

LEARNING WITH INFINITELY MANY KERNELS VIA SEMI-INFINITE PROGRAMMING

Süreyya Özögür-Akyüz, Gerhard Wilhelm Weber

Institute of Applied Mathematics, Middle East Technical University
06531 Ankara, Turkey

E-mail: sozogur@metu.edu.tr, gweber@metu.edu.tr

Abstract. In recent years, learning methods are desirable because of their reliability and efficiency in real-world problems. We propose a novel method to find infinitely many 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. 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 conditions which satisfy the Reduction Ansatz and discuss the type of distribution functions of the kernel coefficients 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.

1. Introduction

By the innovation and development of the technology, high processor computers took place to do the work which the human did in the past. For instance, the classification or detection problems in real-world such as credit card frauding, account management, portfolio optimization, or in life sciences such as biological experiments (Özögür-Akyüz, 2007), prediction of cancer risk, finding pattern in genes or proteins can be identified without need any costing experiments. This helped human beings and industry to save time and finances. By mathematical modelling and computer science, experimental data are analyzed and data mining tools (Weber, 2007) developed. As the demands increase, new constraints are added, risks are to be minimized, etc.. Thus, it turns out to be an 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. Here, it is one of the most efficient tools and it bases on maximizing the margin γ between two classes of objects with some constraints. The classes are separated by an affine function via $\langle w, x \rangle + b = 0$, where w is a normal vector for the hyperplane and $w, x \in \mathbb{R}^N$, $b \in \mathbb{R}$ (Cristianini, 2000). Given a set of examples $S = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where $y_i \in \{\pm 1\}$ and $x_i \in \mathbb{R}^N$ ($i = 1, 2, \dots, N$), two groups of points are separated by a hyperplane as shown in Figure 1 (here, $\langle \cdot, \cdot \rangle$ denotes scalar product).

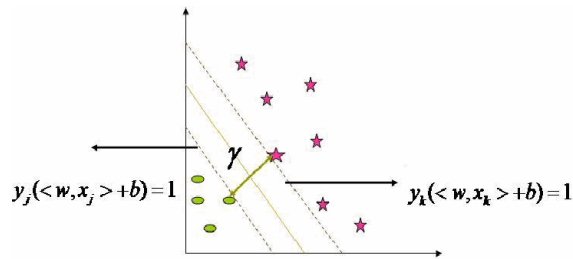


Figure 1: Maximum margin between two classes.

In most real-world problems (Özögür-Akyüz, 2007), data are not linearly separable. To use the facilities of the linear separable case, one can define a *non-linear mapping* ϕ which transforms the input space into the higher dimensional *feature space* such that the points are separable there.

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 (Cristianini, 2000), a *kernel function* is defined as an inner product of two points under the mapping ϕ , i.e., $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, which can also be explained as the similarity between two points. The optimization problem for separating two classes is expressed as follows (Cristianini, 2000):

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

its dual problem reads

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

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, variables are introduced that allow the maximal margin criterion to be violated. Then, this classifier is called a *soft margin classifier*. Here, a vector ξ_i of some slack variables is inserted into the constraints and, equipped with a regularization constant C , into the objective function as well ($\|\cdot\|_2$ denoting Euclidean norm):

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

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

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

Real-world data can be supplied from heterogeneous kinds of sources and it can be noisy. In such cases, multiple kernels are more convenient to use for a good accuracy. Recent applications (Lanckriet, 2004b) showed the need for *multiple kernel learning (MKL)* by its interpretability and efficiency. The common approach to MKL is a convex combination of kernels. In (Lanckriet, 2004a), the kernel-based SVM is formulated by combinations of multiple kernels and solved by quadratically-constrained quadratic programming (QCQP) applied to solve dual conic optimization problem. Likewise, (Sonnenburg, 2006) uses the adapted multiple kernel learning to large scale problems which is tractable, and applies the method to biological sequence analysis. Since the biological sequences have different motifs inside and for each subsequences different type of kernels are used, the combination is taken over the whole sequence. In (Sonnenburg, 2006), kernel coefficients are maximized beyond a minimization with respect to the dual variables, which is a *max-min* type of a problem. It can become canonically represented as a semi-infinite problem (Weber, 1992; Weber, 1993). The classical SVM is solved iteratively with linear programming and increasing the number of constraints iteratively (Sonnenburg, 2006). (Rakotomamonjy, 2007) proposes a different form of objective function in MKL by adapted weighted 2-norm regularization for each function f induced by κ_k ($k = 1, 2, \dots, M$) instead of using the 1-norm block regularization of (Sonnenburg, 2006) (M denoting the number of kernels). Sparsity of linear combinations of kernels is controlled by adding 1-norm regularization constants on these kernel weights.

