

## *Learning with Infinitely Many Kernels via Semi-Infinite Programming*

S. Özögür-Akyüz<sup>a</sup> \* and G.-W. Weber <sup>a, b</sup>

<sup>a</sup>*Institute of Applied Mathematics, Middle East Technical University, METU, Ankara,  
Turkey*

<sup>b</sup>*Faculty of Economics, Management Science and Law, University of Siegen, Germany*

(v4.2 released May 2008)

In recent years, learning methods are desirable because of their reliability and efficiency in real-world problems. *Machine Learning (ML)* is one of the powerful subfields of *Artificial Intelligence (AI)* which deals with finding hidden patterns in data or classifying objects in the data. In this study, we are interested in *Support Vector Machines (SVMs)* which is one of the powerful methods in classification methods in supervised learning. In ML algorithms, one of the crucial issues is the representation of the data. The linearity of the different classes plays an important role in ML. If the data is not linearly separable, a *kernel function* transforms the nonlinear data into a higher-dimensional space in which the nonlinear data are linearly separable. As the data become heterogeneous and large-scale, single kernel methods become insufficient to classify nonlinear data. Convex combinations of kernels are developed to classify this kind of data [5]. Nevertheless, the finite combination of kernels are limited up to a finite choice. In order to overcome this discrepancy, we propose a novel method of "infinite" kernel combinations for learning problems with the help of infinite and semi-infinite optimization regarding all elements in kernel space. This will provide to study variations of combinations of kernels when considering heterogeneous data in real-world applications. Combination of kernels can be done, e.g., along a homotopy parameter or a more specific parameter. Looking at all infinitesimally fine convex combinations of the kernels from the infinite kernel set, the margin is maximized subject to an infinite number of constraints with a compact index set and an additional (Riemann-Stieltjes) integral constraint due to the combinations. After a parametrisation in the space of probability measures it becomes semi-infinite. We analyze the regularity conditions which satisfy the Reduction Ansatz and discuss the type of distribution functions within the structure of the constraints and our bilevel optimization problem.

**Keywords:** machine Learning; semi-infinite optimization; infinite programming; support vector machines; continuous optimization; data mining.

**AMS Subject Classification:** 90C34; 90C06; 68T01

### 1. Introduction

By the innovation and development of the technology, high processor computers took the place of human workload. For instance, the classification or detection problems in the real world such as credit card frauding, account management, portfolio optimization, or in life sciences such as biological experiments [16], prediction of cancer risk, finding pattern in genes [12] or identification of proteins without needing any costing experiments. This provides to save time in industry. By mathematical modelling and computer science, experimental data are analyzed and data mining and learning tools [28] are developed. As the demands increase, new constraints are added, risks are minimized, etc.. Thus, it turns out to be an

---

\*Corresponding author. Email: sozogur@metu.edu.tr

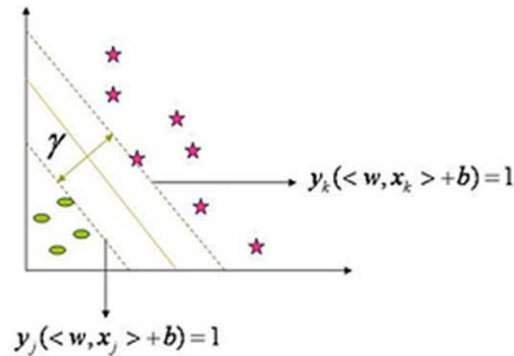


Figure 1. Maximum margin between two classes [7].

optimization problem in which data mining tools are used. The contribution of continuous optimization methods opens a new field of research.

In this study, we will focus on optimization methods for solving *binary* classification problems. As a tool for classification problems, *support vector machines (SVMs)* will be used; they are one of the most efficient classification tools and base on maximizing the margin  $\gamma$  between two classes of objects with some constraints. The two classes are separated by an affine function via  $\langle w, x \rangle + b = 0$ , where  $w$  is a normal vector of the hyperplane and  $\langle \cdot, \cdot \rangle$  denotes scalar product [7]. Given a set of data (examples)  $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ , where  $y_i \in \{\pm 1\}$  and  $x_i \in \mathbb{R}^n$ , groups of points are separated by a hyperplane as shown in Figure 1. From these data points, the so called *training set*, a classifier function or, i.e., hyperplane, is found by SVM in order to predict the class of unlabeled data points, which are unknown data points.

Linearity of the classes of data is one of the essential issues in SVM theory since a hyperplane is a tool to discriminate the classes which is linear itself. The pattern of the data can be discretely nonconvex or some part of the data can be belong to one class or group of data, surrounded by the data of the another class. In most of the real-world problems [16], data are not linearly separable. Thus the data need to be transformed into another space in which they become linearly separable. The representation of nonlinear data is changed with a nonlinear mapping  $\phi$  which transforms the input space into a higher dimensional feature space such that the data points are linearly separable. But the mapping can be very high dimensional and sometimes infinite dimensional. Hence, it is hard to interpret decision (classification) functions which are expressed as  $f(x) = \langle w, \phi(x) \rangle + b$ . In [7], a kernel function is defined as an inner product of two points under the mapping  $\phi$ , i.e.,  $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ , so that the nonlinear mapping is implicitly defined in an inner product which does not need to be found explicitly. The kernel function physically denotes the similarity between the points. Hence, the distribution of the data can determine the pattern of the data, thus, the similarity and kernel function, e.g., Gaussian bellshaped function. The optimization problem for separating two

classes is expressed as follows [7]:

$$\begin{array}{ll}
 \min_{w,b} & \langle w, w \rangle \\
 \text{subject to} & y_i \cdot (\langle w, \phi(x_i) \rangle + b) \geq 1 \\
 \text{Primal Hard} & \\
 \text{Margin Problem} & (i = 1, 2, \dots, l);
 \end{array} \quad (1)$$

its dual problem reads

$$\begin{array}{ll}
 \max_{\alpha} & \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \\
 \text{subject to} & \sum_{i=1}^l \alpha_i y_i = 0, \\
 \text{Dual Hard} & \\
 \text{Margin Problem} & \alpha_i \geq 0 \quad (i = 1, 2, \dots, l).
 \end{array} \quad (2)$$

Note that the primal problem is a convex quadratic problem. Concerning the dual problem, the requirement that the kernel satisfies Mercer's condition [7] is equivalent to the requirement that the kernel matrix  $(\kappa(x_i, x_j))_{i,j=1}^l$  is positive definite for all training points. This implies that the objective function of (2) is strictly concave since the matrix  $(y_i y_j \kappa(x_i, x_j))_{i,j=1}^l$  is positive definite. Since  $\mathcal{K}$  is positive semi-definite, some matrix  $Y^T \mathcal{K} Y$  is, of course, positive (semi) definite, too. In our case,  $Y^T = Y$  is a diagonal matrix with the numbers  $y_i = \{\pm\}$  at the diagonal.

Then, any local solution of primal and dual problem is also a global one. By the strong duality theorem [7], which says that if the domain is convex and the constraints are affine, then there is no duality gap between (1) and (2).

It is not satisfactory to apply strictly perfect maximal margin classifiers without any error term, since they will not be applicable to noisy real data. Therefore, allowing that a *maximal margin criterion* which states that the two classes are separated such that the distance between them are maximum, is violated. This is mathematically written in (3), with a vector  $\xi$  of some slack variables introduced into the constraints and, equipped with an error constant  $C \geq 0$ . Here,  $\xi = (\xi_1, \dots, \xi_l)^T$  measures the degree of misclassification and the constant  $C$  controls the misclassification. In the equation (3), some points are allowed to be in maximum margin with a distance of  $\xi_i > 0$  and an error constant  $C$ . Obviously, if  $\xi_i > 1$ , the point  $x_i$  is missclassified and if  $0 < \xi_i < 1$ , the point  $x_i$  is correctly classified, as shown by Figure 2.

$$\begin{array}{ll}
 \min_{w,b} & \langle w, w \rangle + C \sum_{i=1}^l \xi_i \\
 \text{subject to} & y_i \cdot (\langle w, \phi(x_i) \rangle + b) \geq 1 - \xi_i \\
 \text{Primal Soft} & \\
 \text{Margin Problem} & (i = 1, 2, \dots, l).
 \end{array} \quad (3)$$

The dual problem in the soft margin case looks as follows [7]:

$$\begin{array}{ll}
 \max_{\alpha} & \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \\
 \text{subject to} & \sum_{i=1}^l \alpha_i y_i = 0, \\
 \text{Dual Soft} & \\
 \text{Margin Problem} & 0 \leq \alpha_i \leq C \quad (i = 1, 2, \dots, l).
 \end{array} \quad (4)$$

Real-world data can be supplied from heterogeneous kinds of sources. In such cases, multiple kernels are more convenient to use for a good accuracy. Recent applications [13] showed the need for *multiple kernel learning (MKL)* by its interpretability and efficiency. The common approach to MKL is a convex combination of several kernels. Those kernels were selected before and combined to serve well

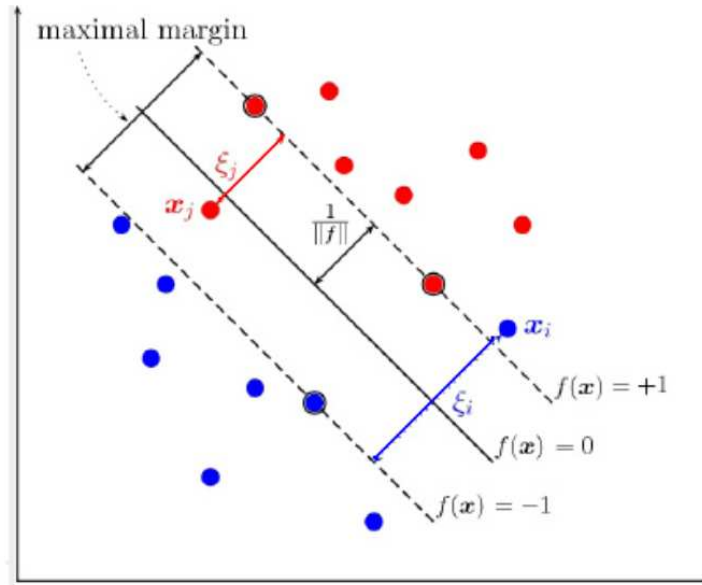


Figure 2. Introduction of slack variables  $\xi_i$ 's.

for the embedding into the feature space to do linear separation there. In [5], the kernel-based SVM is formulated by a combination of multiple kernels and solved by *quadratically-constrained quadratic programming (QCQP)* which is applied to solve dual conic optimization problem. Likewise, [21] uses adapted multiple kernel learning to large-scale problems which applies the method to biological sequence analysis. Since the biological sequences have different motifs inside and for each subsequence, different types of kernels are used, the combination is taken over the whole sequence. In [21], kernel coefficients are maximized beyond a minimization with respect to the dual variables, which is a max-min type of a problem. It can be canonically represented as a semi-infinite problem [26, 27]. The classical SVM is solved iteratively with linear programming and increasing the number of constraints iteratively in [21]. A different form of objective function is proposed in [15] for MKL by adapted weighted 2-norm regularization for each function  $f$  induced by kernels  $k_\kappa$  ( $\kappa = 1, 2, \dots, K$ ) instead of using the 1-norm block regularization [21] ( $K$  denoting some finite number of kernels). Sparsity of linear combinations of kernels is controlled by adding a 1-norm regularization term on these kernel weights.

