case, we call the hyperplane $\{x \mid a^T x = \beta\}$ classifying, as it "correctly separates" all data points, and the criterion of "similarity" of data reduces to whether $a^T x > \beta$ holds or not.

To maximize the "prospect" that the chosen hyperplane correctly classifies new points, the hyperplane should be chosen so that it correctly separates the given points on the one hand and is as far away as possible from all points on the other hand — i.e., as few points as possible are boundary cases that would switch classes under a slight perturbation. Finding the best hyperplane in this sense should be automated.

To determine the hyperplane, conditions (1) are written equivalently in the compact form

$$\delta_i := \zeta_i (a^T x^{(i)} - \beta) > 0 \quad \forall\, i. \tag{2}$$

If a point $x^{(i)}$ satisfies condition (2) and a number $\lambda \ge 1/\delta_i$ is chosen, then $x^{(i)}$ also satisfies the condition

$$\zeta_i (\lambda a^T x^{(i)} - \lambda \beta) \ge 1.$$

By changing from a to λa and β to $\lambda \beta$, the strict inequalities in (2) can therefore be replaced by the weak inequalities

$$\zeta_i (a^T x^{(i)} - \beta) \ge 1 \quad \forall\, i. \tag{3}$$

with the right side 1. Finally, note that the distance of a point $\bar{x}$ from the hyperplane $\{x \mid a^T x = \beta\}$ is given by $|a^T \bar{x} - \beta| / \|a\|_2$. Maximizing the distance of $\bar{x}$ from the hyperplane under the condition that $|a^T x - \beta| \ge 1$ therefore is equivalent to minimizing the norm of a. Thus, the problem of maximizing the minimum distance of all points from a classifying hyperplane can be written in the form

$$\min_{a,\ \beta} \{\ \frac{1}{2}\|a\|_2^2 \ \mid\ \zeta_i (a^T x^{(i)} - \beta) \ge 1 \ \ \forall\, 1 \le i \le m\ \}. \tag{4}$$

Here, the standard notation is used, where the term "min" in Problem (4) is to be understood as "minimize"; if the data is such that there is no separating hyperplane, the minimum does not exist, but otherwise it is uniquely defined. If a, β are optimally determined from (4), then the label $\tilde{\zeta}$ for a new point $\tilde{x}$ can be estimated by

$$\tilde{\zeta} := \operatorname{sign}(a^T \tilde{x} - \beta)$$

since $(\tilde{x}, \tilde{\zeta})$ then also satisfies the relationship $\tilde{\zeta}(a^T\tilde{x} - \beta) \geq 0$.

In solving (4), many of the constraints $\zeta_i(a^T x^{(i)} - \beta) \geq 1$ typically prove to be superfluous. Only the points with the smallest values $|a^T x^{(i)} - \beta|$ are relevant for determining the optimal hyperplane. Those training points that are not redundant, i.e., that have the minimum distance to the hyperplane, are called "support vectors," which explains the name SVM.

3 Soft margin SVM

Often, the situation arises where the given data cannot be exactly separated by a hyperplane because, for example, not all training data points were correctly classified. In this case, one can use a so-called "soft margin" SVM, where the restrictions (3) are relaxed to $\zeta_i(a^T x^{(i)} - \beta) \geq 1 - s_i$ with $s_i \geq 0$, and minimize[1] the expression

$$\frac{1}{2}\|a\|_2^2 + C\sum_{i=1}^{m} s_i \tag{5}$$

for a fixed "penalty parameter" $C > 0$. For a large value C, higher priority is given to minimizing the "tolerated error terms" $s_i \geq 0$ (hoping that only the incompatible, misclassified data points $x^{(i)}$ retain positive values $s_i > 0$) and lower priority is given to minimizing the norm of a, whose inverse describes the distance of the support vectors from the separating hyperplane. The sum in (5) is given by $\sum_{i=1}^{m} s_i = e^T s$, and the overall soft margin SVM problem is given by: Find μ^* and a, β, s with

$$\mu^* = \min_{a,\ \beta,\ s} \{\ \frac{1}{2}\|a\|_2^2 + Ce^T s\ \mid\ \zeta_i(a^T x^{(i)} - \beta) \geq 1 - s_i,\quad \forall\ 1 \leq i \leq m,\quad s \geq 0\}. \tag{6}$$

The solution to (6) can be reformulated as follows. Let L be the so-called Lagrangian for (6), i.e.,

$$L((a,\beta,s),(u,v)) := \frac{1}{2}\|a\|_2^2 + Ce^T s + \sum_{i=1}^{m} u_i \underbrace{(1 - s_i + \zeta_i\beta - \zeta_i a^T x^{(i)})}_{\leq 0 \text{ in } (6)} + v^T \underbrace{(-s)}_{\leq 0 \text{ in } (6)}$$

for so-called Lagrange multipliers $u, v \geq 0$. The motivation for defining the Lagrangian function is that problem (6) can be written as

$$\mu^* = \inf_{a,\ \beta,\ s} \left(\sup_{u \geq 0,\ v \geq 0} L((a,\beta,s),(u,v)) \right) \tag{7}$$

because when forming the infimum, only those (a, β, s) are selected for which the supremum is finite, and those are exactly the ones for which the constraints from (6) are satisfied.

(For example, if $s_i < 0$ for some i, then taking the limit $u_i \to \infty$ for this i would result in the inner supremum having the value $+\infty$. Therefore, only vectors $s \geq 0$ are considered when forming the infimum. Similarly, only (a, β, s) are considered for which $1 - s_i + \zeta_i\beta - \zeta_i a^T x^{(i)} \leq 0$. For such choice of (a, β, s) the optimal value of the inner supremum is obtained for $u = v = 0$ so that indeed the objective function of (6) is minimized in (7).)

Problem (6) is a convex problem, and since only linear constraints are present, the so-called Slater condition is trivially satisfied, and the "Lagrange duality" holds, i.e.

$$\inf_{a,\ \beta,\ s} \sup_{u \geq 0,\ v \geq 0} L((a,\beta,s),(u,v)) = \sup_{u \geq 0,\ v \geq 0} \inf_{a,\ \beta,\ s} L((a,\beta,s),(u,v)). \tag{8}$$

[1] Another formulation of the soft margin is discussed, for example, in [3].

(The fact that the left side in (8) is greater than or equal to the right follows from elementary calculations; and the fact that both sides are equal – when the Slater condition is satisfied – is a standard result of convex optimization, see, for example, [17].) The inner problem on the right (the formation of the infimum) now has no more constraints and due to convexity it can be solved explicitly (for given $u, v \geq 0$):

Rewriting the Lagrangian equivalently as

$$L((a,\beta,s),(u,v)) = \frac{1}{2}\|a\|_2^2 - \left(\sum_{i=1}^m u_i\zeta_i x^{(i)}\right)^T a + \left(\sum_{i=1}^m u_i\zeta_i\right)\beta + (Ce-u-v)^T s + e^T u$$

we consider the minimizer of the inner infimum problem on the right in (8). By setting the derivatives with respect to (a,β,s) to zero it follows that

$$a = \sum_{i=1}^m u_i\zeta_i x^{(i)}, \quad \sum_{i=1}^m u_i\zeta_i = 0, \quad \text{and} \quad Ce-u-v=0 \tag{9}$$

must be satisfied. With these conditions, the terms

$$(\sum_{i=1}^m u_i\zeta_i)\beta + (Ce-u-v)^T s$$

vanish in the Lagrangian, and the first equation in (9) states that the first two terms in the Lagrangian reduce to

$$\frac{1}{2}\|a\|_2^2 - (\sum_{i=1}^m u_i\zeta_i x^{(i)})^T a = -\frac{1}{2}\|\sum_{i=1}^m u_i\zeta_i x^{(i)}\|_2^2,$$

where the variable a has been eliminated.

The conditions (9) are equivalent to forming the inner infimum of the right side of (8) and can therefore be formulated as constraints on the supremum problem,

$$\mu^* = \sup_{u\geq 0,\ v\geq 0}\left\{-\frac{1}{2}\|\sum_{i=1}^m u_i\zeta_i x^{(i)}\|_2^2 + e^T u \ \mid\ u^T\zeta = 0,\ Ce-u-v=0\right\} \tag{10}$$

where $\zeta \in \mathbb{R}^m$ is the vector with components ζ_i. The "slack vector" $v \geq 0$ simply indicates that $Ce-u \geq 0$ holds. It can be eliminated above, and with a change of sign in the objective function, one obtains

$$-\mu^* = \inf_{u\in\mathbb{R}^m}\left\{\frac{1}{2}\|\sum_{i=1}^m u_i\zeta_i x^{(i)}\|_2^2 - e^T u \ \mid\ u^T\zeta = 0,\ Ce\geq u\geq 0\right\}. \tag{11}$$

(The constraint $u \geq 0$ from the term "$\sup_{u\geq 0}$" in (10) is explicitly listed on the right in (11) again.) Problem (11) is also called the *dual problem* to (6).

In general, similar as for problem (4), many data points $x^{(i)}$ are also unnecessary in the original soft-margin formulation (6), i.e., for many indices i, one obtains $\zeta^i(a^T x^{(i)} - \beta) > 1$ in any optimal solution a, β, s of (6). These $x^{(i)}$ and the corresponding s_i can then simply be omitted in the Lagrangian, or the corresponding multipliers u_i can be fixed to zero. And these $x^{(i)}$ are then also unnecessary in the equivalent transformation (11), i.e., the corresponding

multipliers u_i in an optimal solution of Problem (11) are zero. Let $\mathcal{B}$ denote the indices for which $u_i > 0$ holds in the given optimal solution. It follows from (9) that

$$a = \sum_{i \in \mathcal{B}} u_i \zeta_i x^{(i)}. \tag{12}$$

The set $\mathcal{B}$ defines the support vectors, $\{x^{(i)}\}_{i \in \mathcal{B}}$. Now, if a is calculated as above from the solution of (11), the corresponding β can be determined based on (6): For a given β and a, the corresponding optimal s in (6) is given by solving

$$\min \; Ce^T s \;\; | \;\; s_i \geq 1 + \zeta_i(\beta - a^T x(i)) \quad \forall \; 1 \leq i \leq m, \quad s \geq 0.$$