2. Multiple Kernel Learning (MKL)

In this study, the kernel space is expanded as an infinite set; the problem will be formulated as an infinite programming and further analyzed by the help of semi-infinite optimization. Firstly, we regard

$$\kappa(x_i, x_j) = \sum_{k=1}^K \beta_k \kappa_k(x_i, x_j) \quad (2.1)$$

where $\beta_l \geq 0$ ($l=1, \dots, K$), $\sum_{k=1}^K \beta_k = 1$ and x_i is translated via mappings $x \mapsto \phi(x) \in \mathbb{R}^{D_l}$ from the input space into the feature spaces \mathbb{R}^{D_l} . In (Sonnenburg, 2006), the following MKL problem is derived by using the convex combination of kernels (2.1):

$$\begin{aligned} \text{Primal Multiple} \quad & \min \frac{1}{2} \left(\sum_{k=1}^K \|w_k\|_2 \right)^2 + C \sum_{i=1}^N \xi_i \quad (w_k \in \mathbb{R}^{D_k}, \xi \in \mathbb{R}^N, b \in \mathbb{R}) \\ \text{Kernel Problem} \quad & \text{subject to } \xi \geq 0, \quad y_i \cdot \left(\sum_{k=1}^K \langle w_k, \phi_k(x_i) \rangle + b \right) \geq 1 - \xi_i \quad (i=1, 2, \dots, N). \end{aligned} \quad (2.2)$$

In (Bach, 2004), the dual of the problem (2.2) is expressed with second-order cones as follows:

$$\begin{aligned} \text{Dual Multiple} \quad & \min \frac{1}{2} \gamma^2 - \alpha^T e \quad (\gamma \in \mathbb{R}, \alpha \in \mathbb{R}^N) \\ \text{Kernel Problem} \quad & \text{subject to } 0 \leq \alpha \leq C, \quad \alpha^T y = 0, \\ & \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \kappa_k(x_i, x_j) \leq \gamma \quad \forall k=1, \dots, K \end{aligned} \quad (2.3)$$

A tractable formulation for large-scale problems was introduced in (Sonnenburg, 2006) by using SILP (*semi-infinite linear programming*) (Goberna, 1998) for (2.3) rather than solving SDP (*semi-definite programming*) as (Bach, 2004). The Lagrange function of (2.3) can be written as

$$L(\alpha, \gamma) := \gamma + \sum_{k=1}^K \beta_k (S_k(\alpha) - \gamma), \quad \text{where } S_k(\alpha) = \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \kappa_k(x_i, x_j) - \sum_{i=1}^N \alpha_i.$$

When minimizing the Lagrangian with respect to $\alpha \in \mathbb{R}^N, \gamma \in \mathbb{R}$ and maximizing it with respect to $\beta \in \mathbb{R}^K$, we find: $\frac{\partial L}{\partial \gamma} = 0 \Rightarrow 1 - \sum_{k=1}^K \beta_k = 0$, i.e., $\sum_{k=1}^K \beta_k = 1$. Minimizing our Lagrangian in α but also maximizing it in β gives the following *max min* problem (Sonnenburg, 2006):

$$\begin{aligned} & \max_{\beta} \min_{\alpha} \sum_{k=1}^K \beta_k S_k(\alpha) \quad (\alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^K) \\ & \text{subject to } 0 \leq \alpha \leq C, \quad 0 \leq \beta, \\ & \sum_{i=1}^N \alpha_i y_i, \quad \sum_{k=1}^K \beta_k = 1, \end{aligned} \quad (2.4)$$

which can be represented as an SIP (semi-infinite programming) problem by a standard argument. Let us assume that α^* is an optimal solution, i.e., $S(\alpha, \beta) \geq \theta^*$ for all α , where $\theta^* := S(\alpha^*, \beta)$.