**Note on Numerical Aspects:** In our previous studies [16, 17], data are classified regarding the margin of the test data points and using all classifiers in the hypothesis set. Thus, this benefits from the information of all classifiers and also from the various kernels by different kernel parameters, e.g., in Gaussian kernel, kernel parameter is a Gaussian width. Hence, using different classifiers in different ways, for example, by voting, by ensemble classifiers, gives comparable accuracy results for each test data point and also it improves the speed [16, 17]. In [17], the classification functions depend on only one kernel, but the classification of the new data depend on the results of different combinations of these classifiers on the test points. This improves the accuracy and the speed of the algorithm in the numerical results.

The finite combinations of kernels are limited up to a finite choice. This limitation does not always allow to represent the similarity or dissimilarity of data points, specifically highly nonlinear and large-scaled ones. A finite combination

may fail, here. In order to overcome this, with the motivation of previous studies [16, 17], we propose a combination of infinitely many kernels in Riemann-Stieltjes integral form for binary classification to allow an infinite wealth of possible choices of kernels in the kernel space. This makes the problem infinite in both its dimension and number of constraints, which is so-called *infinite programming (IP)*. Our IP problem formulation consists in the limiting case of infinite kernel coefficients  $\beta_\kappa$ 's where  $\kappa \rightarrow \infty$ , which is defined as a monotonically increasing function or a probability measure,  $\beta$ , and an infinite number of constraints coming from maximal margin principle of SVM. Allowing infinitely many kernels might make our problem ill-posed for real-world problems because of the enormous complexity of the model also called *overfitting*. To penalize this curse of dimensionality, we introduce the regularization terms and approximate "differentiability" in the penalizing term by first- and second-order difference quotients. On the other hand, to solve IP more tractably, we reduced the IP to a semi-infinite problem, by parametrizing infinite variables (measures) by probability density functions (pdfs). We will illustrate parametrization with examples for pdf. The organization of this paper is as follows: In Section 2, we will motivate our approach by giving a brief introduction to MKL. In Section 3, we will introduce our approach so called *infinite kernel learning (IKL)* and we will find regularity conditions for reduction ansatz for both primal and dual problem. To define regularity conditions for the reduction ansatz, the neighbourhood of optimal solution needs to be defined since optimal points are implicitly depending on measures. Thus, we will discuss the topology of parameters of the lower level problem, which are defined as measures in IP. In Section 4, examples of different parametrizations will be given to reduce the problem into IP. In Section 5, regularization of infiniteness will be discussed by means of adding a term which penalizes infiniteness in the model. Finally, in Section 6, a conclusion and an outlook of future studies will be given.

## 2. Multiple Kernel Learning

In this section, we will give an intuition of MKL and problem formulations. Heterogeneous kinds of data in real-world examples have let kernel learning algorithms become generalized by the combination of kernels in a compact form [21]. A weighted combination of kernels allows to define similarity measurement of heterogeneous data. Firstly, we regard a convex combination of kernels  $k_\kappa$  ( $\kappa = 1, \dots, K$ ):

$$k_\beta(x_i, x_j) = \sum_{\kappa=1}^K \beta_\kappa k_\kappa(x_i, x_j), \quad (5)$$

where  $\beta_\kappa \geq 0$  ( $\kappa = 1, 2, \dots, K$ ),  $\sum_{\kappa=1}^K \beta_\kappa = 1$ ,  $x_i$  ( $i = 1, 2, \dots, l$ ) is translated via  $K$  mappings  $\phi_\kappa : x_i \mapsto \phi_\kappa(x_i) \in \mathbb{R}^{D_\kappa}$  ( $\kappa = 1, \dots, K$ ), from the input space  $\mathbb{R}^n$  into  $K$  feature spaces  $\mathbb{R}^{D_\kappa}$ ,  $D_\kappa$  being the dimension of the  $k$ -th feature space [21] and  $k_\kappa(x_i, x_j) = \langle \phi_\kappa(x_i), \phi_\kappa(x_j) \rangle$ .

In [21], the following MKL problem is derived by using the convex combination of kernels (5):

$$\begin{aligned} \text{Primal Multiple} \quad & \min \frac{1}{2} \left( \sum_{\kappa=1}^K \|w_\kappa\|_2 \right)^2 + C \sum_{i=1}^l \xi_i \quad (w_\kappa \in \mathbb{R}^{D_\kappa}, \xi \in \mathbb{R}^l, b \in \mathbb{R}) \\ \text{Kernel Problem} \quad & \text{subject to } y_i \cdot \left( \sum_{\kappa=1}^K \langle w_\kappa, \phi_\kappa(x_i) \rangle + b \right) \geq 1 - \xi_i, \\ & \xi \geq 0 \quad (i = 1, 2, \dots, l). \end{aligned} \quad (6)$$

In [5], the dual of the problem (6) is expressed with second-order cones as follows:

$$\begin{aligned}
\text{Dual Multiple} & \quad \min \quad \frac{1}{2}\gamma^2 - \sum_{i=1}^l \alpha_i \quad (\gamma \in \mathbb{R}, \alpha \in \mathbb{R}^l) \\
\text{Kernel Problem} & \quad \text{subject to} \quad 0 \leq \alpha_i \leq C, \sum_{i=1}^l \alpha_i y_i = 0, \\
& \quad \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j k_\kappa(x_i, x_j) \leq \gamma \quad (\kappa = 1, 2, \dots, K).
\end{aligned} \tag{7}$$

A numerical solution for large-scale problems is introduced in [21] by using a *semi-infinite linear programming (SILP)* [10] at the place of (7) rather than solving an SDP (semidefinite programming) problem as done in [5]:

$$\begin{aligned}
\max_{\beta} \min_{\alpha} & \quad \sum_{\kappa=1}^K \beta_\kappa S_\kappa(\alpha) \quad (\alpha \in \mathbb{R}^l, \beta \in \mathbb{R}^K) \\
\text{subject to} & \quad 0 \leq \alpha_i \leq C, \beta \geq 0 \quad (\text{componentwise}), \\
& \quad \sum_{i=1}^l \alpha_i y_i = 0, \text{ and } \sum_{k=1}^K \beta_k = 1,
\end{aligned} \tag{8}$$

where  $S_k(\alpha) := \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) - \sum_{i=1}^l \alpha_i$ . Let us denote  $S(\alpha, \beta) := \sum_{\kappa=1}^K \beta_\kappa S_\kappa(\alpha)$ . Problem (8) can be represented as an *SIP (semi-infinite programming)* problem by a standard argument [21]. Indeed, by maximizing the height variable  $\theta$  under the min term (epigraph argument), problem (8) reduces to the following smooth maximization problem of SILP kind:

$$\begin{aligned}
\max \theta & \quad (\theta \in \mathbb{R}, \beta \in \mathbb{R}^K) \\
\text{subject to} & \quad \beta \geq 0, \sum_{\kappa} \beta_\kappa = 1 \\
& \quad \sum_{\kappa=1}^K \beta_\kappa S_\kappa(\alpha) \geq \theta \quad \forall \alpha \in \mathbb{R}^l \text{ with } 0 \leq \alpha \leq C\mathbf{1} \text{ and } \sum_{i=1}^l y_i \alpha_i = 0.
\end{aligned} \tag{9}$$

Here,  $\mathbf{1} = (1, 1, 1, \dots, 1)^T \in \mathbb{R}^l$ .

### 3. Learning with Infinite Kernels

#### 3.1. Combination of Infinitely Many Kernels

Due to the limited case of multiple kernels as it is discussed in Section 1 and based on the motivation of multiple kernel learning, we propose a different formulation by introducing infinitely many kernels in the Riemann-Stieltjes [2] integral form which covers an infinite dimensional kernel space. Mathematically, an infinite combination will be represented by the following formula:

$$k_\beta(x_i, x_j) := \int_{\Omega} k(x_i, x_j, \omega) d\beta(\omega), \tag{10}$$

where  $\omega \in \Omega$  is a kernel parameter and  $\beta$  is a monotonically increasing function of integral 1, or just a probability measure on  $\Omega$ . For example, infinite combination of Gaussian kernels with different widths from a set  $\Omega$  will be  $\kappa_\beta(x_i, x_j) = \int_{\Omega} \exp(-\omega \|x_i - x_j\|_2^2) d\beta(\omega)$ . Hereby, we use the wealth by means of infinite kernels to overcome the limitation of kernel combination given by finitely pre-chosen kernels. The question of which combination of kernels and the structure of mixture of kernels could be considered and optimized, e.g., be answered by *homotopies*. More formally, let us define a function which provides the combination of kernels as follows:

$$H_{x_i, x_j}(\omega) := k(x_i, x_j, \omega) \quad (\omega \in [0, 1]). \tag{11}$$

In short, we write  $H_{x_i, x_j}(\omega) =: H(\omega)$ , and we illustrate this homotopy by an example.

**Example 1:** Given  $k(x_i, x_j, \omega) = \omega \exp(-w^* \|x_i - x_j\|_2^2) + (1 - \omega)(1 + x_i^T x_j)^d$  and a Gaussian width  $w^*$ , then,

$$H(0) = (1 + x_i^T x_j)^d = k^1(x_i, x_j) \quad (\text{polynomial kernel}),$$

$$H(1) = \exp(-w^* \|x_i - x_j\|_2^2) = k^2(x_i, x_j) \quad (\text{Gaussian kernel}).$$

Herewith,  $\int_{\Omega} k(x_i, x_j, \omega) d\beta(\omega) = k_{\beta}(x_i, x_j)$  with  $\Omega = [0, 1]$ .