Setting $b \in \mathbb{R}^m$ the vector with components $b_i := 1 - \zeta_i a^T x^{(i)}$, the above inequalities on the variables s_i can be written as $s_i \geq 1 - \zeta_i a^T x^{(i)} + \beta \zeta_i = b_i + \beta \zeta_i$. Therefore, for a given β (and a) the optimal $s \in \mathbb{R}^m$ in (6) is explicitly representable as

$$s(\beta) = \max\{b + \beta\zeta, \; 0\},$$

where the maximum is applied component-wise. The mapping

$$\beta \mapsto e^T s(\beta) = e^T(\max\{b + \beta\zeta, \; 0\}) =: \sigma(\beta)$$

is piecewise linear, as a maximum of linear functions it is also convex, and the value of β that minimizes $e^T s(\beta)$ thus solves (6),

$$\beta = \operatorname{argmin}_{\hat{\beta} \in \mathbb{R}} \sigma(\hat{\beta}) = \operatorname{argmin}_{\hat{\beta} \in \mathbb{R}} \; \{e^T \max\{b + \hat{\beta}\zeta, \; 0\}\}. \tag{13}$$

Minimizing the convex, piecewise linear function $\sigma : \mathbb{R} \to \mathbb{R}$ is possible with low computational effort. The classification of a new data point $\tilde{x}$ can then be obtained as

$$\tilde{\zeta} := \operatorname{sign}(\tilde{x}^T a - \beta) = \operatorname{sign}(\tilde{x}^T (\sum_{i \in \mathcal{B}} u_i \zeta_i x^{(i)}) - \beta) = \operatorname{sign}((\sum_{i \in \mathcal{B}} u_i \zeta_i \tilde{x}^T x^{(i)}) - \beta).$$

For the derivation the kernel SVM, a reformulation of the objective function in (11) is presented next. Let $Z := \operatorname{Diag}(\zeta)$ be the diagonal matrix with diagonal $\zeta \in \mathbb{R}^m$. The quadratic term in the objective function of (11) can then be written as

$$\| \sum_{i=1}^{m} u_i \zeta_i x^{(i)} \|_2^2 = u^T (ZQZ) u \tag{14}$$

with the matrix Q having entries $Q_{i,j} = (x^{(i)})^T x^{(j)}$. As a Gram matrix, i.e.,

$$Q = (x^{(1)}, \ldots, x^{(m)})^T (x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}^{m \times m}, \tag{15}$$

the matrix Q is symmetric positive semidefinite, $Q \succeq 0$.

3.1 Summary, soft margin SVM

Given Q as in (15) and a diagonal matrix Z with diagonal entries ζ_i solve (11), i.e.

$$\min_{u \in \mathbb{R}^m} \left\{ \frac{1}{2} u^T ZQZu - e^T u \;\; | \;\; u^T \zeta = 0, \; Ce \geq u \geq 0 \right\},$$

set $b = e - ZQZu$, find β as in (13), and classify a new point $\tilde{x}$ via

$$\tilde{\zeta} := \text{sign}((\sum_{i \in \mathcal{B}} u_i \zeta_i x^{(i)})^T \tilde{x} - \beta)$$

where $\mathcal{B}$ is the set of components i with $u_i > 0$. The steps for the classification via kernel SVM in the next section are quite similar when replacing Q with some other matrix $K \succeq 0$.

4 Kernel SVM

Now, there are also many applications where the data, in the given form, cannot generally be separated by a hyperplane, i.e., the "border" that separates the two classes is not a hyperplane, but rather a somewhat more complicated nonlinear boundary. In these cases, a nonlinear mapping is sought

$$\phi : \Omega \to \mathcal{W},$$

which maps the points $x^{(i)} \in \Omega$ to a scalar product space $\mathcal{W}$ that is typically higher-dimensional (it can also be a function space) with scalar product $\langle ., . \rangle_{\mathcal{W}}$ and induced norm $\| \, . \, \|_{\mathcal{W}}$, so that the images $\phi(x^{(i)})$ of the two classes can be separated by a hyperplane. On the one hand, the choice of ϕ should preserve the "similarity" of data mentioned earlier i.e., ϕ should satisfy certain continuity properties, and on the other hand, the unknown "separation" into two classes should be made possible. By maintaining the "soft margin", one then obtains the problem to find

$$\mu^* = \min_{\tilde{a} \in \mathcal{W},\ \beta,\ s \geq 0} \{ \ \frac{1}{2} \|\tilde{a}\|_{\mathcal{W}}^2 + Ce^T s \ \mid \ \zeta_i \left(\langle \tilde{a}, \phi(x^{(i)}) \rangle_{\mathcal{W}} - \beta \right) \geq 1 - s_i, \quad \forall \ 1 \leq i \leq m \} \tag{16}$$

instead of (6), where the vectors $\tilde{a}$ now lie in the higher-dimensional image space $\mathcal{W}$ of ϕ. (Below it is established that the minimum actually exists.) Due to the nonlinearity of the mapping ϕ, the linear separation in the image space of ϕ translates into a nonlinear separation in the original data space with the data $x^{(i)}$.

When the optimal solution $\tilde{a}, \beta$ from (16) is given, the label $\tilde{\zeta}$ for a new data point $\tilde{x}$ is estimated by

$$\tilde{\zeta} := \text{sign} \left(\langle \, \tilde{a}, \phi(\tilde{x}) \rangle_{\mathcal{W}} - \beta \right). \tag{17}$$

Even if the dimension of $\mathcal{W}$ should be infinite, problem (16) still is a finite-dimensional optimization problem, as will be explained briefly:

To this end, let $M := \text{Span}\{\phi(x^{(i)})\}_{1 \leq i \leq m} \subset \mathcal{W}$ (the linear hull of the $\phi(x^{(i)})$ for $1 \leq i \leq m$). Then $\langle ., . \rangle_{\mathcal{W}}$ is also a scalar product on M and thus, for any fixed $\tilde{a} \in \mathcal{W}$, the function $\psi : M \to \mathbb{R}$ with $\psi(x) := \|\tilde{a} - x\|_{\mathcal{W}}^2$ is strictly convex. The function ψ is convex on the finite-dimensional space M and thus it is continuous on M. Furthermore, for $\bar{x} \in M$ the level set $\{x \in M \mid \psi(x) \leq \psi(\bar{x})\}$ is bounded. Thus, ψ has a unique minimizer on M, which is denoted by $\tilde{a}_M$. (This statement does not require the completeness of $\mathcal{W}$.) For $\lambda \in \mathbb{R}$ and fixed $i \in \{1, \ldots, m\}$ it follows from the definition of M and $\tilde{a}_M$ that

$$\|\tilde{a} - \tilde{a}_M\|_{\mathcal{W}}^2 \leq \|\tilde{a} - \tilde{a}_M + \lambda \phi(x^{(i)})\|_{\mathcal{W}}^2 = \|\tilde{a} - \tilde{a}_M\|_{\mathcal{W}}^2 + 2\lambda \langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} + \lambda^2 \|\phi(x^{(i)})\|_{\mathcal{W}}^2,$$

i.e., $0 \leq 2\lambda \langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} + \lambda^2 \|\phi(x^{(i)})\|_{\mathcal{W}}^2$ for $\lambda \in \mathbb{R}$. Setting

$$\lambda := \begin{cases} -\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} & \text{if} \quad \|\phi(x^{(i)})\|_{\mathcal{W}}^2 = 0, \\ -\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} / \|\phi(x^{(i)})\|_{\mathcal{W}}^2 & \text{otherwise,} \end{cases}$$

the assumption that $\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}} \neq 0$, leads to a contradiction. So, $\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}} = 0$.

Setting $\tilde{a}_{M^\perp} := \tilde{a} - \tilde{a}_M$ it follows

$$\langle \tilde{a}_{M^\perp}, \phi(x^{(i)})\rangle_{\mathcal{W}} = 0 \quad (1 \leq i \leq m), \qquad \text{and therefore also} \qquad \langle \tilde{a}_{M^\perp}, \tilde{a}_M\rangle_{\mathcal{W}} = 0. \tag{18}$$

Now let $\tilde{a}, \beta, s$ be feasible for (16). From (18) it then follows that

$$\|\tilde{a}\|^2_{\mathcal{W}} = \|\tilde{a}_{M^\perp} + \tilde{a}_M\|^2_{\mathcal{W}} = \|\tilde{a}_M\|^2_{\mathcal{W}} + \|\tilde{a}_{M^\perp}\|^2_{\mathcal{W}}$$

and $\langle \tilde{a}, \phi(x^{(i)})\rangle_{\mathcal{W}} = \langle \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}}$ for all i. Hence, $\tilde{a}_M, \beta, s$ is also feasible for (16) and the objective function value in (16) is reduced or remains unchanged when replacing $\tilde{a}$ with $\tilde{a}_M$. Therefore (16) is equivalent to

$$\min_{\tilde{a}_M \in M,\ \beta,\ s \geq 0} \{\ \frac{1}{2}\|\tilde{a}_M\|^2_{\mathcal{W}} + Ce^T s \ \mid\ \zeta_i \left(\langle \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}} - \beta \right) \geq 1 - s_i, \quad \forall\ 1 \leq i \leq m\}. \tag{19}$$

This problem is finite-dimensional and has the same structure as problem (6).[2] It can therefore be reformulated analogously to (11) with the objective function from (14). The dual problem then results in finding

$$-\mu^* = \inf_{u \in \mathbb{R}^m} \left\{ \frac{1}{2} u^T Z K Z u - e^T u \ \mid\ u^T \zeta = 0,\ Ce \geq u \geq 0 \right\}, \tag{20}$$

with the term $u^T ZKZu$ in place of the term $u^T ZQZu$ in (14). The entries of K are given by $K_{i,j} = \langle \phi(x^{(i)}), \phi(x^{(j)})\rangle_{\mathcal{W}}$. Just like the above matrix $Q \in \mathbb{R}^{m \times m}$, also $K \in \mathbb{R}^{m \times m}$ is a Gram matrix and thus positive semidefinite, $K \succeq 0$, and K does not depend on the dimension of the image space $\mathcal{W}$ — but on the number of support points m. When ϕ is the identity mapping then $K = Q$ and problem (20) coincides with problem (11).