Then, problem (2.4) reduces to the smooth SILP *max* problem

$$\begin{aligned} & \max_{\theta, \beta} \theta \quad (\theta \in \mathbb{R}, \beta \in \mathbb{R}^K) \\ & \text{such that } 0 \leq \beta, \quad \sum_{k=1}^K \beta_k = 1, \\ & \sum_{k=1}^K \beta_k S_k(\alpha) \geq \theta \quad \forall \alpha \in \mathbb{R}^N \text{ with } 0 \leq \alpha \leq C \text{ and } \sum_{i=1}^N \alpha_i y_i = 0. \end{aligned} \quad (2.5)$$

3. Learning with Infinite Kernels

Based on the motivation of multiple kernel learning, we introduce a different formulation by introducing infinitely many kernels in the Riemann-Stieltjes (Apostol, 1974) integral form which covers an infinite dimensional kernels space. Let us assume that $(\eta_k)_{k \in \mathbb{N}_0}$ is a monotonically increasing sequence in the bounded interval $\Omega := [0, 1]$ tending to 1 as $k \rightarrow \infty$ and, say, $\eta_0 = 0$.

Then, $\sum_{k=1}^{\infty} (\eta_k - \eta_{k-1}) = 1$. We can refine the summation by *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

infinitesimal $d\beta(\omega)$ after limit calculus with weights $\beta_k = \beta(\omega_k) - \beta(\omega_{k-1})$, i.e., the incremental weights related to an index β of kernels $\kappa_\beta = \int_\Omega \kappa(x_i, x_j, \omega) d\beta(\omega)$, where β is a probability measure on Ω . If we introduce Riemann-Stieltjes integrals to the problem (2.6), we get the following general problem formulation:

$$\begin{aligned} & \max_{\theta, \beta} \theta \quad (\theta \in \mathbb{R}, \beta: [0,1] \rightarrow \mathbb{R} : \text{monotonically increasing}) \\ & \text{subject to} \quad \int_0^1 d\beta(\omega) = 1, \\ & \int_\Omega \left(\frac{1}{2} S(\omega, \alpha) - \sum_{i=1}^N \alpha_i \right) d\beta(\omega) \geq \theta \quad \forall \alpha \in \mathbb{R}^N \text{ with } 0 \leq \alpha \leq C, \sum_{i=1}^N \alpha_i y_i = 0. \end{aligned} \quad (3.1)$$

Here, S is defined by $S(\omega, \alpha) := \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j, \omega)$. Let us introduce $T(\omega, \alpha) := \frac{1}{2} S(\omega, \alpha) - \sum_{i=1}^N \alpha_i$, recall $\Omega = [0,1]$ and for the index set of inequality constraints we write: $A := \left\{ \alpha \mid 0 \leq \alpha \leq C, \sum_{i=1}^N \alpha_i y_i = 0 \right\}$. Herewith, (3.1) turns into the following form:

$$\begin{aligned} & \max_{\theta, \beta} \theta \quad (\theta \in \mathbb{R}, \beta: \text{a positive measure on } \Omega) \\ & \text{such that} \quad \theta - \int_\Omega T(\omega, \alpha) d\beta(\omega) \leq 0 \quad \forall \alpha \in A, \quad \int_\Omega d\beta(\omega) = 1. \end{aligned} \quad (3.2)$$