The intuition behind the above example is illustrated in Figure 3 and Figure 4. We can go from polynomial to Gaussian via a defined homotopy by infinitesimal coefficients  $d\beta(\omega)$ .

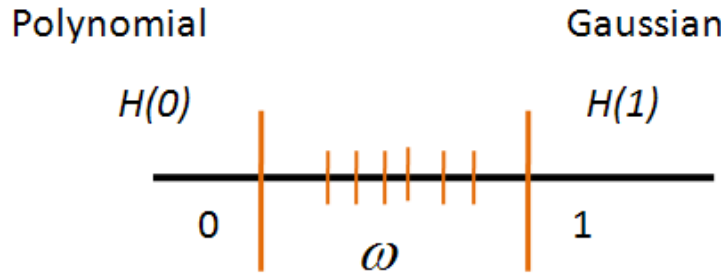


Figure 3. On the homotopy between two kernels, example.

To come to infinitely many and infinitesimal coefficients, let us assume that  $(\eta_{\kappa})_{\kappa \in \mathbb{N}_0}$  is a monotonically increasing sequence in the bounded interval  $\Omega := [0, 1]$  tending to 1 as  $\kappa \rightarrow \infty$  and, say,  $\eta_0 = 0$ . Then  $\sum_{\kappa=1}^{\infty} (\eta_{\kappa} - \eta_{\kappa-1}) = 1$ . We can refine the summation by a Riemann-Stieltjes integral with any monotonically increasing function  $\beta : [0, 1] \rightarrow \mathbb{R}$  such that  $\int_0^1 d\beta(\omega) = 1$ . Indeed, we obtain an infinitesimal increment  $d\beta(\omega)$  after limit calculus with weights  $\beta_{\kappa} = \beta(\omega_{\kappa}) - \beta(\omega_{\kappa-1})$ , i.e., the incremental weights related to a convex combination  $\beta$  of kernels as in (5).

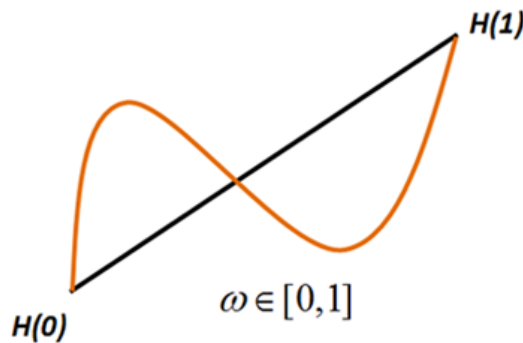


Figure 4. Homotopy function which starts at  $H(0)$  and combines kernels until  $H(1)$  is reached.

Another form of a combination is having just one kernel with its various parameters which are to be in infinite dimensional space. More formally, this can be written in the following form:

**Example 2:** Given a kernel  $k(x_i, x_j, \omega) = \exp(-\omega \|x_i - x_j\|_2^2)$ , the infinite combination of kernel in a Riemann-Stieltjes integral form is

$$\begin{aligned} k_\beta(x_i, x_j, \omega) &= \int_{\Omega} k(x_i, x_j, \omega) d\beta(\omega) \\ &= \int_{\Omega} \exp(-\omega \|x_i - x_j\|_2^2) d\beta(\omega), \end{aligned}$$

where  $\Omega = [a, b]$  ( $0 \leq a < b$ ) is the set in which  $\omega$  lies. Here, we allow different combination of Gaussian widths.

The difference between the first example and second one is that, in the first one, the Gaussian width is fixed and different types of kernels are combined by a homotopy. But, in the second one, the kernel parameter is allowed to be a specific nonlinearly implied variable.

After giving an information about the structure of the combination of infinitely many kernels, we introduce these combinations in the form of Riemann-Stieltjes integrals to the problem (9) as follows:

$$\begin{aligned} \max_{\theta, \beta} \quad & \theta \quad (\theta \in \mathbb{R}, \beta : [a, b] \rightarrow \mathbb{R}, \text{ monotonically increasing function}) \\ \text{subject to} \quad & \int_{\Omega} \left( \frac{1}{2} S(\omega, \alpha) - \sum_{i=1}^l \alpha_i \right) d\beta(\omega) \geq \theta \quad \forall \alpha \in \mathbb{R}^l \text{ with } 0 \leq \alpha \leq C\mathbf{1}, \\ & \sum_{i=1}^l \alpha_i y_i = 0, \quad \int_{\Omega} d\beta(\omega) = 1. \end{aligned} \quad (12)$$

Here,  $S(\omega, \alpha)$  is defined by

$$S(\omega, \alpha) := \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j k(x_i, x_j, \omega). \quad (13)$$

Let us introduce  $T(\omega, \alpha) := S(\omega, \alpha) - \sum_{i=1}^l \alpha_i$ , recall  $\Omega = [0, 1]$  and for the index set of inequality constraints we write

$$A := \{ \alpha \in \mathbb{R}^l \mid 0 \leq \alpha \leq C\mathbf{1} \text{ and } \sum_{i=1}^l \alpha_i y_i = 0 \},$$

where  $\mathbf{1} = (1, 1, \dots, 1)^T$ . Herewith, (12) turns into the following form:

$$\begin{aligned} \max_{\theta, \beta} \quad & \theta \quad (\theta \in \mathbb{R}, \beta : \text{ a positive measure on } \Omega) \\ \text{subject to} \quad & \theta - \int_{\Omega} T(\omega, \alpha) d\beta(\omega) \leq 0 \quad (\alpha \in A). \end{aligned} \quad (14)$$

Since there are infinitely many inequality constraints and the state variable  $\beta$  is from an infinite dimensional space, our problem is a one of *infinite programming (IP)* [1]. Now, we get a dual of (14) as

$$\begin{aligned} \min_{\sigma, \rho} \quad & \sigma \quad (\sigma \in \mathbb{R}, \rho : \text{ a positive measure on } A) \\ \text{subject to} \quad & \sigma - \int_A T(\omega, \alpha) d\rho(\alpha) \leq 0, \quad (\omega \in \Omega), \quad \int_A d\rho(\alpha) = 1. \end{aligned} \quad (15)$$

Let us assume that there exist  $(\beta, \theta)$  and  $(\rho, \sigma)$  that are feasible for their respective problems, and are complementary slack, i.e.,

$$\sigma^* = \int_A T(\omega, \alpha) d\rho^*(\alpha) \quad \text{and} \quad \theta^* = \int_A T(\omega, \alpha) d\beta^*(\omega).$$

Then,  $\beta$  has measure only where  $\sigma = \int_A T(\omega, \alpha) d\rho(\alpha)$  and  $\rho$  has measure only where  $\theta = \int_{\Omega} T(\omega, \alpha) d\beta(\omega)$  which implies that both solutions are optimal for their

respective problems.

The interesting theoretical problem with this is to find conditions which ensure that solutions are point masses (i.e., original monotonic  $\beta$  is a step function). Problem (15) is a linear infinite one, i.e., from *ILP* (*infinite linear programming*), an *SILP* (semi-infinite linear programming) one up to the infinite dimensions of  $\rho$  space. Because of this insight and problem, and in view of the compactness of the feasible (index) sets at the lower levels,  $A$  and  $\Omega$ , we are interested in the nondegeneracy of the local minima of the lower level problem to get finitely many local minimizers [26]. We note that on the lower levels,  $\theta$  and  $\sigma$  are just shift terms which do not affect the local solutions there.

For sake of simplicity from now on, Gaussian kernel combination is used in the form given in Example 2.

### 3.2. Dual Problem

In this section, regularity conditions will be analyzed for the dual problem on its lower level. Let us focus on problem (15), employ the language of bilevel programming known from *SIP* (*semi-infinite programming*), introduce the function  $g((\sigma, \rho), \omega) := \sigma - \int_{\Omega} T(\omega, \alpha) d\rho(\alpha)$ , parametric in  $(\sigma, \rho)$ , and state the

**Lower Level Problem (Dual):** For a given parameter  $(\sigma, \rho)$  we consider

$$\begin{aligned} \min_{\omega} \quad & g((\sigma, \rho), \omega) \\ \text{subject to} \quad & \omega \in \Omega. \end{aligned} \quad (16)$$

Indeed, we denote the defining inequality constraint functions of  $\Omega$  by  $v_1((\sigma, \rho), \omega) := \omega, v_2((\sigma, \rho), \omega) := -\omega + 1$ . We write  $L := \{1, 2\}$ ,  $L_0(\omega) := \{\ell \in L \mid v_{\ell}(\omega) = 0\}$  and briefly denote  $v_{\ell}(\omega) := v_{\ell}((\sigma, \rho), \omega)$  ( $\ell = 1, 2$ ). Consequently, for any critical point  $\bar{\omega}$ , the Lagrange function reads

$$\mathcal{L}^{\mathcal{D}}(\sigma, \rho; \omega, \gamma) := g((\sigma, \rho), \omega) - \sum_{\ell \in L_0(\bar{\omega})} \gamma_{\ell} v_{\ell}(\omega).$$

We briefly write  $\mathcal{L}^{\mathcal{D}}(\omega, \gamma) := \mathcal{L}^{\mathcal{D}}(\sigma, \rho; \omega, \gamma)$ . Since  $\Omega$  is compact, for any  $(\sigma, \rho)$ , local (global) minimizer(s) of (16) exists. We analyze the three conditions in [29], of the *nondegeneracy* of a critical point  $\bar{\omega}$  of the lower level problem which establish the *reduction ansatz* [8]. For any given  $(\sigma, \rho)$  and  $\bar{\omega} \in \Omega$  we note:

- (1) *LICQ*:  $\nabla v_{\ell}(\bar{\omega})$  ( $\ell \in L_0(\bar{\omega})$ ) is a family with not more than one element since an active  $v_{\ell}$  can either be  $\omega$  or  $-\omega + 1$  in the interval  $0 \leq \omega \leq 1$  and  $\nabla v_1(\omega) = 1$  and  $\nabla v_2(\omega) = -1$  do not vanish respectively.
- (2) *Karush Kuhn-Tucker (KKT) condition with strictly positive Lagrange multipliers*: There exists a multiplier  $\bar{\gamma} \in \mathbb{R}^{|L_0(\bar{\omega})|}$  such that  $\nabla_{\omega} \mathcal{L}^{\mathcal{D}}(\bar{\omega}, \bar{\gamma}) = 0$  and  $\bar{\gamma}_{\ell} > 0$  ( $\ell \in L_0(\bar{\omega})$ ). We evaluate this subsequently. If we rewrite  $g((\sigma, \rho), \omega)$ , it will have the following form:

$$\begin{aligned} g((\sigma, \rho), \omega) &= \sigma - \int_A T(\omega, \alpha) d\rho(\alpha) \\ &= \sigma - \sum_{i,j=1}^l k(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) + \int_A \sum_{i=1}^l \alpha_i d\rho(\alpha). \end{aligned}$$