To determine a suitable function ϕ, the so-called kernel trick is applied: instead of the transformation ϕ, only a symmetric continuous mapping

$$\kappa : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

is defined in such a way that $\kappa(x, y)$ could be interpreted as $\kappa(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{W}}$ for a function ϕ. To do this, it is required that for all $m \in \mathbb{N}$ and all $x^{(1)}, \ldots, x^{(m)}$ from Ω, the matrices $K \in \mathbb{R}^{m \times m}$ with entries $K_{i,j} = \kappa(x^{(i)}, x^{(j)})$ for $1 \leq i, j \leq m$ always satisfy $K \succeq 0$. In this case, κ is called a[3]

$$\textit{positive definite kernel.} \tag{21}$$

In summary, the matrix K is formed with $K_{i,j} := \kappa(x^{(i)}, x^{(j)})$ for $1 \leq i, j \leq m$ and (20) is solved. As in (12), the optimal solution u of (20) defines the set $\mathcal{B} := \{i \mid u_i > 0\}$ as well as the optimal solution $\tilde{a} = \sum_{i \in \mathcal{B}} u_i \zeta_i \phi(x^{(i)})$ of (16). However, the vector $\tilde{a}$ is not required explicitly (i.e., the function ϕ is not evaluated explicitly). The classification of a new data point $\tilde{x}$ is performed according to the rule

$$\tilde{\zeta} = \operatorname{sign}(\langle \tilde{a}, \phi(\tilde{x})\rangle_{\mathcal{W}} - \beta) = \operatorname{sign}(\langle \sum_{i \in \mathcal{B}} u_i \zeta_i \phi(x^{(i)}), \phi(\tilde{x})\rangle_{\mathcal{W}} - \beta) = \operatorname{sign}(\sum_{i \in \mathcal{B}} u_i \zeta_i \kappa(x^{(i)}, \tilde{x}) - \beta), \tag{22}$$

[2] Because the scalar product $\langle\ .\ ,\ .\ \rangle_{\mathcal{W}}$ in (17) is applied to arbitrary $\tilde{x} \in \Omega$, it is reasonable to formulate problem (16) over the general space $\mathcal{W}$ and not to restrict oneself to the finite-dimensional formulation (19) to begin with.

[3] At this point, the notation "positive semidefinite kernel" would be appropriate. The fact that the kernel defines a norm in a certain space justifying for the notation "positive definite kernel" will be provided later in (27).

where β is obtained by exploiting $\langle \tilde{a}, \phi(x^{(i)})\rangle_{\mathcal{W}} = \sum_{j\in\mathcal{B}} u_j\zeta_j\kappa(x^{(i)}, x^{(j)})$ as with the soft-margin SVM: First, $\zeta \in \mathbb{R}^m$ is set to the vector with components ζ_i, and $b \in \mathbb{R}^m$ as the vector with components $b_i = 1 - \zeta_i \sum_{j\in\mathcal{B}} u_j\zeta_j\kappa(x^{(i)}, x^{(j)})$. Then, again

$$\beta = \text{argmin } \{e^T \max\{b + \hat{\beta}\zeta,\ 0\} \mid \hat{\beta} \in \mathbb{R}\}.$$

In the next paragraph, the question is considered whether for a given positive definite kernel $\kappa : \Omega \times \Omega \to \mathbb{R}$, there also is a function $\phi : \Omega \to \mathcal{W}$ such that for given data points $x^{(i)} \in \Omega$ for $1 \le i \le m$, the relationship

$$\kappa(x^{(i)}, x^{(j)}) = \langle \phi(x^{(i)}), \phi(x^{(j)})\rangle_{\mathcal{W}} \quad \text{for} \quad 1 \le i, j \le m \tag{23}$$

holds true. Since the number of equations in (23) to be satisfied by ϕ increases with m and no upper bound is set for m, it is natural to allow an infinite value for the degrees of freedom of ϕ, i.e., for the dimension of $\mathcal{W}$.

4.1 Interpretation in the feature space

The image space $\mathcal{W}$ of the above function ϕ is also called the "feature space"; it is the space where the linear separation of the two classes is performed. For a given κ, neither the mapping ϕ nor its image space $\mathcal{W}$ is unique. The existence and desirable properties of ϕ will be considered below.

4.1.1 Existence

If the order of reasoning in (23) is reversed, and for a given κ and for points $x^{(i)}$ that are given "first", a function ϕ is sought, then the existence of ϕ can be established easily: For $K \succeq 0$, there exists an eigenvalue decomposition, $K = U^T DU$ with an orthogonal matrix U with columns $u^{(i)}$ and a diagonal matrix D. Setting $\tilde{u}^{(i)} := D^{1/2}u^{(i)}$, it holds that $K_{i,j} = (\tilde{u}^{(i)})^T\tilde{u}^{(j)}$. Thus, a mapping $\phi : \mathbb{R}^n \to \mathbb{R}^m =: \mathcal{W}$ can be defined arbitrarily, so that for $1 \le i \le m$, the interpolation conditions $\phi(x^{(i)}) = \tilde{u}^{(i)}$ are fulfilled, and thus also the desired relationship $K_{i,j} = \langle \phi(x^{(i)}), \phi(x^{(j)})\rangle_{\mathcal{W}}$ holds true with $\langle ., .\rangle_{\mathcal{W}}$ being the standard 2-norm scalar product. (This reasoning uses the fact that the dimension of $\mathcal{W}$ was assumed to be chosen at will.) However, compared to (23), the order is reversed here: The points $x^{(i)}$ are used to define the function ϕ without requiring any form of continuity of ϕ. The reason why such approach is problematic will be explained below using a simple example.

4.1.2 Overfitting

If, for example, many measurements (x_i, y_i) of a function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto y$ are given, which approximately lie on a straight line, and if the value of f at a "new" point $\tilde{x}$ needs to be estimated, one could on the one hand define a line $g : x \mapsto ax + b$ so that g approximates the measurements in a certain sense as accurately as possible. This leads to the so-called least squares problem for determining the two parameters $a, b \in \mathbb{R}$, and to the approximation $f(\tilde{x}) \approx g(\tilde{x})$. On the other hand, one could also determine a polynomial p of high degree such that all measurements are exactly interpolated, and then approximate $f(\tilde{x}) \approx p(\tilde{x})$. Typically, such a polynomial oscillates strongly, and therefore provides a very unreliable prediction of $f(\tilde{x})$. The higher number of adjustable parameters in p compared to just two parameters in g does not provide a more reliable approximation. This well-known fact is also called overfitting.

For support vector machines, often there are also many data points available, and the separation into two classes is to be made based on unknown "similarity properties". As in the example above, here as well, the approach of constructing a function ϕ so that all training data can be correctly separated does not guarantee a reliable classification by itself. The goal is to find a mapping ϕ that, on the one hand, does not depend on the specific choice of $x^{(i)}$ (these only determine the separating hyperplane in the space $\mathcal{W}$) and, on the other hand, is chosen so that ϕ does not behave "too chaotically" (does not "oscillate too much") but rather possesses certain continuity properties that preserve the assumed but unknown "similarity properties" of the original data points.

4.1.3 Cross-validation

The so-called cross-validation provides an approach to estimate the reliability of the classification. A simple approach to cross-validation is as follows: Assume that the training data were generated randomly and independently. Then one can randomly divide the training data into two parts, e.g., 70% in one part and the rest in the other, and then calculate the separation using only the 70%. The 30% that were not used in generating the SVM but for which the classification is known, are then used to estimate which percentage of the correctly classified data points, and this can be used as an estimated error rate for the separation. The actual separation can then be performed using all training data. By monotonicity of the separation quality with increasing size of the training data the "70% estimator" can serve as an estimate of the overall error rate – but only for the error rate for new data generated from the same distribution.

4.1.4 Example: Gaussian kernel

Definition 1. *Let a constant $c > 0$ be given. The function $\kappa : \Omega \times \Omega \to \mathbb{R}$ with*

$$\kappa(x, y) := e^{-c\|x-y\|_2^2}$$

is referred to as the Gaussian kernel, inspired by the Gaussian distribution curve.

The associated kernel matrix K has the entries $K_{i,j} = \kappa(x^{(i)}, x^{(j)}) := e^{-c\|x^{(i)}-x^{(j)}\|_2^2}$. In Section 7.1 in the appendix, it is briefly explained that the above K is always positive semidefinite. However, it is not immediately obvious how to determine the scalar product $\langle ., . \rangle_{\mathcal{W}}$ and a mapping ϕ such that

$$\langle \phi(x), \phi(y) \rangle_{\mathcal{W}} \equiv e^{-c\|x-y\|_2^2}$$

always holds. If such a mapping ϕ exists, it immediately follows that

$$\|\phi(x)\|_{\mathcal{W}}^2 = \langle \phi(x), \phi(x) \rangle_{\mathcal{W}} = \kappa(x, x) = e^0 = 1$$

for all $x \in \Omega$.

Definition 2. *In the following we call κ with $\|\phi(x)\|_{\mathcal{W}}^2 = \kappa(x, x) \equiv 1$ an*

iso-normalized *kernel.*

Any kernel for which $\kappa(x, x) > 0 \quad \forall x \in \Omega$ holds true can be scaled diagonally in the manner described in Section 7.1, such that it is iso-normalized. The scaling is typically nonlinear and thus also alters the separation properties.

4.2 Determining the kernel and the scalar product

The construction of a map ϕ and associated scalar product $\langle .,. \rangle_{\mathcal{W}}$ that is more suitable than the motivation given in Section 4.1.1 is based on [6]: Let a continuous, symmetric positive definite kernel κ be given on the compact convex set $\Omega \subset \mathbb{R}^n$. For a fixed $x \in \Omega$, we define the mapping $\phi(x) := K_x : \Omega \to \mathbb{R}$ as follows:

$$K_x := \kappa(x, \ . \ ), \qquad \text{i.e.,} \qquad \phi(x)[z] \equiv K_x(z) \equiv \kappa(x,z) \ \text{ for } z \in \Omega.$$

To avoid confusion regarding the fact that $\phi(x)$ itself is a function, we use the more intuitive notation K_x instead of $\phi(x)$. The finite linear combinations of such functions K_x then form the space

$$\mathcal{W} := \mathrm{Span}(\{K_x \mid x \in \Omega\}) = \{f \mid \exists k \in \mathbb{N},\ x^{(i)} \in \Omega,\ \alpha_i \in \mathbb{R}\ (1 \le i \le k) :\ f = \sum_{i=1}^{k} \alpha_i K_{x^{(i)}}\}.$$

Furthermore, for $f, g \in \mathcal{W}$ with $f := \sum_{i=1}^{k} \alpha_i K_{x^{(i)}}$ and $g := \sum_{j=1}^{\ell} \beta_j K_{x^{(j)}}$ let the mapping $\langle .,. \rangle : \ \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be defined as follows:

$$\langle f, g \rangle := \sum_{i=1}^{k} \sum_{j=1}^{\ell} \alpha_i \beta_j \kappa(x^{(i)}, x^{(j)}). \tag{24}$$

It is shown next that $\langle .,. \rangle$ indeed is a scalar product: First, we need to justify that $\langle .,. \rangle$ is well-defined: Since it is not assumed that all $K_{x^{(i)}}$ are linearly independent, there could be different representations for a given function $g \in \mathcal{W}$. However, the above mapping $\langle .,. \rangle$ is independent of the representation of g since from (24) it follows that

$$\langle f, g \rangle = \sum_{i=1}^{k} \alpha_i \underbrace{\left(\sum_{j=1}^{\ell} \beta_j \kappa(x^{(i)}, x^{(j)}) \right)}_{= \sum_{j=1}^{\ell} \beta_j K_{x^{(j)}}(x^{(i)}) = g(x^{(i)})} = \sum_{i=1}^{k} \alpha_i g(x^{(i)}).$$

The right-hand side does not depend on the chosen coefficients $x^{(j)}$ and β_j for the representation of g, but only on the function values $g(x^{(i)})$. In addition, the right-hand side is linear in g. Analogously, it follows that the mapping is also independent of the representation of f and linear in f, i.e., $\langle f, g \rangle$ is bilinear – and, like κ, also symmetric. Finally, positive semidefiniteness is inherited:

$$\langle f, f \rangle = \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j \kappa(x^{(i)}, x^{(j)}) \ge 0$$

according to the definition of a positive definite kernel. Therefore, the Cauchy-Schwarz inequality also holds, $\langle f, g \rangle^2 \le \langle f, f \rangle \langle g, g \rangle$ (with the usual proof[4]).