Since there are infinitely many constraints $\alpha \in A$ and the state variable β is from an infinite dimensional space, our problem is a one of infinite programming (IP) (Anderson, 1987). Now, we get a dual of (3.2) as

$$\begin{aligned} & \min_{\sigma, \rho} \sigma \quad (\sigma \in \mathbb{R}, \rho: \text{a positive measure on } A) \\ & \text{such that} \quad \sigma - \int_A T(\omega, \alpha) d\rho(\alpha) \geq 0 \quad \forall \omega \in \Omega, \quad \int_A d\rho(\alpha) = 1. \end{aligned} \quad (3.3)$$

Duality Conditions: Let us assume that there exist (θ, β) and (σ, ρ) that are feasible for their respective problems, and are complementary slack, so β has measure only where $\sigma = \int_A T(\omega, \alpha) d\rho$ and ρ has measure only where $\theta = \int_\Omega T(\omega, \alpha) d\beta$, then 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 β is a step function). Problem (3.3) is a linear infinite one, i.e., from ILP (infinite linear programming), an SILP one up to the infinite dimensions of ρ space. Because of this insight and problem, and in view of the compactness of the feasible (index) sets at the lower levels, A and Ω , we are interested in the *nondegeneracy* of the local minima of the lower level problem to get *finitely* many local minimizers (Weber, 1992). We note that on the lower levels, θ and σ are just shift terms which do not affect the local solutions there. Let us focus on problem (3.3), employ the language of bilevel programming known from SIP, introduce the function $g((\sigma, \rho), \omega) := \sigma - \int_\Omega T(\omega, \alpha) d\rho$ parametric in (σ, ρ) , and state the

Lower Level Problem: For a given parameter (σ, ρ) we consider

$$\min_{\omega} g((\sigma, \rho), \omega) \quad \text{subject to } \omega \in \Omega, \quad (3.4)$$

denote the defining inequality constraint functions of Ω by $v_1(\omega) := \omega$, $v_2(\omega) := -\omega + 1$, write $L := \{1, 2\}$ and $L_0(\bar{\omega}) := \{l \in L \mid v_l(\bar{\omega}) = 0\}$. Consequently, for any critical point $\bar{\omega}$, the Lagrange function reads $L(\omega) := L((\sigma, \rho), \omega, \gamma) := g((\sigma, \rho), \omega) - \sum_{l \in L_0(\bar{\omega})} \gamma_l v_l(\omega)$. Since Ω is compact,

local (and global) minimizer(s) $\bar{\omega}$ of (3.4) exists. We analyze the conditions of the *nondegeneracy* and, hence, *reduction ansatz* (Hettich, 1982), at any such an $\bar{\omega}$.

1. **LICQ:** $D_{\omega}v_l(\bar{\omega})$ ($l \in L_0(\bar{\omega})$) is a family with not more than 1 element and $D_{\omega}v_1(\bar{\omega})=1$ and $D_{\omega}v_2(\bar{\omega})=-1$ do not vanish; hence, it is linearly independent.

2. **Kuhn-Tucker (KKT) condition:** There exists a multiplier $\bar{\gamma} \in \mathbb{R}^{L_0(\bar{\omega})}$ such that $D_{\omega}L(\bar{\omega})=0$ and $\bar{\gamma}_l > 0$ ($l \in L_0(\bar{\omega})$). If we rewrite $g((\sigma, \rho), \omega)$, we get:

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

Then Lagrange function is parametric in (σ, ρ) , and fully it looks as follows: $L((\sigma, \rho), \omega, \gamma) = \sigma - \frac{1}{2} \sum_{i,j=1}^N \kappa(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) + \int_A \sum_{i=1}^N \alpha_i d\rho(\alpha) - \sum_{l \in L_0(\bar{\omega})} \gamma_l v_l(\omega)$. Let us find the conditions which satisfy the KKT condition $D_{\omega}L((\sigma, \rho), \omega, \gamma) = 0$, where $D_{\omega}L((\sigma, \rho), \omega, \gamma) = D_{\omega}Z - D_{\omega}\left(\sum_{l \in L_0(\bar{\omega})} \gamma_l v_l(\omega)\right)$, $Z := -\frac{1}{2} \sum_{i,j=1}^N \kappa(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha)$, hence, $D_{\omega}Z = -\frac{1}{2} \sum_{i,j=1}^N D_{\omega}\kappa(x_i, x_j, \omega) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha)$. If we assume that we have a *Gaussian kernels*, i.e.,