Our Lagrange function is parametric in  $(\sigma, \rho)$  and, fully, it looks as follows:

$$\mathcal{L}^{\mathcal{D}}(\omega, \gamma) = \sigma - \frac{1}{2} \sum_{i,j=1}^l k(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) + \int_A \sum_{i=1}^l \alpha_i d\rho(\alpha) - \sum_{\ell \in L_0(\bar{\omega})} \gamma_{\ell} v_{\ell}(\omega).$$

Let us find the conditions which satisfy the KKT condition with strictly positive Lagrange multipliers, to ensure the nondegeneracy:

$$\nabla_{\omega} \mathcal{L}^{\mathcal{D}}(\omega, \gamma) = \nabla Z - \nabla_{\omega} \left( \sum_{\ell \in L_0(\bar{\omega})} \gamma_{\ell} v_{\ell}(\omega) \right),$$

where, in this case, gradients are reals, and

$$Z := -\frac{1}{2} \sum_{i,j=1}^l k(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha),$$

$$\Rightarrow \nabla Z = -\frac{1}{2} \sum_{i,j=1}^l \nabla_{\omega} k(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha).$$

To closer illustrate this, let us take a Gaussian kernel, i.e.,  $k(x_i, x_j, \omega) = \exp(-\omega \|x_i - x_j\|_2^2)$ , and denote

$$\mathcal{I}(\ell \in L_0(\bar{\omega})) := \begin{cases} 1, & \text{if } \ell \in L_0(\bar{\omega}), \\ 0, & \text{if } \ell \notin L_0(\bar{\omega}), \end{cases}$$

at some critical point  $\bar{\omega}$ . Then, we get

$$\nabla Z = \frac{1}{2} \sum_{i,j=1}^l \|x_i - x_j\|_2^2 \exp(-\omega \|x_i - x_j\|_2^2) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha)$$

and

$$\nabla_{\omega} \left( \sum_{\ell \in L_0(\bar{\omega})} \gamma_{\ell} v_{\ell}(\omega) \right) = \mathcal{I}(1 \in L_0(\bar{\omega})) \cdot \gamma_1 - \mathcal{I}(2 \in L_0(\bar{\omega})) \cdot \gamma_2.$$

Now, we come back to our KKT conditions and evaluate

$$\nabla Z = -\mathcal{I}(1 \in L_0(\bar{\omega})) \cdot \gamma_1 + \mathcal{I}(2 \in L_0(\bar{\omega})) \cdot \gamma_2. \quad (17)$$

There are three cases to be discussed to find strictly positive Lagrange multipliers as given below:

**Case 1:** If  $v_1(\bar{\omega}) = 0$ , i.e.,  $1 \in L_0(\bar{\omega})$ , equation (17) will be

$$\frac{1}{2} \sum_{i,j=1}^l \|x_i - x_j\|_2^2 \exp(-\omega \|x_i - x_j\|_2^2) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) = \gamma_1,$$

$$\gamma_1 > 0 \Leftrightarrow \sum_{i,j=1}^l \|x_i - x_j\|_2^2 \exp(-\omega \|x_i - x_j\|_2^2) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) > 0. \quad (18)$$

**Case 2:** If  $v_2(\bar{\omega}) = 0$ , i.e.,  $2 \in L_0(\bar{\omega})$ , equation (17) will be

$$\frac{1}{2} \sum_{i,j=1}^l \|x_i - x_j\|_2^2 \exp(-\omega \|x_i - x_j\|_2^2) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) = -\gamma_2,$$

$$\gamma_2 > 0 \Leftrightarrow \sum_{i,j=1}^l \|x_i - x_j\|_2^2 \exp(-\omega \|x_i - x_j\|_2^2) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) < 0. \quad (19)$$

**Case 3:** If  $L_0(\bar{\omega}) = \emptyset$ , the solution lies in the interior of the feasible region and then the necessary condition for optimality is the same as for unconstrained case:  $\nabla g((\sigma, \rho), \bar{\omega}) = 0$ . It leads to solve  $\bar{\omega}$  from

$$\sigma - \sum_{i,j=1}^l k(x_i, x_j, \bar{\omega}) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) + \int_A \sum_{i=1}^l \alpha_i d\rho(\alpha) = 0.$$

For the following, we introduce

$$\bar{\gamma} := \begin{cases} \gamma_1, & \text{if case 1 holds,} \\ \gamma_2, & \text{if case 2 holds,} \\ 0, & \text{if case 3 holds.} \end{cases} \quad (20)$$

(3) *Second Order Condition (SOC)*: With our value  $\bar{\gamma}$  introduced it is fulfilled

$$\eta^T \nabla_{\omega}^2 \mathcal{L}^D(\sigma, \rho; \bar{\omega}, \bar{\gamma}) \eta > 0, \quad \text{for all } \eta \in \mathcal{T}^D(\bar{\omega}) \setminus \{0\},$$

where  $\mathcal{T}^D(\bar{\omega}) = \{\eta \in \mathbb{R} \mid \nabla^T v_\ell(\bar{\omega}) \eta = 0 \ (\ell \in L_0(\bar{\omega}))\}$ .

Let us find the tangent space  $\mathcal{T}^D(\bar{\omega})$  for all cases, and evaluate (SOC) with respect to them by the following cases. Here, we write  $\mathcal{L}_j^D, \mathcal{T}_j^D$  ( $j = 1, 2, 3$ ) according to those cases. (The same later on the dual case.)

**Case 1:** If  $v_1(\bar{\omega}) = 0$ , then  $\mathcal{T}_1^D(\bar{\omega}) = \{0\}$ .

(SOC)  $\eta^T \nabla_{\omega}^2 \mathcal{L}_1^D(\bar{\omega}, \bar{\gamma}) \eta > 0 \ \forall \eta \in \mathcal{T}_1^D(\bar{\omega}) \setminus \{0\}$   
is fulfilled, since  $\forall \eta \in \emptyset$ .

**Case 2:** If  $v_2(\bar{\omega}) = 0$ , then  $\mathcal{T}_2^D(\bar{\omega}) = \{0\}$ .

(SOC)  $\eta^T \nabla_{\omega}^2 \mathcal{L}_2^D(\bar{\omega}, \bar{\gamma}) \eta > 0 \ \forall \eta \in \mathcal{T}_2^D(\bar{\omega}) \setminus \{0\}$   
is fulfilled since  $\mathcal{T}_2^D(\bar{\omega}) \setminus \{0\} = \emptyset$ .

**Case 3:**  $L_0(\bar{\omega}) = \emptyset \Rightarrow \mathcal{T}_3^D(\bar{\omega}) = \mathbb{R}$ .

Then, the Lagrange function consists only of the objective function  $g((\sigma, \rho), \omega)$  which gives

$$\mathcal{L}_3^D(\omega) = \sigma - \sum_{i,j=1}^l k(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) + \int_A \sum_{i=1}^l \alpha_i d\rho(\alpha).$$

$$\begin{aligned} (SOC) \quad \nabla^2 \mathcal{L}_3^D(\bar{\omega}) = \\ - \frac{1}{2} \sum_{i,j=1}^l \|x_i - x_j\|_2^4 \exp\left(-\bar{\omega} \|x_i - x_j\|_2^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) > 0. \end{aligned} \quad (21)$$

Thus  $\bar{\omega}$  is nondegenerate if and only if the sign conditions (on the multipliers) and, in case 3,

$$\sum_{i,j=1}^l \|x_i - x_j\|_2^4 \exp\left(-\bar{\omega} \|x_i - x_j\|_2^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) < 0$$

are fulfilled.

We underline that this essentially depends on the data given. One of the important differences between the dual and primal problem is that the dual problem (15) reduces the dimension in the lower level from  $l$  to 1. Observe that the infinitely many inequality constraints of the dual problem depends on *one-dimensional* variable  $\omega$ , whereas in the primal problem they depend on the  $l$  dimensional variable  $\alpha$ .

Hence, working with dual problem is analytically more easy and computationally more tractable. However, the interpretation of the classification function for SVM is difficult if we solve (15) because of the the infinite dimension of nonlinear mapping  $\phi(x)$ . For example, even when we have one kernel, in particular, a Gaussian kernel,

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \exp(-\omega \|x_i - x_j\|_2^2),$$

and  $w = \sum_{i=1}^l \alpha_i y_i \phi(x_i)$ , it is difficult to interpret  $f(x) = \langle w, \phi(x) \rangle + b$  since we do not know the explicit form of  $\phi(x)$  and its dimension is infinite [7]. Because of these reasons, we propose to solve the primal problem to use kernel function implicitly without applying  $\phi(x)$  with primal variables  $\alpha_i$  in our problem.

### 3.3. Primal Problem

In this section, regularity conditions will be analyzed for the lower level of primal problem given as follows:

$$\begin{aligned} \max_{\theta, \beta} \quad & \theta \quad (\theta \in \mathbb{R}, \beta \text{ is a positive measure on } \Omega) \\ \text{subject to} \quad & \theta - \int_{\Omega} T(\omega, \alpha) d\beta(\omega) \leq 0 \quad (\alpha \in A). \end{aligned} \quad (22)$$

The standard form of (22) can be easily written by

$$\begin{aligned} \min_{\theta, \beta} \quad & (-\theta) \quad (\theta \in \mathbb{R}, \beta \text{ is a positive measure on } \Omega) \\ \text{subject to} \quad & \int_{\Omega} T(\omega, \alpha) d\beta(\omega) - \theta \geq 0 \quad (\alpha \in A). \end{aligned} \quad (23)$$

Using the language of bilevel programming of SIP, introduce the function  $g((\theta, \beta), \alpha) := \int_{\Omega} T(\omega, \alpha) d\beta(\omega) - \theta$  which is parametric in  $(\theta, \beta)$ . We state the

**Lower Level Problem (Primal):** For a given  $(\theta, \beta)$  we consider

$$\begin{aligned} \min_{\alpha} \quad & g((\theta, \beta), \alpha) \\ \text{subject to} \quad & \alpha \in A. \end{aligned} \quad (24)$$