[4] Due to bilinearity and semidefiniteness, we have $0 \le \langle f - \lambda g, f - \lambda g \rangle = \langle f, f \rangle - 2\lambda \langle f, g \rangle + \lambda^2 \langle g, g \rangle$ for $\lambda \in \mathbb{R}$. So $2\lambda \langle f, g \rangle \le \langle f, f \rangle + \lambda^2 \langle g, g \rangle$ for all λ. If $\langle g, g \rangle = 0$, this implies $\langle f, g \rangle = 0$, i.e., $\langle f, g \rangle^2 = 0 \le \langle f, f \rangle \langle g, g \rangle$, and if $\langle g, g \rangle > 0$, then, with the choice $\lambda := \langle f, g \rangle / \langle g, g \rangle$, we get

$$2 \tfrac{\langle f,g \rangle^2}{\langle g,g \rangle} \le \langle f, f \rangle + \langle g, g \rangle \tfrac{\langle f,g \rangle^2}{\langle g,g \rangle^2} \qquad \text{i.e.,} \qquad \tfrac{\langle f,g \rangle^2}{\langle g,g \rangle} \le \langle f, f \rangle \qquad \text{or} \qquad \langle f, g \rangle^2 \le \langle f, f \rangle \langle g, g \rangle.$$

Now, for $x \in \Omega$, $f = \sum_{i=1}^{k} \alpha_i K_{x^{(i)}} \in \mathcal{W}$, and $g := K_x$ in (24), we also have

$$\langle f, \kappa(.,x)\rangle = \langle f, K_x\rangle \overset{(24)}{=} \sum_{i=1}^{k} \alpha_i \kappa(x^{(i)}, x) = \sum_{i=1}^{k} \alpha_i K_{x^{(i)}}(x) = f(x), \tag{25}$$

a central property known as the "reproducing kernel" property. Let $x, z \in \Omega$ and $g := \kappa(.,z)$, then we also have

$$\langle K_x, K_z\rangle = \langle \kappa(.,x), \kappa(.,z)\rangle = \langle \kappa(.,x), g\rangle = \langle g, \kappa(.,x)\rangle \overset{(25)}{=} g(x) = \kappa(x,z). \tag{26}$$

Using (25) and the Cauchy-Schwarz inequality, for $f \in \mathcal{W}$ and $x \in \Omega$, we further have

$$f(x)^2 = (\langle \kappa(.,x), f\rangle)^2 \leq \langle \kappa(.,x), \kappa(.,x)\rangle \langle f, f\rangle = \kappa(x,x)\langle f, f\rangle,$$

where the last equation follows from (26). Therefore, if $\langle f, f\rangle = 0$, then $f(x) \equiv 0$ for $x \in \Omega$, i.e.,

$$\langle ., .\rangle_{\mathcal{W}} := \langle ., .\rangle \text{ is a scalar product,} \tag{27}$$

which induces a norm $\| \, . \, \|$ on $\mathcal{W}$)[5]. In general, $\mathcal{W}$ is not complete with respect to this norm. What is important, as motivated earlier, is the continuity of $\phi : \Omega \to \mathcal{W}$, which is discussed in Section 5.1. But first, the existence of a separation is addressed.

4.2.1 Separability in the feature space

In the case where κ is chosen so that the functions $\phi(x^{(1)}), \ldots, \phi(x^{(m)})$ are linearly independent[6], the existence of a separating hyperplane can be established explicitly even without a soft margin:

To do this, we consider the problem (19) and use the notation

$$\tilde{a}_M = \sum_{i=1}^{m} \alpha_i \phi(x^{(i)}) = \left[\phi(x^{(1)}), \ldots, \phi(x^{(m)})\right] \alpha$$

and fix $C = \infty$, i.e., $s = 0$. Then (19) is given by

$$\begin{aligned} &\min_{\tilde{a}_M \in M,\ \beta} \{ \ \frac{1}{2}\|\tilde{a}_M\|_{\mathcal{W}}^2 \ \mid \ \zeta_i \left(\langle \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}} - \beta \right) \geq 1, \quad \forall\ 1 \leq i \leq m \} \qquad (28) \\ = \quad &\min_{\alpha \in \mathbb{R}^m,\ \beta} \{ \ \frac{1}{2}\alpha^T K \alpha \ \mid \ ZK\alpha - \beta\zeta \geq e \}, \end{aligned}$$

where the entries of the matrix K are again given by $K_{i,j} = \langle \phi(x^{(i)}), \phi(x^{(i)})\rangle_{\mathcal{W}}$. Since K is invertible by assumption, and $Z = Z^{-1}$, the latter problem has the feasible solution $\beta = 0$ and $\alpha = K^{-1}Ze$. Together with $s := 0$ this is also feasible for (19) with objective value $\frac{1}{2}e^T ZK^{-1}Ze$. This feasible solution generally is not optimal, but it can be seen that the corresponding objective value generally increases as the smallest eigenvalues of K approach zero, an observation that often also applies to the optimal solution of (28).

If one chooses a kernel for which the $\phi(x^{(i)})$ are always linearly independent for pairwise different $x^{(i)}$, i.e., for which an exact separation of the data is always achievable, then the problem of overfitting from Section 4.1.2 arises again. Therefore, it is also usually advisable to choose an approach with a soft margin for such kernels.

[5] The property of being a norm may seem surprising at first, since the $K_{x^{(i)}}$ were not assumed to be linearly independent and only $K \succeq 0$ was demanded, but for linearly dependent $K_{x^{(i)}}$, the space $\mathcal{W}$ is also smaller. On $\mathcal{W}$, the scalar product is (strictly) positive definite. However, as motivated in Section 4.2.1, linear independence of the $K_{x^{(i)}}$ for pairwise distinct points $x^{(i)}$ is a desirable property.

[6] As discussed in the Appendix 7.2, this assumption is always fulfilled for the Gaussian kernel.

4.3 Summary, kernel SVM

Choose a kernel function $\kappa : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ so that the matrix K with entries $K_{i,j} = \kappa(x^{(i)}, x^{(j)})$ satisfies $K \succeq 0$, and let Z be the diagonal matrix with diagonal entries ζ_i. Solve

$$\min_{u \in \mathbb{R}^m} \left\{ \frac{1}{2} u^T Z K Z u - e^T u \;\mid\; u^T \zeta = 0,\ Ce \geq u \geq 0 \right\}, \tag{29}$$

set $b = e - ZKZu$, find β as in (13), i.e. as the minimizer of "$e^T \max\{b + \hat{\beta}\zeta, 0\}$" for $\hat{\beta} \in \mathbb{R}$, and classify a new point $\tilde{x}$ via

$$\tilde{\zeta} := \operatorname{sign}((\sum_{i \in \mathcal{B}} u_i \zeta_i \kappa(x^{(i)}, \tilde{x})) - \beta)$$

where $\mathcal{B} := \{i \mid u_i > 0\}$.

5 Kernel properties

5.1 Relative Lipschitz condition

In the following, it is assumed that the data space has been rescaled beforehand, such that the maximum norm of the data from Ω is on the order of 1 and that the similarity of data from $\Omega \subset \mathbb{R}^n$ can be measured in the 2-norm. The latter assumption depends on the specific application.

Since "similar data" generally should be classified similarly and the separation of the data in the kernel approach is done via the scalar product $\langle ., . \rangle_{\mathcal{W}}$, we are now looking for a mapping ϕ such that for small $\|x - y\|_2$, also $\|\phi(x) - \phi(y)\|_{\mathcal{W}}$ is small. A stronger requirement regarding continuity of ϕ (e.g., a global Lipschitz property that holds for large $\|x - y\|_2$ as well) generally cannot be justified for the SVM approach.

When considering the continuity properties of ϕ in (16), it should also be noted that ϕ and β can be multiplied by an arbitrary factor $\lambda > 0$ without changing the separation (17). However, if ϕ were, for example, locally Lipschitz continuous, then multiplication by λ would also change the Lipschitz constant by a factor of λ. Therefore, the Lipschitz constant should be considered relative to the norm of ϕ. As a possible requirement on the kernel, the (local) relative Lipschitz condition can be considered

$$\frac{\|\phi(x) - \phi(y)\|_{\mathcal{W}}}{\|\phi(x)\|_{\mathcal{W}}} \leq \gamma \|x - y\|_2 \tag{30}$$

for small $\|x - y\|_2$ with a local relative Lipschitz constant $\gamma > 0$. This condition implies $\phi(x) \neq 0$ for $x \in \Omega$, a requirement that is always fulfilled for iso-normalized mappings. By construction, it is invariant under the transition from $\phi(\ .\ )$ to $\lambda\phi(\ .\ )$ for some fixed $\lambda \neq 0$. But it is not invariant under a scaling of the data when replacing Ω with $\lambda\Omega$ for some fixed $\lambda \neq 0$, i.e., under the transition from $\phi(\ .\ )$ to $\phi(\lambda(\ .\ ))$. In particular, the right-hand-side changes by a factor $|\lambda|$ under such scaling. To limit the effects of such scaling, it was assumed in the beginning of this section that the maximum norm of the elements in Ω is on the order of 1.

Small values of γ in (30) guarantee "high data consistency" in the sense that closely neighboring data points x, y will have closely neighboring images $\phi(x), \phi(y)$, while larger values allow more flexibility in the form of separation.

Note that ϕ and $\langle\ .\ ,\ .\ \rangle_{\mathcal{W}}$ used in Condition (30) are *not* determined uniquely[7] for a given kernel κ. Nevertheless, the following theorem holds:

[7] One may change $\langle\ .\ ,\ .\ \rangle_{\mathcal{W}}$ for example using a bounded invertible linear operator and adjust ϕ accordingly.