$$\kappa(x_i, x_j, \omega) = \exp\left(-\omega \|x_i - x_j\|_2^2\right), \text{ if we denote } I(l \in L_0(\bar{\omega})) = \begin{cases} 1, & \text{if } l \in L_0(\bar{\omega}) \\ 0, & \text{if } l \notin L_0(\bar{\omega}) \end{cases} \text{ at some}$$

critical point $\bar{\omega}$, then we get $D_{\omega}Z = \frac{1}{2} \sum_{i,j=1}^N \|x_i - x_j\|^2 \exp\left(-\omega \|x_i - x_j\|^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha)$ and $D_{\omega}\left(\sum_{l \in L_0(\bar{\omega})} \gamma_l v_l(\omega)\right) = I(1 \in L_0(\bar{\omega})) \cdot \gamma_1 - I(2 \in L_0(\bar{\omega})) \cdot \gamma_2$. Now, we come back to our KKT conditions and evaluate $D_{\omega}Z = -I(1 \in L_0(\bar{\omega})) \cdot \gamma_1 + I(2 \in L_0(\bar{\omega})) \cdot \gamma_2$:

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

$$\begin{aligned} &\frac{1}{2} \sum_{i,j=1}^N \|x_i - x_j\|^2 \exp\left(-\omega \|x_i - x_j\|^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) = \gamma_1, \\ \gamma_1 > 0 &\Leftrightarrow \frac{1}{2} \sum_{i,j=1}^N \|x_i - x_j\|^2 \exp\left(-\omega \|x_i - x_j\|^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) > 0. \end{aligned} \quad (3.5)$$

Case 2: If $v_2(\bar{\omega}) = 0$, i.e., $2 \in L_0(\bar{\omega})$, our equation becomes

$$\begin{aligned} &\frac{1}{2} \sum_{i,j=1}^N \|x_i - x_j\|^2 \exp\left(-\omega \|x_i - x_j\|^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) = -\gamma_2, \\ \gamma_2 > 0 &\Leftrightarrow \frac{1}{2} \sum_{i,j=1}^N \|x_i - x_j\|^2 \exp\left(-\omega \|x_i - x_j\|^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) < 0, \end{aligned} \quad (3.6)$$

For the following, we introduce $\bar{\gamma} := \begin{cases} \gamma_1, & \text{if case 1 holds} \\ \gamma_2, & \text{if case 2 holds} \end{cases}$.

3. **Second Order Condition (SOC):** With our value $\bar{\gamma}$ introduced, it is fulfilled: $\eta^T D_{\omega}^2 L(\bar{\omega}, \bar{\gamma}) \eta > 0 \quad \forall \eta \in T(\bar{\omega}) \setminus \{0\}$, where $T(\bar{\omega}) := \{\eta \in \mathbb{R}^r \mid D_{\omega}v_l(\bar{\omega})\eta = 0 \ (l \in L_0(\bar{\omega}))\}$.

For SOC, $D_{\omega}^2 L(\bar{\omega}, \bar{\gamma}) = -\frac{1}{2} \sum_{i,j=1}^N \|x_i - x_j\|^4 \exp\left(-\omega \|x_i - x_j\|^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha)$ should be nonnegative to satisfy nondegeneracy. Thus, $\bar{\omega}$ is non-degenerate if and only if the sign

conditions (on the multipliers) and $\frac{1}{2} \sum_{i,j=1}^N \|x_i - x_j\|^4 \exp\left(-\omega \|x_i - x_j\|^2\right) y_i y_j \int_A \alpha_i \alpha_j d\rho(\alpha) < 0$ are fulfilled. We underline that this essentially depends on the data given. Finally, we note that solving the dual problem (3.3) reduces the dimension in the lower level from N to 1.