We write the defining inequality constraint functions of  $A$  by  $v_r((\theta, \beta), \alpha) =: \alpha_i$ ,  $v_s((\theta, \beta), \alpha) =: -\alpha_{l-s} + C$ , where  $r \in \{1, \dots, l\}$  and  $s \in \{l+1, \dots, 2l\}$ , and equality constraints by  $u((\theta, \beta), \alpha) =: \sum_{i=1}^l \alpha_i y_i$ . Let us briefly denote  $v_r((\theta, \beta), \alpha) =: v_r(\alpha)$ ,  $v_s((\theta, \beta), \alpha) =: v_s(\alpha)$  and  $u((\theta, \beta), \alpha) =: u(\alpha)$ , and  $L_0(\bar{\alpha}) := \{\ell \in L \mid v_{\ell}(\bar{\alpha}) = 0\}$ , where  $L := \{1, 2, \dots, 2l\}$ . Consequently, for any critical point  $\bar{\alpha}$ , the Lagrange function reads

$$\mathcal{L}^P(\theta, \beta; \alpha, \zeta, \gamma) := g((\theta, \beta), \alpha) - \zeta u(\alpha) - \sum_{\ell \in L_0(\bar{\alpha})} \gamma_{\ell} v_{\ell}(\alpha).$$

Let us shortly write  $\mathcal{L}^P(\alpha, \zeta, \gamma) := \mathcal{L}^P(\theta, \beta; \alpha, \zeta, \gamma)$ . Since  $A$  is compact, for any local  $(\theta, \beta)$ , (global) minimizer(s) of (24) exists. We analyze the conditions of the nondegeneracy and *reduction ansatz* [8], at any such an  $\bar{\alpha}$ . For all  $(\theta, \beta)$  and each candidate  $\bar{\alpha} \in A$ , we evaluate:

- (1) *LICQ*: We have to check linear independence of  $\nabla v_r(\alpha)$ ,  $\nabla v_s(\alpha)$  and  $\nabla u(\alpha)$ , where  $r \in \{1, \dots, l\}$  and  $s \in \{l+1, \dots, 2l\}$  are active. In other words, variables  $\alpha \in \mathbb{R}^l$  can satisfy either  $v_r(\alpha) = \alpha_r$  or  $v_s(\alpha) = -\alpha_{l-s} + C$ . The Jacobian of the (active) inequalities can be calculated simply as follows:  $\nabla v_r(\bar{\alpha}) = (0, \dots, 0, 1, 0, \dots, 0)^T$  and  $\nabla v_s(\bar{\alpha}) = (0, \dots, 0, -1, 0, \dots, 0)^T$ . For simplicity, we introduce  $\mathcal{A}(\alpha)$  as the vector of all active constraints, the

equality constraint included:

$$\mathcal{A}(\alpha) = \begin{bmatrix} u(\alpha) \\ v_{\ell_1}(\alpha) \\ v_{\ell_2}(\alpha) \\ \vdots \\ v_{\ell_k}(\alpha) \end{bmatrix}, \text{ where } L_0(\bar{\alpha}) = \{\ell_1, \ell_2, \dots, \ell_k\}, |L_0(\bar{\alpha})| = k.$$

Then, the Jacobi matrix is a  $(k + 1) \times l$  matrix and looks as follows:

$$D\mathcal{A}(\alpha) = \begin{bmatrix} y_1 & y_2 & y_3 & \dots & \dots & \dots & y_l \\ 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & -1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & -1 \end{bmatrix},$$

where  $y_i \in \{\pm 1\}$  ( $i = 1, 2, \dots, l$ ) and  $k = |L_0(\bar{\alpha})|$ . On the right-hand side, we took the example of one matrix for illustration.

Observe that the matrix  $\text{rank}(D\mathcal{A}(\alpha)) = l$  if  $l < k + 1$ , which means that LICQ condition is violated since rank of  $D\mathcal{A}$  is smaller than the number of rows (i.e., constraints involved). This shows linear dependence of the row vectors, i.e., linear dependence of gradients of (active) constraints.

Let us geometrically analyze this condition in 2 dimensions, i.e.,  $l = 2$ . In Figure 5, two *different* examples of nondegeneracy cases are given such that at the origin and at the upper right corner, three active constraints meet and these points (corners) are degenerate because of the linear dependencies. At these points, we have three equations in two dimensions.

Let us introduce a sequence  $\xi_\nu > 0$  ( $\nu \in \mathbb{N}_0$ ) which is monotonically decreasing to zero such that the inequalities  $-\xi_\nu \leq \sum_{i=1}^l \alpha_i y_i \leq \xi_\nu$  are requested. Regarding active inequality constraints as equality constraints will lead to lines which do not pass through the origin and cannot produce a corner. This is shown as two examples in Figure 5 and Figure 6. At the blue points which are feasible points for our perturbed problem, the gradients of active constraints are linearly independent. Thus, by decreasing  $\xi_\nu$  to zero, for non degeneracy, LICQ can be forced while perturbation.

- (2) *Kuhn-Tucker condition with strictly positive Lagrange multipliers (for active inequalities):*

There has to exist a multiplier vector  $\bar{\gamma} \in \mathbb{R}^{|L_0(\bar{\alpha})|}$  such that  $\nabla_\alpha \mathcal{L}^P(\theta, \beta; \alpha, \zeta, \gamma) = 0$  and  $\bar{\gamma}_\ell > 0$  ( $\ell \in L_0(\bar{\alpha})$ ).

Let us consider all cases which make Lagrange multiplier strictly positive:

**Case 1:**  $L_0(\bar{\alpha}) \neq \emptyset$ . If we rewrite  $g((\theta, \beta), \alpha)$ , it has the following form:

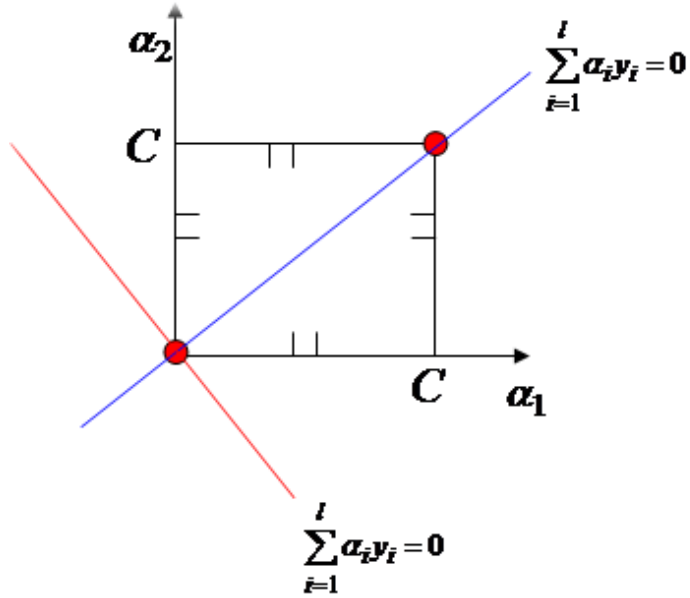


Figure 5. Active constraints, red dots are degenerate points, two examples.

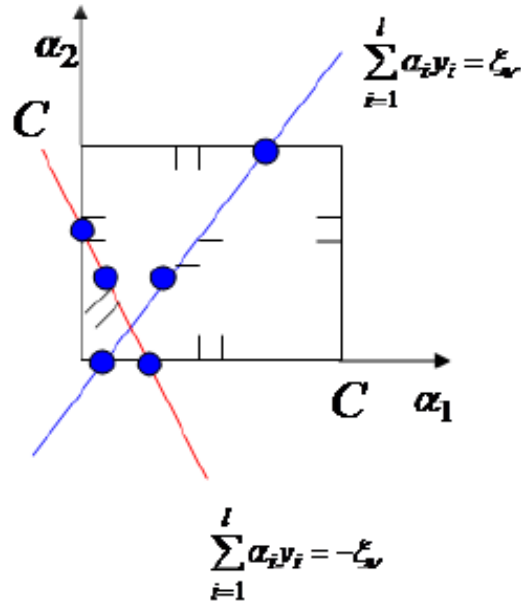


Figure 6. Active constraints with regular points in the perturbed problem, two examples.

$$\begin{aligned}
 g((\theta, \beta), \alpha) &= \int_{\Omega} T(\omega, \alpha) d\beta(\omega) - \theta \\
 &= \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \underbrace{\int_{\Omega} \kappa(x_i, x_j, \omega) d\beta(\omega)}_{=: M_{i,j}} - \sum_{i=1}^l \alpha_i - \theta \\
 &= \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j M_{i,j} - \sum_{i=1}^l \alpha_i - \theta.
 \end{aligned}$$

Here, we used a special condition of a probability measure  $\beta$ :  $\int_{\Omega} d\beta(\omega) = 1$ . Note that  $M_{i,j}$  is constant with respect to  $\alpha$  but dependent on  $(\theta, \beta)$ . If we substitute  $g((\theta, \beta), \alpha)$  into the Lagrange function, we will get the following representation:

$$\begin{aligned} \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma) &= \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j M_{i,j} - \sum_{i=1}^l \alpha_i - \theta - \zeta u(\alpha) - \sum_{\ell \in L_0(\bar{\alpha})} \gamma_{\ell} v_{\ell}(\alpha) \\ &= \frac{1}{2} \sum_{i=1}^l \alpha_i^2 y_i^2 M_{i,i} - \frac{1}{2} \sum_{i \neq j}^l \alpha_i \alpha_j y_i y_j M_{i,j} - \sum_{i=1}^l \alpha_i - \theta - \zeta u(\alpha) - \sum_{\ell \in L_0(\bar{\alpha})} \gamma_{\ell} v_{\ell}(\alpha). \end{aligned}$$

In the second line, we assumed that our kernel function is a Gaussian kernel which is  $\kappa(x_i, x_j, \omega) = \exp(-\omega \|x_i - x_j\|_2^2)$ . For  $i = j$ , we get  $\kappa(x_i, x_i, \omega) = 1$ .