Theorem 1. *If the function ϕ is given by a three times continuously differentiable positive definite kernel κ (see (21)) and the relative Lipschitz condition (30) holds for some constant $\gamma > 0$,*

$$\frac{\|\phi(x) - \phi(y)\|_{\mathcal{W}}}{\|\phi(x)\|_{\mathcal{W}}} \leq \gamma \|x - y\|_2 \qquad \text{for small } \|x - y\|_2,$$

then, for the mixed second derivative of κ at $z \in \Omega$:

$$\|D_x(\nabla_y \kappa(x, y))|_{x=z,\ y=z}\|_2 \leq \gamma^2 \kappa(z, z). \tag{31}$$

Conversely, if (31) is satisfied, then the condition (30) holds in the following form:

$$\frac{\|\phi(x) - \phi(y)\|_{\mathcal{W}}}{\|\phi(x)\|_{\mathcal{W}}} \leq e^{\gamma \|x-y\|_2} - 1 \qquad (\approx \gamma \|x - y\|_2 \ \text{ for small } \|x - y\|_2).$$

A proof is provided in Section 7.3. The condition (31) is slightly more precise than (30), as the latter does not specify what exactly is meant by "small $\|x - y\|_2$". Therefore, (31) will always be used in the following.

Definition 3. *Following [9], a kernel satisfying (31) will also be referred to as*[8]

γ-self-concordant kernel

with a local relative Lipschitz constant $\gamma > 0$.

From the representation (31), the following construction guideline for γ-self-concordant kernels can be derived directly:

Lemma 1. *If κ_1 and κ_2 are kernels that satisfy the conditions of Theorem 1 with Lipschitz constants γ_1 and γ_2, respectively, then for $\rho > 0$, $\rho\kappa_1$ is also a γ_1-self-concordant kernel. Furthermore, $\kappa_1 + \kappa_2$ is a γ-self-concordant kernel with $\gamma = \max\{\gamma_1, \gamma_2\}$. If κ_1 and κ_2 are iso-normalized kernels, then γ can be tightened to $\gamma = \frac{1}{2}(\gamma_1 + \gamma_2)$.*

5.2 Condition number of the kernel matrix

The condition number of the kernel matrix may play a crucial role for the quality of the separation: Consider an iso-normalized $m \times m$-kernel matrix K. Then, by positive semidefiniteness, all matrix entries have absolute value at most one, and thus, by Gershgorin's theorem (see, for example, [16]), the largest eigenvalue of K is at most m. By the interlacing property (see, for example, [5]) it is at least 1 (the eigenvalue of a 1×1 principal submatrix), but the smallest eigenvalue can be tiny. If the condition number of K is a moderate number M, a "reasonable" feasible solution of (28) is given by choosing $\alpha = K^{-1}\zeta$ and $\beta = 0$. It is feasible because $Z\zeta = e$, and a straightforward calculation shows that the objective value of (28) is bounded by the moderate value $Mm/2$ independent of ζ. In this case, a separation with a moderately wide margin is always possible. The perfect condition number of K (with respect to the 2-norm) is obtained if, and only if[9], K is iso-normalized and

$$\langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle_{\mathcal{W}} = \kappa(x^{(i)}, x^{(j)}) = K_{i,j} = 0 \tag{32}$$

[8] The concept that the derivatives of a function are bounded by constant multiples of other derivatives, and hence are in "concordance" with themselves, was introduced in [9]. There, Newton's method for θ-self-concordant barrier functions is examined. In the case of the kernel functions considered here, a bound of the mixed second derivatives by the "zeroth" derivative (the function value) is of interest.

[9] (Because positive multiples of the identity matrix are the only symmetric positive definite matrices with condition number 1.)

for all $i \neq j$.

However, the aim of generating a well conditioned kernel may be in conflict with the relative Lipschitz condition with a small Lipschitz constant: For data $x^{(i)}$ and $x^{(j)}$ with small value $\|x^{(i)} - x^{(j)}\|_2$ the Lipschitz condition requires that also $\|\phi(x^{(i)}) - \phi(x^{(j)})\|_{\mathcal{W}}$ is small so that K has two nearly linearly dependent columns and thus K has a small positive eigenvalue and a large condition number. In such situation a separation with wide margin is no longer possible for all choices of ζ. (In particular, it is not possible when $x^{(i)} \approx x^{(j)}$ but $\zeta_i = -\zeta_j$.) A relaxation of (32) that is compatible with the γ-self-concordance is the requirement of choosing κ such that

$$\kappa(x^{(i)}, x^{(j)}) \approx 0 \quad \text{whenever} \quad \|x^{(i)} - x^{(j)}\|_2 \quad \text{is large.} \tag{33}$$

This requirement is satisfied, for example, for the Gaussian kernel. And as shown below, the Gaussian kernel is also optimal with respect to the Lipschitz constant.

5.2.1 Effects of iso-normalization

The following guidelines hold: Iso-normalization always[10] improves the condition number of all 2×2 principal submatrices of K (unless they are already iso-normalized), and by the interlacing theorem the condition number of K always is at least as large as the largest condition number of any 2×2 principal submatrix. By Theorem 4.1 in [15], when the dimension n is larger than 2, the iso-normalized scaling is not too far (namely by a factor at most n) from the optimal scaling. On the other hand, there are examples where the condition number of an iso-normalized matrix indeed is larger than $(n/2 - \epsilon)$-times the value of an optimally rescaled matrix, [14]. Summarizing, iso-normalization not only improves the worst-case bound of the condition number compared to arbitrary scaling, but as shown in Section 5.4 below, it may also improve the Lipschitz constant γ.

5.3 The Gaussian kernel

5.3.1 Lipschitz constant:

Consider again the Gaussian kernel $\kappa(x, y) = e^{-c\|x-y\|_2^2}$. Here, $D_y\kappa(x, y) = 2c(x - y)^T\kappa(x, y)$ and

$$D_x\nabla_y\kappa(x, y)|_{x=y=z} \equiv ((4c^2(x - y)(x - y)^T + 2cI)\kappa(x, y))|_{x=y=z} = 2c\kappa(z, z)I.$$

The requirement of the relative Lipschitz condition therefore is

$$\|2cI\|_2 \leq \gamma^2,$$

which means $\gamma = \sqrt{2c}$ can be chosen here. Observe that for this choice of γ, Condition (31) is satisfied with equality for all z, i.e., the local relative Lipschitz constant γ for ϕ (in Theorem 1) is equal to the maximum allowed value at all points in Ω. (As will be discussed below, large values γ allow kernels with lower condition numbers.) Summarizing we obtain the following lemma:

Lemma 2. *The Gaussian kernel $\kappa(x, y) \equiv e^{-c\|x-y\|_2^2}$ with $c > 0$ satisfies the relative Lipschitz condition uniformly for all $x \in \Omega$ with Lipschitz constant $\sqrt{2c}$.*

For the Gaussian kernel with $c > 0$, it is shown in Section 7.2 that for any pairwise distinct $x^{(i)}$, the functions $K_{x^{(i)}}$ are linearly independent, meaning the dimension of $\mathcal{W}$ is infinite and

[10] The straightforward proof of this claim is left as an exercise.

exact separation always is possible. However, the space $\mathcal{W}$ depends on the choice of c. For $x \in \Omega$, the function $\phi(x)$ has the form

$$\phi(x)[y] \equiv e^{-c\|x-y\|_2^2} \quad \text{for } y \in \Omega.$$

As $c \to \infty$, the function $\phi(x)$ converges to the characteristic function of the point $x \in \Omega$, and the matrix K in Section 4.2.1 converges to the identity matrix for any choice of (pairwise distinct) data points $x^{(i)}$. On the other hand, for $c \to 0$, $\phi(x)$ converges to the constant function 1 on Ω, and the matrix K tends to the rank-1 matrix ee^T. For small $c > 0$, the optimal value of (28) generally tends to infinity.

5.3.2 Adjusting the soft margin:

In addition to the parameter c in the Gaussian kernel (or the Lipschitz constant $\gamma = \sqrt{2c}$), the constant C from the soft-margin approach in (5), which penalizes the violation of the separation properties, is a freely adjustable parameter when using Gaussian kernels. For large finite values of C, it may be the case that the optimal solution of (19) leads to a separation that correctly separates all training data points, but some of the training data points (i.e. some of the support vectors) are closer to the separating hyperplane (in the space $\mathcal{W}$) than others. (In the case of separation with $C = \infty$, within the space $\mathcal{W}$ all support vectors are equidistant from the separating hyperplane, and there are no data points lying closer.)

Figures 1-3 are intended to illustrate the effects of both parameters for an example where there are no misclassified data points. The black stars indicate the data points. The label for each data point was assigned along a 3×3 checkerboard pattern with $\zeta = -1$ in the middle and in the four corners, and $\zeta = +1$ in the remaining fields. 50 data points were randomly chosen with higher probability to lie near the center. The green areas in the plots below mark those points that will be assigned the label $+1$ while red points will be assigned to -1.

Larger values of c generally may lead to a "more curvy" boundary between the two areas, and lower values of C allow more misclassifications. Also, smaller values of c generally match larger values of C. The right patterns in the three rows of images below do classify all training data points correctly, i.e. each of the right patterns might depict the true division between red and green points. (One of the other patterns might be correct if some of the training data were misclassified.)

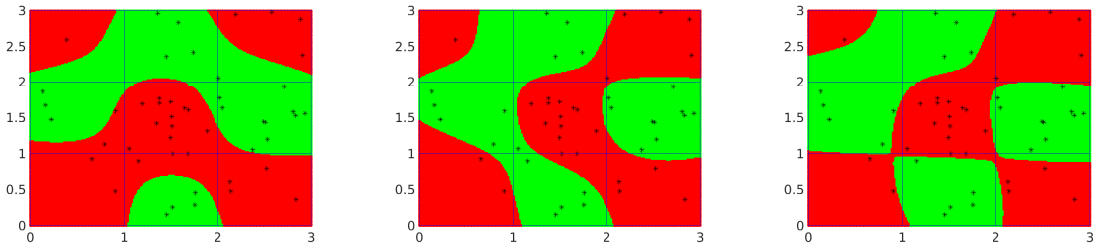

Figure 1: Gaussian kernel with $c = 1$ and $C = 10$, $C = 100$, $C = 1000$ ($C = \infty$) from left to right.