4. Conclusions

The method we proposed in this study leads to the selection of kernels from infinite space Ω which enabled to enrich learning process SVM under the range interval $[0,1]$ of ω . Hence, we are not limited to choose kernel parameter(s), Gaussian kernels in our case, to discrete values with cross validation method so 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 finitely problem, except of the fact that measures are the state variables. By focussing on measures which possess a Radon-Nikodym density, we turn to a space of functions (Wan, 2007), and by looking at parametric *density* functions, e.g., with parameters (μ, σ^2) we get semi-infinite and, via reduction ansatz, a finite program indeed. Besides of that ansatz, also discretization and exchange methods are parts of our future studies.

Acknowledgement: The authors thank the professors E. Anderson, M. Goberna and J. Shawe-Taylor for their valuable advice.

References

- Anderson, E.J., and Nash, P. (1987), *Linear Programming in Infinite-Dimensional Spaces*, John Wiley and Sons Ltd.
- Apostol, T. M (1974), *Mathematical Analysis: A Modern Approach to Advanced Calculus*, Addison Wesley.
- Bach, F.R. and Lanckriet, G.R.G. (2004), “Multiple kernel learning, conic duality, and the smo algorithm”, *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada.
- Cristianini, N. and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press.
- Hettich, R., and Zencke, P. (1982), *Numerische Methoden der Approximation und semi-infiniten Optimierung*, Teubner, Stuttgart.
- Goberna, M.A. and Lopez, M.A., (1998), *Linear Semi-Infinite Optimization*. John Wiley and Sons Ltd.
- Lanckriet, G.R.G., Bie, T.D., Cristianini, N., Jordan, M. and Noble, W. (2004a), “A statistical framework for genomic data fusion”, *Bioinformatics* 20, pp. 2626–2635.
- Lanckriet, G.R.G., Cristianini, N., Ghaoui, L. E., Bartlett, P. and Jordan, M.I. (2004b), “Learning the kernel matrix with semidefinite programming”, *J. Machine Learning Research* 5, pp. 27–72.
- Rakotomamonjy, A., Bach, F., Canu, S. and Grandvalet, Y. (2007), “More efficiency in multiple kernel learning”, *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR.
- Özögür-Akyüz, S., Shawe-Taylor, J., Weber, G.-W., and Ogel, Z.B. (2007), “Pattern analysis for the prediction of eukaryotic pro-peptide cleavage sites“, to appear in the special issue of *Discrete Applied Mathematics Networks in Computational Biology*.
- Sonnenburg, S., Raetsch, G., Schafer, C. and Schoelkopf, B. (2006), “Large scale multiple kernel learning”, *J. Machine Learning Research* 7, pp. 1531–1565.
- Weber, G.-W. (1992), *Charakterisierung Struktureller Stabilität in der nichtlinearen Optimierung*. Aachener Beiträge zur Mathematik 5, eds.: Bock, H.H., Jongen, H.Th., and Plesken, W., Augustinus publishing house (now: Mainz publishing house), Aachen.
- Weber, G.-W., (1993), *Minimization of a max-type function: Characterization of structural stability*. *Parametric Optimization and Related Topics III*, eds.: Guddat, J., Jongen, H.Th., Kummer, B., and Nozicka, F., Peter Lang publishing house, Frankfurt a.M., Bern, pp. 519–538, New York.
- Weber, G.-W., Taylan, P., Ozogur, S. and Akteke-Ozturk, B., (2007), “Statistical learning and optimization methods in data mining“, *Recent Advances in Statistics*, eds, Ayhan, H.O., and Batmaz, I., Turkish Statistical Institute Press, Ankara, pp. 181-195.
- Wan, Z., Wu, S.Y. and Teo, K.L. (2007), “Some Properties on quadratic infinite programs of integral type”, *Applied Mathematics Letters* 20, pp. 676-680.