To find KKT points  $(\bar{\alpha}, \bar{\zeta}, \bar{\gamma})$ , we need to solve  $\nabla_{\alpha} \mathcal{L}^{\mathcal{P}}(\alpha, \zeta, \gamma) = 0$ , which is a system of linear equations in  $(\zeta, \gamma)$  with

$$\nabla_{\alpha} \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma) = \left[ \frac{\partial \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_1}, \frac{\partial \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_2}, \dots, \frac{\partial \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_l} \right]^T, \quad (25)$$

where for all  $i = 1, \dots, l$ ,

$$\frac{\partial \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_i} = \alpha_i y_i^2 M_{i,i} - \frac{1}{2} \sum_{\substack{j=1, \\ j \neq i}}^l \alpha_j y_i y_j M_{i,j} - 1 - \zeta \frac{\partial u(\alpha)}{\partial \alpha_i} - \sum_{\ell \in L_0(\alpha)} \gamma_{\ell} \frac{\partial v_{\ell}(\alpha)}{\partial \alpha_i}.$$

Let us for the sake of simplicity assume that  $L_0(\bar{\alpha}) = \{1, \dots, k\}$ , renumbering the active inequalities otherwise. Then, from  $\nabla_{\alpha} \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma) = 0$  we get the following equations:

$$\begin{aligned} \alpha_1 y_1^2 - \frac{1}{2} \sum_{\substack{j=1 \\ j \neq 1}}^l \alpha_j y_i y_j M_{i,j} - 1 &= \zeta \frac{\partial u(\alpha)}{\partial \alpha_1} + \sum_{\ell=1}^k \gamma_{\ell} \frac{\partial v_{\ell}(\alpha)}{\partial \alpha_1}, \\ \alpha_2 y_2^2 - \frac{1}{2} \sum_{\substack{j=1 \\ j \neq 2}}^l \alpha_j y_i y_j M_{i,j} - 1 &= \zeta \frac{\partial u(\alpha)}{\partial \alpha_2} + \sum_{\ell=1}^k \gamma_{\ell} \frac{\partial v_{\ell}(\alpha)}{\partial \alpha_2}, \\ &\vdots \\ \alpha_l y_l^2 - \frac{1}{2} \sum_{\substack{j=1 \\ j \neq l}}^l \alpha_j y_i y_j M_{i,j} - 1 &= \zeta \frac{\partial u(\alpha)}{\partial \alpha_l} + \sum_{\ell=1}^k \gamma_{\ell} \frac{\partial v_{\ell}(\alpha)}{\partial \alpha_l}. \end{aligned} \quad (26)$$

The systems of equations (26) can be written in the matrix-vector multiplication form as follows:

$$\begin{bmatrix} \frac{\partial u(\alpha)}{\partial \alpha_1} & \frac{\partial v_1}{\partial \alpha_1} & \cdots & \frac{\partial v_k}{\partial \alpha_1} \\ \frac{\partial u(\alpha)}{\partial \alpha_2} & \frac{\partial v_1}{\partial \alpha_2} & \cdots & \frac{\partial v_k}{\partial \alpha_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u(\alpha)}{\partial \alpha_l} & \frac{\partial v_1}{\partial \alpha_l} & \cdots & \frac{\partial v_k}{\partial \alpha_l} \end{bmatrix} \begin{bmatrix} \zeta \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_l \end{bmatrix}, \quad (27)$$

where  $A_i = \alpha_i y_i^2 - \frac{1}{2} \sum_{j \neq i} \alpha_j y_i y_j M_{i,j} - 1$  ( $i = 1, \dots, l$ ). If we solve (27) restricted to  $\gamma_{\ell} > 0$  ( $\ell = 1, \dots, k$ ), we can specify rank and conditioning properties for  $\alpha = \bar{\alpha}$  being a candidate of a locally optimal solution.

**Case 2:**  $L_0(\bar{\alpha}) = \emptyset$ , i.e., the equality constraint is the only active constraint. Our Lagrangian will take the following form:

$$\begin{aligned}\mathcal{L}_2^{\mathcal{P}}(\alpha, \zeta) &= \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j M_{i,j} - \sum_{i=1}^l \alpha_i - \theta - \zeta u(\alpha) \\ &= \frac{1}{2} \sum_{i=1}^l \alpha_i^2 y_i^2 M_{i,i} - \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^l \alpha_i \alpha_j y_i y_j M_{i,j} - \sum_{i=1}^l \alpha_i - \theta - \zeta u(\alpha).\end{aligned}$$

Here, to illustrate the Case 2, we assumed that we have a Gaussian kernel as in Case 1. Let us find  $\zeta \in \mathbb{R}$  which satisfies  $\nabla_{\alpha} \mathcal{L}_2^{\mathcal{P}}(\alpha, \zeta) = 0$ , i.e.,

$$\nabla_{\alpha} \mathcal{L}_2^{\mathcal{P}}(\alpha, \zeta) = \left[ \frac{\partial \mathcal{L}_2^{\mathcal{P}}(\alpha, \zeta)}{\partial \alpha_1}, \frac{\partial \mathcal{L}_2^{\mathcal{P}}(\alpha, \zeta)}{\partial \alpha_2}, \dots, \frac{\partial \mathcal{L}_2^{\mathcal{P}}(\alpha, \zeta)}{\partial \alpha_l} \right]^T = 0, \quad (28)$$

where  $\frac{\partial \mathcal{L}_2^{\mathcal{P}}(\alpha, \zeta)}{\partial \alpha_i} = \alpha_i y_i^2 M_{i,i} - \frac{1}{2} \sum_{i \neq j} \alpha_j y_i y_j M_{i,j} - 1 - \zeta \frac{\partial u(\alpha)}{\partial \alpha_i}$  and  $M_{i,i} = 1$  as in previous case.

If we expand (28), we get the following system of equations:

$$\begin{aligned}\alpha_1 y_1^2 - \frac{1}{2} \sum_{j \neq 1} \alpha_j y_i y_j M_{i,j} - 1 &= \zeta \frac{\partial u(\alpha)}{\partial \alpha_1}, \\ \alpha_2 y_2^2 - \frac{1}{2} \sum_{j \neq 2} \alpha_j y_i y_j M_{i,j} - 1 &= \zeta \frac{\partial u(\alpha)}{\partial \alpha_2}, \\ &\vdots \\ \alpha_l y_l^2 - \frac{1}{2} \sum_{j \neq l} \alpha_j y_i y_j M_{i,j} - 1 &= \zeta \frac{\partial u(\alpha)}{\partial \alpha_l}.\end{aligned} \quad (29)$$

The above system of equations (29) can be written in matrix-vector multiplication form as follows:

$$\begin{bmatrix} \frac{\partial u(\alpha)}{\partial \alpha_1} \\ \frac{\partial u(\alpha)}{\partial \alpha_2} \\ \vdots \\ \frac{\partial u(\alpha)}{\partial \alpha_l} \end{bmatrix} \zeta = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_l \end{bmatrix}, \quad (30)$$

where  $\frac{\partial u(\alpha)}{\partial \alpha_i} = y_i$ . Hence, (30) becomes the following linear system:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix} \zeta = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_l \end{bmatrix}. \quad (31)$$

If we can solve (31), we can find the conditions for  $\alpha = \bar{\alpha}$  to be the optimal solution.

(3) *Second Order Condition (SOC)*: With our value  $\bar{\gamma}$  introduced it is fulfilled:

$$\eta^T \nabla_{\alpha}^2 \mathcal{L}^{\mathcal{P}}(\bar{\alpha}, \bar{\gamma}) \eta > 0 \quad \text{for all } \eta \in \mathcal{T}^{\mathcal{P}}(\bar{\alpha}) \setminus \{0\},$$

where  $\mathcal{T}^{\mathcal{P}}(\bar{\alpha}) = \{\eta \in \mathbb{R}^l \mid \nabla^T u(\bar{\alpha})\eta = 0, \nabla^T v_l(\bar{\alpha})\eta = 0 \ (l \in L_0(\bar{\alpha}))\}$ .

Now, let us find tangent space and conditions for SOC to be satisfied for all cases:

**Case 1:** If  $L_0(\bar{\alpha}) \neq \emptyset$ , then the tangent space of the form

$\mathcal{T}_1^{\mathcal{P}}(\bar{\alpha}) = \{\eta \in \mathbb{R}^l \mid D\mathcal{A}(\bar{\alpha})\eta = 0\}$ , with the condition (written a bit like an example again)

$$\begin{bmatrix} y_1 & y_2 & y_3 & \dots & \dots & \dots & y_l \\ 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & -1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & -1 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \eta_l \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \quad (32)$$

Here,  $\nabla v_\ell(\alpha) = (0, \dots, +1, 0, \dots, 0)^T \ (l \in L_0(\bar{\alpha}))$  and  $k = |L_0(\bar{\alpha})|$ . Equation (32) yields the following condition:

$$\eta_r = 0, \quad \forall r \in L_0(\bar{\alpha}) \cap \{1, \dots, l\}, \quad (33)$$

$$\eta_s = 0, \quad \forall l + s \in L_0(\bar{\alpha}) \cap \{l + 1, \dots, 2l\}, \quad (34)$$

$$\sum_{i=1}^l \eta_i y_i = 0. \quad (35)$$

From (33)-(35), it follows that  $\sum_{\substack{i=1 \\ i, l+i \notin L_0(\bar{\alpha})}}^l y_i \eta_i = 0$ .

Let us note that  $\nabla_\alpha^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \gamma, \zeta)$  explicitly as follows:

$$\nabla_\alpha^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial^2 \alpha_1} & \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_1 \partial \alpha_2} & \dots & \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_1 \partial \alpha_l} \\ \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_2 \partial \alpha_1} & \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial^2 \alpha_2} & \dots & \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_2 \partial \alpha_l} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_l \partial \alpha_1} & \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_l \partial \alpha_2} & \dots & \frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial^2 \alpha_l} \end{bmatrix},$$

with

$$\frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial^2 \alpha_i} = y_i^2 = 1 > 0, \quad (36)$$

and

$$\frac{\partial^2 \mathcal{L}_1^{\mathcal{P}}(\alpha, \zeta, \gamma)}{\partial \alpha_i \partial \alpha_j} = -\frac{1}{2} y_i y_j M_{i,j} \quad (i \neq j), \quad (37)$$

and recall:

**THEOREM 3.1** [6]. *A symmetric  $n \times n$  matrix  $M$  is positive definite (positive semi-definite) if and only if any one of the following conditions holds.*

a) *Every eigenvalue of  $M$  is positive (zero or positive).*

- b) All the leading principal minors of  $M$  are positive (all the principal minors of  $M$  are positive).
- c) There exists an  $n \times n$  nonsingular matrix  $N$  (an  $n \times n$  singular matrix  $N$  or an  $m \times m$  matrix  $N$  with  $m < n$ ) such that  $M = N^T N$ .

**COROLLARY 3.2** [24]. If  $A = [a_{i,j}]$  is symmetric  $n \times n$  strictly diagonally dominant matrix with positive real diagonal entries, then  $A$  is positive definite.