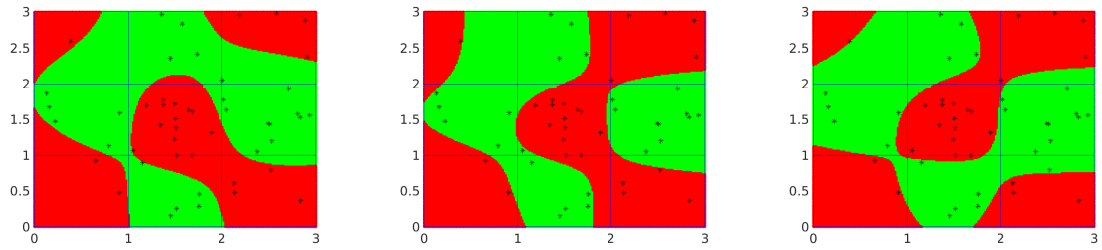

Figure 2: Gaussian kernel with $c = 0.03$ and $C = 10^7$, $C = 10^8$, $C = 10^9$ ($C = \infty$) from left to right.

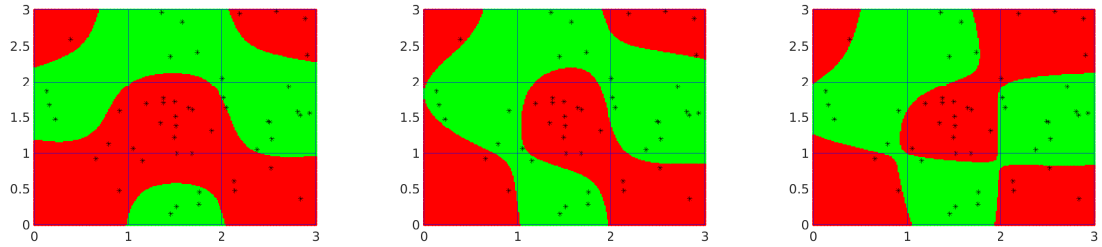

Figure 3: Gaussian kernel with $c = 0.001$ and $C = 10^{12}$, $C = 10^{13}$, $C = 10^{14}$ from left to right.

5.3.3 Modifying the cross-validation strategy:

A simple procedure for identifying c and C uses the following approach. If the set of training data is divided into 3 disjoint parts, then the concept of cross-validation from Section 4.1.3 can be generalized. First, with the first part, several choices of the parameters c and C can be used to determine the respective SVM. The parameters that best classify the second part are then chosen, and with this choice, the error rate is estimated using the third part. This estimate is kept when ultimately the SVM is formed over all training data points with these parameters. (This approach is not optimal; there are a number of newer works on hyper-parameter optimization using repeated cross-validation that are more sophisticated than outlined here. The outlined approach is intended to demonstrate the basic feasibility of how hyperparameters[11] can be chosen appropriately and to sensitize for the risk of not using data for error estimation that were already used for the design of the algorithm to be evaluated.)

5.3.4 Orthogonal invariance:

Next to Theorem 1, there is often another desirable invariance property: the classification of the SVM should be invariant when training and test data are all equally reflected at a hyperplane passing through the origin or rotated about the origin. Since every orthogonal transformation can be represented by a sequence of reflections and rotations, this requirement implies that the classification of the SVM should be invariant under orthogonal linear transformations. For the Gaussian kernel, $\kappa(x, y)$ depends only on $\|x - y\|_2$. It follows that κ remains unchanged

[11] In general, algorithmic parameters such as step sizes, etc. are referred to as hyperparameters as opposed to problem parameters like the number of data points.

with input x, y and with input $U(x), U(y)$ if U is an orthogonal linear transformation. Similarly, the calculation of $\tilde{\zeta}$ based on the optimal solution of (20) is the same whether $\kappa(x^{(i)}, \tilde{x})$ or $\kappa(U(x^{(i)}), U(\tilde{x}))$ is used. Thus, the Gaussian kernel is independent of orthogonal transformations of the input. This is noteworthy since the discrete Fourier transform (DFT), which is frequently used in sound and image processing, is an orthogonal transformation (up to a constant factor which can be accounted for in the term "c" of the Gaussian kernel). Hence, it does not matter whether the original data or their DFT are used as input for the SVM; the separation remains the same!

5.3.5 Preprocessing:

In many real-world applications, preprocessing of the data is a tedious but essential detail, and this may also be true for SVMs. While a given SVM may be invariant under certain simultaneous preprocessing of *all* data points, it may still be useful to transform each *individual* data point to a certain standard form, for example, by translating, rotating, or scaling it before using it as input for the SVM. This may change the separation significantly. In case of such preprocessing the input to the SVM should be augmented by the parameters that were used for each individual transformation.

For example, consider the task of detecting a given digit in a text. If it is also necessary to distinguish between the letters | and /, and if the preprocessing includes a rotation that aligns each new data points before being put into the SVM, then the aligned "/" would resemble a "|". Hence, the parameters used in the preprocessing (such as parameters for a possible rotation) should also be provided as additional input variables to the SVM.

5.4 Other commonly used kernels

1) A commonly used polynomial kernel κ is given by

$$\kappa(x, y) \equiv (x^T y + 1)^p \quad \text{with} \quad p \in \mathbb{N}. \tag{34}$$

As noted in Section 7.1, this kernel is also positive definite. Here,

$$D_y \kappa(x, y) = p x^T (x^T y + 1)^{p-1},$$

and

$$\begin{aligned} D_x(\nabla_y \kappa(x, y))|_{x=y=z} &= D_x(p x (x^T y + 1)^{p-1})|_{x=y=z} \\ &= p(p-1)(z^T z + 1)^{p-2} z z^T + p(z^T z + 1)^{p-1} I \end{aligned}$$

with the identity matrix I. Due to $\|z z^T\|_2 = z^T z$ the relation (31) then reads as

$$p(z^T z + 1)^{p-2} \left((p-1) z^T z + (z^T z + 1)\right) \leq \gamma^2 (z^T z + 1)^p.$$

or

$$p(p z^T z + 1) \leq \gamma^2 (z^T z + 1)^2. \tag{35}$$

This requirement shall be satisfied for all $z \in \Omega$. When $p \geq 2$ this is true for $\gamma = p/\sqrt{2}$)[12]. Here, $\phi(x) \in \mathcal{W}$ is a polynomial in n variables of maximum degree p, so the dimension of $\mathcal{W}$ is bounded by n^{p+1}. Higher values of p improve the separation capabilities at the expense of a larger Lipschitz constant. Summarizing the following lemma is true:

[12] For $z = 0$, the requirement (35) states that $p \leq \gamma^2$, which is fulfilled with $\gamma := p/\sqrt{2}$ because $p \geq 2$. Setting $t := z^T z$ and $\ell : \mathbb{R} \to \mathbb{R}$ with $\ell(t) := \gamma^2 (t+1)^2 - p(pt+1)$ it suffices to show that $\ell(t) \geq 0$ for $t \geq 0$. Because $\ell'(0) = 2\gamma^2 - p^2 = 0$ and because ℓ is a convex quadratic function, it follows that $\ell(t) \geq \ell(0) \geq 0$ for $t \geq 0$. #

Lemma 3. *The polynomial kernel (34) satisfies the self-concordance condition (31) with the local relative Lipschitz constant* $\gamma = p/\sqrt{2}$.

2) More generally, one can also consider kernels of the form

$$\kappa(x, y) := (x^T y + \alpha)^p$$

with a parameter $\alpha > 0$ and integer $p \geq 2$. With the same calculations this results in a Lipschitz constant of $\gamma = p/\sqrt{2\alpha}$, i.e., the Lipschitz constant also is a continuously adjustable hyperparameter. (Large values of α result in a nearly constant kernel comparable to tiny values of $c > 0$ in the Gaussian kernel, while large values of p in some form have an opposite effect.) When $\alpha = 0$ this kernel still is a positive definite kernel, but the condition (31) for this kernel takes the form

$$p^2 \leq \gamma^2 z^T z$$

and (for small $\|z\|_2$) this cannot be satisfied. (This kernel is also rarely used.)

3) One can also consider kernels κ of the form

$$\kappa(x, y) \equiv h(x)h(y)g(x^T y) \tag{36}$$

with smooth functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$. If g can be represented as a power series with nonnegative coefficients, $g(t) \equiv \sum_{j=0}^{\infty} a_j t^j$ with $a_j \geq 0$ for all j, then (as shown in Section 7.1) the kernel is positive definite. Condition (31) is then derived as follows:

$$D_y \kappa(x, y) = D_y(h(x)h(y)g(x^T y)) = h(x)Dh(y)g(x^T y) + h(x)h(y)x^T g'(x^T y),$$

For $D_x \nabla_y \kappa(x, y)|_{x=y=z}$, one obtains

$$\nabla h(z)Dh(z)g(z^T z)+h(z)\nabla h(z)z^T g'(z^T z)+zDh(z)h(z)g'(z^T z)+h(z)^2(Ig'(z^T z)+zz^T g''(z^T z)).$$

By omitting the arguments z or $z^T z$ for h or g and their derivatives, the requirement is

$$\|g\nabla h\nabla h^T + hg'(\nabla h\ z^T + z\ \nabla h^T) + h^2(g'I + g''zz^T)\|_2 \leq \gamma^2 h^2 g. \tag{37}$$

For the choice $h(z) = g^{-1/2}(z^T z)$, leading to an iso-normalized kernel, the requirement (37) then is

$$\|g\nabla h\nabla h^T + \tfrac{g'}{g^{1/2}}(\nabla h\ z^T + z\ \nabla h^T) + \tfrac{1}{g}(g'I + g''zz^T)\|_2 \leq \gamma^2.$$

Using $\nabla h(z) = \nabla_z(g(z^T z)^{-1/2}) = -g(z^T z)^{-3/2}g'(z^T z)z$, we obtain

$$\| - \tfrac{1}{g^2}(g')^2 zz^T + \tfrac{1}{g}(g'I + g''zz^T)\|_2 \leq \gamma^2.$$

For the polynomial kernel (34) with $g(t) \equiv (t + 1)^p$ we obtain for the above choice of h,

$$\| - \tfrac{p^2}{(t+1)^2} zz^T + \tfrac{p}{t+1} I + \tfrac{p(p-1)}{(t+1)^2} zz^T\|_2 \leq \gamma^2$$

with $t = z^T z \geq 0$, i.e.,

$$p\|I - \tfrac{1}{(t+1)} zz^T\|_2 \leq \gamma^2(t + 1),$$

which is satisfied for $\gamma = \sqrt{p}$, a significant improvement compared to (35). In addition to the improvement of the Lipschitz constant, also the matrix K generally is better scaled. The improvement of the Lipschitz constant compared to Lemma 3 is summarized in the next lemma.

Lemma 4. *For the kernel (36) with* $g(t) \equiv (t + 1)^p$ *as in (34) and* $h(z) \equiv g(z^T z)^{-1/2}$ *the local relative Lipschitz constant in (31) can be chosen as* $\gamma = \sqrt{p}$.

4) Other kernels that are used include $\kappa(x, y) = e^{-c\|x-y\|_2}$ or $\kappa(x, y) = e^{-c\|x-y\|_1}$. These kernels are non-differentiable at $x = y$; so the above analysis is not applicable.