In particular, anyone of the conditions is given in Theorem 3.1 and Corollary 3.2 is satisfied accordingly, then the Hessian to be positive definite:  $\eta^T \nabla_{\alpha}^2 \mathcal{L}_1^{\mathcal{P}}(\bar{\alpha}, \bar{\zeta}, \bar{\gamma}) \eta > 0$  for all  $\eta \in \mathcal{T}_1^{\mathcal{P}}(\bar{\alpha}) \setminus \{0\}$ , so that (SOC) is satisfied.

**Case 2:**  $L_0(\bar{\alpha}) = \emptyset$ . Then,

$$\begin{aligned} \mathcal{T}_2^{\mathcal{P}}(\alpha) &= \{ \eta \in \mathbb{R}^l \mid \nabla^T u(\bar{\alpha}) \eta = 0 \}, \\ &= \left\{ \eta \in \mathbb{R}^l \mid \sum_{i=1}^l y_i \eta_i = 0 \right\}. \end{aligned}$$

The Hessian is the same as in case 1, with the same entries (36) and (37); and there are the same (SOC) conditions referring to  $\mathcal{T}_2^{\mathcal{P}}(\alpha)$ .

Under these assumptions, the following theorem assures the optimal solution locally in a neighbourhood of the optimal solution on the lower level. It is an extension of the results given in [8, 9], where now the parameter space is infinite dimensional.

**THEOREM 3.3** Let at some feasible point  $(\bar{\theta}, \bar{\beta})$  of (14) the condition reduction ansatz be satisfied. Then:

(a) The active index set is finite, in symbols:  $A_0 = \{\alpha_1, \dots, \alpha_{\chi}\}$ , and there exist neighbourhoods  $U_{(\bar{\theta}, \bar{\beta})}$  of  $(\bar{\theta}, \bar{\beta})$  and  $V_{\bar{\alpha}_j}$  of  $\bar{\alpha}_j$ , and continuous mappings

$$\alpha_j : U_{(\bar{\theta}, \bar{\beta})} \rightarrow V_{\bar{\alpha}_j}, \quad \alpha_j(\bar{\theta}, \bar{\beta}) = \bar{\alpha}_j \quad \text{and} \quad \gamma_j : U_{(\bar{\theta}, \bar{\beta})} \rightarrow \mathbb{R}^{|L_0((\bar{\theta}, \bar{\beta}), \alpha_j)|}, \quad \gamma_j(\bar{\theta}, \bar{\beta}) = \bar{\gamma}_j,$$

such that for every  $(\theta, \beta) \in U_{(\bar{\theta}, \bar{\beta})}$  the value  $\alpha_j(\theta, \beta)$  is the unique local minimizer of (24) in  $V_{\bar{\alpha}_j}$ , with corresponding Lagrange multiplier vector  $\gamma_j(\theta, \beta)$  ( $j = 1, 2, \dots, \chi$ ).

(b) With the functions in (a) the following finite reduction holds:  $(\theta, \beta) \in U_{(\bar{\theta}, \bar{\beta})} \cap \mathcal{F}$ , where  $\mathcal{F}$  is the feasible set of upper level problem (14), is a local solution of (14), if and only if  $(\bar{\theta}, \bar{\beta})$  is a local solution of the so-called reduced problem

$$\begin{aligned} P_{red}(\theta, \beta) : & \min_{(\theta, \beta) \in U_{(\bar{\theta}, \bar{\beta})}} (-\theta) \\ & \text{such that } G_j(\theta, \beta) := g((\theta, \beta), \alpha_j(\theta, \beta)) \geq 0 \quad (j = 1, \dots, \chi). \end{aligned} \quad (38)$$

*Remark 1* An analogous theorem holds for dual problem (15) with respect to dual variables. We underline that by this theorem the reduced problem has (locally) finitely many constraints. Then, our task becomes a finitely constrained optimization problem locally around optimal solution. This insight is based on *Implicit Function Theorem (IFT)* and the neighbourhood notion defined by the *Prokhorov*

distance introduced below.

For very general versions of Inverse and, hence, Implicit Function Theorem, we refer, e.g., to [11].

Let us start with some definitions which are necessary to define neighbourhood in terms of measures, including the probability measures of our study.

**DEFINITION 3.4** [18]. *Let  $(X, T)$  be a Hausdorff topological space and let  $\Sigma$  be a  $\sigma$ -algebra on  $X$  that contains the topology  $T$  (so that every open set is a measurable set, and  $\Sigma$  is at least as fine as the Borel  $\sigma$ -algebra on  $X$ ). A measure  $\mu$  defined on  $\Sigma$  is called locally finite if, for every point  $p$  of the space  $X$ , there is an open neighbourhood  $N_p$  of  $p$  such that a measure  $\mu$  of  $N_p$  is finite.*

*In more condensed notation,  $\mu$  is locally finite if and only if*

$$\forall p \in X, \exists N_p \in T \text{ such that } p \in N_p \text{ and } |\mu(N_p)| < +\infty.$$

*With the same assumptions, a measure  $\mu$  on the measurable space  $(X, \Sigma)$  is called inner regular if, for every set  $A \in \Sigma$ ,*

$$\mu(A) = \sup\{\mu(K) \mid K \subseteq A \text{ compact}\}.$$

This property is sometimes referred to in words as approximation from within by compact sets.

After giving these definitions, we define a *Radon measure* and the distance metric needed for neighbourhoods in Theorem 3.3:

**DEFINITION 3.5** [14]. *Let  $(E, d)$  be the metric space. A Radon measure is a measure on the  $\sigma$ -algebra of Borel sets of  $E$  that is locally finite and inner regular.*

We denote the set of Radon measure by  $\mathcal{H}(E)$ . In our problems, we look at the subspaces of the all probability measures  $\rho$  for the dual problem (15), and  $\beta$  for the primal problem (22).

**DEFINITION 3.6** [14]. *Let  $f_i : E \rightarrow \mathbb{R}$  ( $i = 1, 2, \dots, n$ ) be continuous bounded functions,  $f_i \in (\mathcal{H}(E))'$ , where  $(\mathcal{H}(E))'$  is the dual space of  $\mathcal{H}(E)$ . A base of neighbourhood can be defined as  $\{\mu \in \mathcal{H}(E) \mid |\int_E f_i d\rho - \int_E f_i d\mu| < \epsilon \text{ (} i = 1, 2, \dots, n \text{)}\}$ .*

In our problems, the elements in the dual space are pdfs. Now, to represent our neighbourhood notion by a metric, let us define *Prokhorov distance*:

**DEFINITION 3.7** [14]. *Let  $(E, d)$  be a metric space, where  $d_0$  is a Prokhorov distance between any  $\mu, \rho \in \mathcal{H}(E)$  is defined by*

$$d_0(\mu, \rho) := \inf \{ \epsilon \geq 0 \mid \mu(A) \leq \rho(A_\epsilon) + \epsilon \text{ and } \rho(A) \leq \mu(A_\epsilon) + \epsilon \text{ (} A \subseteq E, \text{ closed)} \},$$

*with  $A_\epsilon := \{x \in E \mid d(x, A) < \epsilon\}$ . Then, the  $\delta$ -open neighbourhood of  $\rho$  is defined by  $B_\delta(\rho) := \{\mu \in \mathcal{H}(E) \mid d_0(\rho, \mu) < \delta\}$ .*

*Remark 2* Definition 3.7 allows to define a neighbourhood of  $(\sigma, \rho)$  in an appropriate topological sense. By Theorem 3.3 and Definition 3.7 we specify the meaning of reduction ansatz and of a local optimal solution, namely, in one of these neighbourhoods.

#### 4. Different Parametrization Functions for Infinite Problem

Until now, we have assumed that parameters  $(\theta, \beta)$  and  $(\sigma, \rho)$  are given for both primal problem and dual problem. In this section, we will introduce and discuss various parametrizations for our problems. We consider a positive measure  $\beta$  such that  $\int_0^1 d\beta(\omega) = 1$ , and we select probability density functions (pdfs)  $f$  such that  $f(\omega)d\omega$  takes the place of  $d\beta(\omega)$ . For example, the pdfs of a *normal*, *exponential*, *uniform*, *beta*, or *Poisson distribution* [20].

**Normal Distribution:** This distribution is also called *Gaussian distribution*; it is very approximate for modelling of various continuous random variables. The sampling distribution of the sample mean is approximately normal, even if the distribution of the population from which the sample is taken is not normal [20]. The pdf of a normal distribution is

$$f(\omega; (\mu, \sigma^2)) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(\omega - \mu)^2}{2\sigma^2}.$$

where  $\omega, \mu, \sigma \in \mathbb{R}$ .

For the multidimensional case, i.e.,  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_l)^T$ , then the pdf becomes

$$f(\alpha; (\mu, \Sigma)) = \frac{1}{(2\pi)^{l/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\alpha - \mu)^T \Sigma^{-1}(\alpha - \mu)\right).$$

**Exponential Distribution:** This distribution is a class of continuous probability distributions which is useful for modelling time between independent events of constant average rate [20]. The pdf of an *exponential distribution* looks as follows:

$$f(\omega; \lambda) = \begin{cases} \lambda \exp(-\lambda\omega), & \omega \geq 0, \\ 0, & \omega < 0. \end{cases}$$

Since  $\omega \in [0, 1]$  in our problem,  $\beta(\omega) = \lambda \exp(-\lambda\omega)d\omega$ , where  $\lambda \in \mathbb{R}$  is a parameter of rate. Of course, translations of the origin 0, e.g., delay, are possible.

**Continuous Uniform Distribution:** This is a family of probability distributions such that for each member of the family, all intervals  $[a, b]$  of the same length on the distribution's support are equally probable. The pdf of a *continuous uniform distribution* looks as follows:

$$f(\omega; (a, b)) = \begin{cases} \frac{1}{b-a}, & a \leq \omega \leq b, \\ 0, & \omega < a \text{ or } \omega > b. \end{cases}$$

**Beta Distribution:** The *Beta distribution* is a family of continuous probability distributions defined on the interval  $[0, 1]$  parameterized by two positive shape parameters, typically denoted by  $\alpha$  and  $\beta$ . (No confusion with the meaning of  $\alpha$  and  $\beta$  in our paper needs to be expected.) The pdf of a Beta distribution looks as follows:

$$f(\omega; (\alpha, \beta)) = \frac{\omega^{\alpha-1}(1-\omega)^{\beta-1}}{\int_0^1 \omega^{\alpha-1}(1-\omega)^{\beta-1}d\omega}.$$

### 5. Regularization of Infinite Programming Model with respect to Kernel Coefficients

In the previous section, our classification problem became modelled with infinitely many kernels by infinite programming. Infinity may cause ill-posedness which is called *overfitting* in regression problems. Here, we consider classification problems which needs to be regularized to penalize overfitting caused by infinity. This could be the case if any positive multiple of a kernel is also a kernel [3]. Argyriou et al. (2006) introduced a regularization term to prevent from overfitting of data by the objective function [3]:

$$Q(f) := \sum_{j=1}^l q(y_j, f(x_j)) + \lambda \|f\|_k^2, \quad (39)$$

where  $q(\cdot, \cdot)$  is a loss function and  $\|\cdot\|_k$  is the norm induced by *reproducing Kernel Hilbert space*. Here,  $f$  is represented by a combination of kernels as  $f = \sum_{j=1}^l k(x_j, \cdot)$ , which is known as *Representer Theorem* [19], and the parameters  $c_j$  become optimized [3]. In our infinite kernel representation with Riemann-Stieltjes integrals or positively defined measures, we need to find a penalization function in terms of measures  $\beta(\omega)$  (or  $\rho(\alpha)$ ) since they represent our continuous convex coefficient for infinite kernel combinations. Motivated by the theory of inverse problems [4, 23], this can be formulated as:

$$\begin{aligned} \min_{\theta, \beta} \quad & (-\theta) + \lambda \sup_{t \in [0,1]} \left| \frac{d^\nu}{dt^\nu} \int_0^t d\beta(\omega) \right| \\ \text{subject to} \quad & \int_\Omega T(\omega, \alpha) d\beta(\omega) - \theta \geq 0 \quad (\alpha \in A), \end{aligned} \quad (40)$$

where the second term in the objective function is the regularization term and  $\lambda$  is a regularization constant. With  $\nu = 1, 2$ , we express that we take into account and penalize first- or second-order derivatives which we can interpret as feathness and energy of our models, respectively.

Another formulation can be done by including the kernel combination  $k_\beta$  derived, e.g., by homotopy discussed in Section 3.1, as follows:

$$\begin{aligned} \min_{\theta, \beta} \quad & (-\theta) + \lambda \sum_{i,j=1}^l \sup_{t \in [0,1]} \left| \frac{d^\nu}{dt^\nu} \int_0^t k(x_i, x_j, \omega) d\beta(\omega) \right| \\ \text{subject to} \quad & \int_\Omega T(\omega, \alpha) d\beta(\omega) - \theta \geq 0 \quad (\alpha \in A), \end{aligned} \quad (41)$$

where  $\lambda$  is a regularization parameter again.

Observe that our regularization function highly depends on parameter  $\beta(\omega)$  and it usually needs to be twice continuously differentiable to be well-defined. To weaken the need of differentiability, we replace the derivatives by first- and second-order coefficient formulas, as offered in the example below, where  $0 = t_0 < t_1 < \dots < t_m = 1$ .

**Example: (dual case)**

- First-order difference quotient:

$$\begin{aligned} \frac{d}{dt} \int_0^t d\beta(\omega) &\approx \frac{\int_0^{t_{\nu+1}} d\beta(\omega) - \int_0^{t_\nu} d\beta(\omega)}{t_{\nu+1} - t_\nu} \\ &= \frac{1}{t_{\nu+1} - t_\nu} \int_{t_\nu}^{t_{\nu+1}} d\beta(\omega) \quad (\nu \in \{0, 1, \dots, m-1\}). \end{aligned}$$

- Second-order difference quotient:

$$\frac{d^2}{dt^2} \int_0^t d\beta(\omega) \approx \frac{\frac{1}{t_{\nu+2} - t_{\nu+1}} \int_{t_{\nu+1}}^{t_{\nu+2}} d\beta(\omega) - \frac{1}{t_{\nu+1} - t_\nu} \int_{t_\nu}^{t_{\nu+1}} d\beta(\omega)}{t_{\nu+1} - t_\nu} \quad (\nu \in \{0, 1, \dots, m-2\}).$$

## 6. Conclusion and Future Study

The method we proposed in this study leads to the selection of kernels from an infinite space which enabled us to enrich the learning process SVM through the range interval  $[0, 1]$  of  $\omega$ . Hence, we are not limited to choose kernel parameter(s), Gaussian kernels in our special case, as discrete values with a cross validation method, but that depending on the examples given beforehand we can learn from data through this infinite process. By reduction ansatz, an infinite problem is turned to a locally finitely constrained problem, except of the fact that probability measures are our main state variables. By focusing on measures which possess a Radon-Nikodym density, we turn to a space of density functions [25]. By looking at parametric density functions, we get semi-infinite and, via reduction ansatz, a finitely constrained program indeed. Besides of that ansatz, also discretization and exchange methods will be analyzed and developed in future studies.

In this paper, the classification problem by SVM is modeled with infinitely many kernels by infinite programming. The proposed dimension is infinite, and it has infinitely many constraints which may cause ill-posedness. To overcome this, we introduced the regularization term into the objective function where the derivative of the regularization term is approximated by first- and second-order difference quotients. This kind of problems can be useful for real-world data which are heterogeneous, e.g., in bioinformatics and financial applications. The proposed method is novel by its kernel definitions in Riemann-Stieltjes integral form. On the other hand, our optimization problems are defined in probability measures as the state variables, which are infinite in dimension. Here, the parametrization is offered by positively defined measures via pdfs. We gave some examples of distribution functions to be applied. Another novelty of the model is to use Prokhorov distances between Radon measures to define neighbourhoods in the state space. In the future, numerical treatments such as exchange methods and gradient descent methods will be studied and presented.

**Acknowledgment:** The authors cordially thank the professors E. Anderson, U. Çapar, M. Goberna, and J. Shawe-Taylor for their valuable advices.

## References

- [1] E.J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley and Sons Ltd, 1987.
- [2] T.M. Apostol, *Mathematical Analysis: A Modern Approach to Advanced Calculus*, Addison Wesley, 1974.

- [3] A. Argyriou, R. Hauser, C.A. Micchelli and M. Pontil, A DC-programming algorithm for kernel selection, 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.
- [4] A. Aster, B. Borchers and C. Thurber, *Parameter Estimation and Inverse Problems*, Academic Press, 2004.
- [5] F.R. Bach and G.R.G. Lanckriet, Multiple kernel learning, conic duality, and the smo algorithm, International Conference on Machine Learning, Banff, Canada, 2004.
- [6] C.-T. Chen, *Linear System Theory and Design*, Oxford University Press, 1999.
- [7] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
- [8] R. Hettich and P. Zencke, *Numerische Methoden der Approximation und semi-infiniten Optimierung*, Teubner, Stuttgart, 1982.
- [9] R. Hettich and H.Th. Jongen, *Semi-infinite programming: conditions of optimality and applications*, in: *Optimization Techniques 2*, J. Stoer, eds., Lecture notes in Control and Information Sci., Springer, Berlin, Heidelberg, /New York, 1978, pp. 1-11.
- [10] M.A. Goberna and M.A. Lopez, *Linear Semi-Infinite Optimization*, John Wiley and Sons Ltd, 1998.
- [11] R.S. Hamilton, *The inverse function theorem of Nash and Moser*, Bulletin (New Series) of American Mathematical Society 7 (1), (1982).
- [12] E. Kropat, G.W. Weber and B. Akteke-Öztürk, Eco-finance networks under uncertainty, EngOpt 2008 - International Conference on Engineering Optimization, Rio de Janeiro, Brazil, 01 - 05 June 2008.
- [13] G.R.G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett and M.I. Jordan, *Learning the kernel matrix with semidefinite programming*, J. Machine Learning Research 5 (2004), pp. 27-72.
- [14] W. Linde, *Probability in Banach Spaces - Stable and Infinitely Divisible Distributions*, John Wiley and Sons, Chichester-New York, 1983.
- [15] A. Rakotomamonjy, F. Bach, S. Canu and Y. Grandvalet, More efficiency in multiple kernel learning, 24th International Conference on Machine Learning, Corvallis, 2007.
- [16] S. Özögür, J. Shawe-Taylor, G.-W. Weber, and Z.B. Ogel, *Pattern analysis for the prediction of eukaryotic pro-peptide cleavage sites*, to appear in the special issue of Discrete Applied Mathematics (DAM) Networks in Computational Biology (2007).
- [17] S. Özögür-Akyüz, Z. Hussain and J. Shawe-Taylor, *Prediction with the SVM using test point margins*, to appear in Annals of Information Systems, Springer, 2008.
- [18] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, AMS Chelsea Publishing, Providence, RI, 2005.
- [19] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.
- [20] A. N. Shiryaev, *Probability*, Springer Verlag, New York, 1995.
- [21] S. Sonnenburg, G. Raetsch, C. Schafer and B. Schölkopf, *Large scale multiple kernel learning*, J. Machine Learning Research 7 (2006), pp. 1531-1565.
- [22] G. Still, *Semi-infinite Programming: An Introduction*, preliminary version, Technical Report, 2004.
- [23] P. Taylan, G.-W. Weber and F. Yerlikaya, *Continuous optimization applied in MARS for modern application in finance, science and technology*, Continuous Optimization and Knowledge Based Technologies, 20th EURO Mini conference, Lithuania, 2008.
- [24] R.S. Varga, *Matrix Iterative Analysis*, Springer, 2000.
- [25] Z. Wan, S.Y. Wu and K.L. Teo, *Some Properties on quadratic infinite programs of integral type*, Applied Mathematics Letters 20 (2007), pp. 676-680.
- [26] G.-W. Weber, *Charakterisierung struktureller stabilität in der nichtlinearen optimierung in Aachener Beiträge zur Mathematik 5*, H.H. Bock, H.Th. Jongen and W. Plesken, eds., Augustinus publishing house (now: Mainz publishing house), Aachen, 1992.
- [27] G.-W. Weber, *Minimization of a max-type function: Characterization of structural stability*, in: *Parametric Optimization and Related Topics III*, J. Guddat, H.Th. Jongen, B. Kummer and F. Nozicka, eds., Peter Lang publishing house, Frankfurt a.M., Bern, 1993, pp. 519-538.
- [28] G.-W. Weber, P. Taylan, S. Ozogur and B. Akteke-Ozturk, *Statistical learning and optimization methods in data mining*, in: *Recent Advances in Statistics*, H.O. Ayhan and I. Batmaz, Turkish Statistical Institute Press, Ankara, 2007, pp. 181-195.
- [29] W.W.E. Wetterling, *Definitheitsbedingungen für r relative Extrema bei Optimierungsaufgaben*, Numer. Math. 12 (1970), pp. 122-136.