6 Conclusion

A self-contained derivation of SVMs with kernel is given along with a new condition for evaluating the continuity of the kernel function. It turns out that the Gaussian kernel in some form is an optimal choice with respect to a given local relative Lipschitz constant $\gamma > 0$ on the one hand and the aim to generate a well conditioned kernel matrix on the other hand. For more general kernels it is demonstrated that a normalization of the diagonal of the kernel may improve both, the condition number of the kernel and the local relative Lipschitz constant.

7 Appendix

To keep this work self-contained, the proofs of two well-known results are repeated below, followed by an elementary proof of Theorem 1.

7.1 Positive kernels

Note that the Hadamard product of two positive semidefinite $n \times n$ matrices again is positive semidefinite, because from the decompositions $A = \sum_i a^{(i)}(a^{(i)})^T \succeq 0$ and $B = \sum_j b^{(j)}(b^{(j)})^T \succeq 0$, it follows that

$$A \circ B = \sum_{i,j} (a^{(i)} \circ b^{(j)})(a^{(i)} \circ b^{(j)})^T \succeq 0.$$

By construction, the matrix $(x^{(1)} \dots x^{(m)})^T (x^{(1)} \dots x^{(m)})$ with entries $(x^{(i)})^T x^{(j)}$ at positions (i,j) is a positive semidefinite Gram matrix.

Thus, Hadamard products of the above Gram matrices are positive semidefinite, and consequently, so is the exponential function as a sum of such products, i.e., the matrix with entries $e^{c(x^{(i)})^T x^{(j)}}$ with $c > 0$ is positive semidefinite. This shows that K with

$$K_{i,j} := e^{-c\|x^{(i)} - x^{(j)}\|^2} = e^{2c(x^{(i)})^T x^{(j)} - c\|x^{(i)}\|^2 - c\|x^{(j)}\|^2}$$

is positive semidefinite, as the terms $e^{-c\|x^{(i)}\|^2}$ and $e^{-c\|x^{(j)}\|^2}$ only cause a symmetric diagonal scaling of the matrix K, i.e., a change from $K \succeq 0$ to $DKD \succeq 0$ with a diagonal matrix D.

Analogously, it follows that $\kappa(x,y) \equiv ((x^T y)+1)^p \equiv \sum_{j=0}^{p} \binom{p}{j} (x^T y)^j$ with $p \in \mathbb{N}$ is a positive definite kernel.

Also, if $a \in \mathbb{R}$, $a > 0$ is given and $g : [-a, a] \to \mathbb{R}$ is a power series with nonnegative coefficients, then for Ω with $\sup_{x \in \Omega} \{\|x\|_2^2\} \leq a$ as above, it follows that the kernel $\kappa(x,y) \equiv g(x^T y)$ from Section 5.1 is a positive definite kernel. And since $K \succeq 0$ implies $DKD \succeq 0$ for any diagonal matrix D, it follows that the kernel $\kappa(x,y) \equiv h(x)h(y)g(x^T y)$ from Section 5.1 is a positive definite kernel as well.

Finally, if $g(x^T x) > 0$ for all $x \in \Omega$, then κ with the choice $h(x) := g(x^T x)^{-1/2}$ is an iso-normalized kernel (as defined in Section 4.1.4).

7.2 Exact separability with Gaussian kernels

To prove that the functions $x \mapsto e^{-c\|x - x^i\|_2^2}$ are always linearly independent for pairwise distinct x^i (with $1 \leq i \leq m$), consider the question to determine a vector $\alpha \in \mathbb{R}^m$ such that

$$0 \equiv \sum_{i=1}^{m} \alpha_i e^{-c\|x - x^i\|_2^2} = \sum_i \alpha_i e^{-c\|x\|_2^2} e^{-c\|x^i\|_2^2} e^{2cx^T x^i} = e^{-c\|x\|_2^2} \sum_i \alpha_i e^{-c\|x^i\|_2^2} e^{2cx^T x^i} \quad \forall x \in \mathbb{R}^n.$$

Setting $\beta_i := \alpha_i e^{-c\|x^i\|_2^2}$, this system is equivalent to

$$0 = \sum_i \beta_i e^{2cx^T x^i} \quad \forall x \in \mathbb{R}^n.$$

(In particular, $\beta_i = 0 \Longleftrightarrow \alpha_i = 0$.) Now choose a vector $\bar{x}$ such that $\bar{x}^T(x^i - x^j) \neq 0$ for all $i \neq j$, and set $x := j\bar{x}$ for $0 \leq j \leq m-1$ in the above system. It follows that

$$0 = \sum_i \beta_i e^{2cj\bar{x}^T x^i} = \sum_i \beta_i (a_i)^j \quad \text{for } 0 \leq j \leq m-1$$

with $a_i := e^{2c\bar{x}^T x^i}$. By choice of $\bar{x}$, the numbers a_i are pairwise distinct. The above system has the only solution $\beta = 0$ if the transposed system

$$0 = \sum_j \tilde{\beta}_j (a_i)^j$$

also has only the solution $\tilde{\beta} = 0$. This latter system implies that the polynomial $t \mapsto \sum \tilde{\beta}_j t^j$ interpolates the zero function at all points a_i. Thus, $\tilde{\beta} = 0$ must hold. (By uniqueness of polynomial interpolation.) As derived above, this implies the desired conclusion $\alpha = 0$.

7.3 Proof of Theorem 1

Assume that the relative Lipschitz condition (30) is satisfied. Rearranging and squaring shows that (30) is equivalent to

$$\frac{\|\phi(x) - \phi(y)\|_{\mathcal{W}}^2}{\|x - y\|_2^2} \leq \gamma^2 \ \|\phi(x)\|_{\mathcal{W}}^2 \qquad \text{for small } \|x - y\|_2. \tag{38}$$

The expression on the left resembles a finite-difference approximation of a second derivative, and as shown below, the bound (38) is indeed equivalent to a bound on the "mixed" part of the second derivative of κ as a function $\mathbb{R}^{2n} \to \mathbb{R}$.

The numerator of the left-hand side of (38) is

$$\|\phi(x) - \phi(y)\|_{\mathcal{W}}^2 = \langle K_x - K_y, K_x - K_y \rangle_{\mathcal{W}} = \kappa(x,x) + \kappa(y,y) - 2\kappa(x,y). \tag{39}$$

Furthermore, assume that κ is three times continuously differentiable. Below, the notation z for the variable pair x, y, i.e., $z := \begin{pmatrix} x \\ y \end{pmatrix}$, and the vectors $\mathbf{m} := \frac{x+y}{2}$ and $\mathbf{d} := \frac{x-y}{2}$ are used. The Landau symbol $a = o(|b|)$ indicates that a depends on b and the limit $\lim_{b \to 0,\ b \neq 0} a/b = 0$. Similarly, if $\limsup_{b \to 0,\ b \neq 0} a/b < \infty$ we write $a = O(|b|)$. With these notations, the Taylor expansion yields

$$\kappa(x,x) = \kappa(\mathbf{m},\mathbf{m}) + D_z\kappa(\mathbf{m},\mathbf{m}) \left[\begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix} \right] + \frac{1}{2} D_{z,z}^2 \kappa(\mathbf{m},\mathbf{m}) \left[\begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix}, \begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix} \right] + o(\|\mathbf{d}\|^2),$$

$$\kappa(y,y) = \kappa(\mathbf{m},\mathbf{m}) - D_z\kappa(\mathbf{m},\mathbf{m}) \left[\begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix} \right] + \frac{1}{2} D_{z,z}^2 \kappa(\mathbf{m},\mathbf{m}) \left[\begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix}, \begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix} \right] + o(\|\mathbf{d}\|^2).$$

Adding these equations yields

$$\kappa(x,x) + \kappa(y,y) = 2\kappa(\mathbf{m},\mathbf{m}) + D_{z,z}^2 \kappa(\mathbf{m},\mathbf{m}) \left[\begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix}, \begin{pmatrix} \mathbf{d} \\ \mathbf{d} \end{pmatrix} \right] + o(\|\mathbf{d}\|^2).$$

Similarly, from the expansions of $\kappa(x,y) = \kappa(y,x)$ around the point $(\mathbf{m},\mathbf{m})$,

$$\kappa(x,y) + \kappa(y,x) = 2\kappa(\mathbf{m},\mathbf{m}) + D^2_{z,z}\kappa(\mathbf{m},\mathbf{m})\left[\begin{pmatrix}\mathbf{d}\\-\mathbf{d}\end{pmatrix},\begin{pmatrix}\mathbf{d}\\-\mathbf{d}\end{pmatrix}\right] + o(\|\mathbf{d}\|^2).$$

The last two equations imply

$$\begin{aligned}&\kappa(x,x) + \kappa(y,y) - 2\kappa(x,y)\\ = \;& D^2_{z,z}\kappa(\mathbf{m},\mathbf{m})\left[\begin{pmatrix}\mathbf{d}\\\mathbf{d}\end{pmatrix}\begin{pmatrix}\mathbf{d}\\\mathbf{d}\end{pmatrix}\right] - D^2_{z,z}\kappa(\mathbf{m},\mathbf{m})\left[\begin{pmatrix}\mathbf{d}\\-\mathbf{d}\end{pmatrix}\begin{pmatrix}\mathbf{d}\\-\mathbf{d}\end{pmatrix}\right] + o(\|\mathbf{d}\|^2).\end{aligned}$$

The second derivatives with respect to x,x and y,y in the variable vector z cancel each other. Due to symmetry, $D^2_{x,y}\kappa(\mathbf{m},\mathbf{m}) = D^2_{y,x}\kappa(\mathbf{m},\mathbf{m})$. Thus, only the mixed terms above remain,

$$\kappa(x,x) + \kappa(y,y) - 2\kappa(x,y) = 4D^2_{x,y}\kappa(\mathbf{m},\mathbf{m})[\mathbf{d},\mathbf{d}] + o(\|\mathbf{d}\|^2)$$

or

$$\kappa(x,x) + \kappa(y,y) - 2\kappa(x,y) = D^2_{x,y}\kappa(\mathbf{m},\mathbf{m})[x-y,x-y] + o(\|x-y\|^2). \tag{40}$$

Observe that the left hand side in (40) coincides with (39). Thus, dividing (40) by $\|x-y\|_2^2$ yields

$$\frac{\|\phi(x)-\phi(y)\|^2_{\mathcal{W}}}{\|x-y\|_2^2} = D^2_{x,y}\kappa(\mathbf{m},\mathbf{m})[\tfrac{x-y}{\|x-y\|_2},\tfrac{x-y}{\|x-y\|_2}] + o(1)$$

The left hand side is positive, and by (38) it is at most $\gamma^2\kappa(z,z)$. Considering the limit $y \to x$ and using that $D_x(D_y\kappa(x,y))|_{x=y=z}$ is symmetric it follows that it is positive semidefinite and

$$\|D_x(D_y\kappa(x,y))|_{x=y=z}\|_2 \le \gamma^2\kappa(z,z)$$

which is equivalent to (30).

To justify the converse direction, i.e. that (31) implies (30) observe that in the derivation of (31), only equations were used, as well as the limit $y \to x$, i.e., (31) also implies the infinitesimal version of (30) or (38), i.e., (31) implies

$$\lim_{y\to x}\frac{\|\phi(x)-\phi(y)\|^2_{\mathcal{W}}}{\|x-y\|_2^2} \le \gamma^2\|\phi(x)\|^2_{\mathcal{W}}$$

For the above proof, the squares of the equivalent inequalities

$$\lim_{y\to x}\frac{\|\phi(x)-\phi(y)\|_{\mathcal{W}}}{\|x-y\|_2} \le \gamma\ \|\phi(x)\|_{\mathcal{W}} \tag{41}$$

were considered above. Since the third derivative of κ is continuous on the compact set $\Omega\times\Omega$ it is bounded (as a trilinear form with respect to the 2-norm) by a constant $\overline{\omega} < \infty$, implying that the terms $o(\|\mathbf{d}\|^2)$ in the above estimates can be replaced with $\overline{\omega}\|\mathbf{d}\|^3$.

For fixed $x,y \in \Omega$ $(x \neq y)$ and $t \in [0,1]$, let's first consider the function

$$l : t \mapsto \frac{\|\phi(x+t(y-x))\|_{\mathcal{W}}}{\|x-y\|_2}.$$

Here, we utilize the assumption that Ω is convex. By assumption, $\phi(x) \neq 0$ in Ω, and thus $l(t) \neq 0$ for $t \in [0,1]$. Thus, the function l is differentiable. Let $\bar{t} \in [0,1)$ and define the point

$x(\bar{t}) := x + \bar{t}(y-x)$. The derivative of l at $t = \bar{t}$ coincides with the following one-sided limit:

$$\begin{aligned} l'(\bar{t}) &= \lim_{t \downarrow 0} \frac{l(\bar{t}+t) - l(\bar{t})}{t} \\ \underbrace{=}_{\text{since } t>0} & \lim_{t \downarrow 0} \frac{\|\phi(x + (\bar{t}+t)(y-x))\|_{\mathcal{W}} - \|\phi(x+\bar{t}(y-x))\|_{\mathcal{W}}}{\|t(x-y)\|_2} \\ = & \lim_{t \downarrow 0} \frac{\|\phi(x(\bar{t}) + t(y-x))\|_{\mathcal{W}} - \|\phi(x(\bar{t}))\|_{\mathcal{W}}}{\|t(x-y)\|_2} \\ \underbrace{\leq}_{(41)} & \gamma \|\phi(x(\bar{t}))\|_{\mathcal{W}} = \gamma \|x-y\|_2 l(\bar{t}). \end{aligned}$$

Thus, the differential inequality $l'(t) \leq \gamma \|x-y\|_2 l(t)$ holds for $t \in [0,1)$. The corresponding differential equation $u'(t) = \gamma \|x-y\|_2 u(t)$ with $u(0) = l(0)$ has the solution $u(t) \equiv l(0) e^{\gamma \|x-y\|_2 t}$. By Gronwall's inequality, it follows that

$$l(t) \leq u(t) = l(0) e^{\gamma \|x-y\|_2 t}$$

for $t \in [0,1]$. For $t = 1$ it therefore follows that

$$\|\phi(y))\|_{\mathcal{W}} \leq e^{\gamma \|x-y\|_2} \|\phi(x)\|_{\mathcal{W}}. \tag{42}$$

Furthermore, according to (40), assumption (31), and the local Lipschitz continuity of κ,

$$\begin{aligned} & \|\phi(x) - \phi(y)\|_{\mathcal{W}}^2 \\ = \; & \kappa(x,x) + \kappa(y,y) - 2\kappa(x,y) = D_{x,y}^2 \kappa(\mathbf{m},\mathbf{m})[x-y, x-y] + O(\|\mathbf{d}\|^3) \\ & \leq \gamma^2 \kappa(\mathbf{m},\mathbf{m}) \|x-y\|^2 + O(\|\mathbf{d}\|^3) = \gamma^2 \|\phi(\mathbf{m})\|_{\mathcal{W}}^2 \|x-y\|^2 + O(\|\mathbf{d}\|^3) \end{aligned} \tag{43}$$

This implies

$$\begin{aligned} \|\phi(x) - \phi(y)\|_{\mathcal{W}} &\leq \sqrt{\gamma^2 \kappa(x,x) \|x-y\|^2 + O(\|\mathbf{d}\|^3)} \\ &= \gamma \sqrt{\kappa(x,x)} \|x-y\| + O(\tfrac{1}{2\sqrt{\xi}} \|\mathbf{d}\|^3) \end{aligned}$$

with $\xi \geq \gamma^2 \kappa(x,x) \|x-y\|^2 = 4\gamma^2 \kappa(x,x) \|\mathbf{d}\|^2$ (mean value theorem). Therefore,

$$\|\phi(x) - \phi(y)\|_{\mathcal{W}} \leq \gamma \sqrt{\kappa(x,x)} \|x-y\| + O(\|\mathbf{d}\|^2) = \gamma \|\phi(x)\|_{\mathcal{W}} \|x-y\| + O(\|\mathbf{d}\|^2). \tag{44}$$

Now, define

$$g(t) := \|\phi(x(t)) - \phi(x)\|_{\mathcal{W}}, \qquad \text{where } x(t) := x + t(y-x).$$

Then g is differentiable as long as $g(t) \neq 0$, and using the fact that $\|a\| - \|b\| \leq \|a-b\|$, and exploiting (44), we have

$$\begin{aligned} g'(t) &= \lim_{\Delta t \downarrow 0} \frac{\|\phi(x(t+\Delta t)) - \phi(x)\|_{\mathcal{W}} - \|\phi(x(t)) - \phi(x)\|_{\mathcal{W}}}{\Delta t} \\ &\leq \lim_{\Delta t \downarrow 0} \frac{\|\phi(x(t+\Delta t)) - \phi(x(t))\|_{\mathcal{W}}}{\Delta t} \\ &\leq \lim_{\Delta t \downarrow 0} \frac{\gamma \|\phi(x(t))\|_{\mathcal{W}} \Delta t \|x-y\| + O(\|\Delta t \; \mathbf{d}\|^2)}{\Delta t} \\ &= \gamma \|\phi(x(t))\|_{\mathcal{W}} \|x-y\|. \end{aligned}$$

The above inequality holds also at $t = 0$ for the right-hand side derivative of g. Substituting $x(t)$ in place of y into (42), we get from the above inequality

$$g'(t) \leq \gamma e^{\gamma\|x-x(t)\|_2}\|\phi(x)\|_{\mathcal{W}}\ \|x-y\| = \gamma\|\phi(x)\|_{\mathcal{W}}\ \|x-y\|e^{\gamma\|x-y\|_2 t}.$$

With $g(0) = 0$, it follows

$$g(t) \leq \int_0^t \gamma\|\phi(x)\|_{\mathcal{W}}\ \|x-y\|e^{\gamma\|x-y\|_2 t} = \|\phi(x)\|_{\mathcal{W}}(e^{\gamma\|x-y\|_2 t} - 1)$$

and for $t = 1$ we obtain

$$\|\phi(y) - \phi(x)\|_{\mathcal{W}} \leq \|\phi(x)\|_{\mathcal{W}}(e^{\gamma\|x-y\|_2} - 1)$$

with

$$e^{\gamma\|x-y\|_2} - 1 \approx \gamma\|x-y\|_2 \qquad \text{for small } \|x-y\|_2. \qquad \#$$

7.4 Acknowledgment

The author wishes to thank Achim Schädle and Melinda Hagedorn for helpful discussions and corrections. He is also indebted to Nikolai Jarre for the English translation of this text that was initially written in German. The thoughtful comments of an anonymous referee and Varvara Turova helped to improve the presentation of the paper.

References

[1] J. Cervantes, F. Garcia-Lamont, and L. Rodríguez-Mazahua. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[2] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.

[3] P-H. Chen, C-J. Lin, and B. Schölkopf. A tutorial on ν-support vector machines. *Appl. Stochastic Models Bus. Ind.*, 21:111–136, 2005.

[4] T. Glasmachers. Recipe for fast large-scale svm training: Polishing, parallelism, and more ram! In T. Calders, C. Vens, J. Lijffijt, and B. Goethals, editors, *Artificial Intelligence and Machine Learning*, volume 1805 of *Communications in Computer and Information Science, CCIS*. Springer Verlag, 2023.

[5] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. JHU Press, 2013.

[6] T. Hofmann, B. Schölkopf, and A.J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

[7] Y-J. Lee and O.L. Mangasarian. Rsvm: Reduced support vector machines. In V. Kumar and R.L. Grossman, editors, *Proceedings of the SIAM International Conference on Data Mining, Chicago, April 5-7*, pages 1–17. SIAM, Philadelphia, 2001. doi: 10.1137/1.9781611972719.13.

[8] O.L. Mangasarian and D.R. Musicant. Lagrangian support vector machines. *Journal of Machine Learning Research*, 1:161–177, 2001.

[9] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Studies in Applied and Numerical Mathematics. SIAM Philadelphia, PA, 1994.

[10] W.S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

[11] C. O'Neil. *Weapons of Math Destruction*. Crown Books, 2016.

[12] D.A. Pisner and D.M. Schnyer. Support vector machine. In A. Mechelli and S. Vieira, editors, *Machine Learning. Methods and Applications to Brain Disorders*, pages 101–121. Academic Press, Elsevier, 2020. doi: 10.1016/B978-0-12-815739-8.00006-7.

[13] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, 2001.

[14] A. Shapiro. Upper bounds for nearly optimal diagonal scaling of matrices. *Linear and Multilinear Algebra*, 29:145–147, 1991.

[15] A. Van Der Sluis. Condition numbers and equilibration of matrices. *Numer. Math.*, 14:14–23, 1969.

[16] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Texts in Applied Mathematics. Springer New York, NY, 2002.

[17] J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions I*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 1969.

[18] S. Suthaharan. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification*, volume 36 of *Integrated Series in Information Systems*, pages 207–235. Springer, Boston, 2016. doi: 10.1007/978-1-4899-7641-3_9.

[19] V.N. Vapnik and A.Y. Chervonenkis. The necessary and suffcient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.