J Glob Optim (2010) 48:215–239
DOI 10.1007/s10898-009-9488-x

# On numerical optimization theory of infinite kernel learning

**S. Özöğür-Akyüz · G.-W. Weber**

Received: 21 January 2009 / Accepted: 2 November 2009 / Published online: 20 November 2009

**Abstract** In Machine Learning algorithms, one of the crucial issues is the representation of the data. As the given data source become heterogeneous and the data are large-scale, multiple kernel methods help to classify "nonlinear data". Nevertheless, the finite combinations of kernels are limited up to a finite choice. In order to overcome this discrepancy, a novel method of "infinite" kernel combinations is proposed with the help of infinite and semi-infinite programming regarding all elements in kernel space. 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 parametrization in the space of probability measures, it becomes semi-infinite. We adapt well-known numerical methods to our infinite kernel learning model and analyze the existence of solutions and convergence for the given algorithms. We implement our new algorithm called "infinite" kernel learning (IKL) on heterogenous data sets by using exchange method and conceptual reduction method, which are well known numerical techniques from solve semi-infinite programming. The results show that our IKL approach improves the classifaction accuracy efficiently on heterogeneous data compared to classical one-kernel approaches.

This study is partially done at Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey and Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey.

S. Özöğür-Akyüz (✉)
Department of Mathematics and Computer Science, Bahçeşehir University, Istanbul, Turkey
e-mail: sureyya.akyuz@bahcesehir.edu.tr

G.-W. Weber
Institute of Applied Mathematics, Middle East Technical University, METU, Ankara, Turkey
e-mail: gweber@metu.edu.tr

G.-W. Weber
Faculty of Economics, Management Science and Law, University of Siegen, Siegen, Germany

**Keywords** Machine learning · Infinite kernel learning · Semi-infinite optimization · Infinite programming · Support vector machines · Continuous optimization · Discretization · Exchange method · Conceptual reduction · Triangulation

## 1 Introduction

In classical kernel learning methods, a single kernel is used to map the input space to a higher dimensional feature space. But for large scale and heterogeneous data in real-world applications [17,19], *multiple kernel learning (MKL)* is developed [4,21]. The main intuition behind *multiple kernel learning* is to combine finitely many pre-chosen kernels in a convex combination [21]

$$k_\beta(\mathbf{x}_i, \mathbf{x}_j) := \sum_{\kappa=1}^{K} \beta_\kappa k_\kappa(\mathbf{x}_i, \mathbf{x}_j). \tag{1}$$

In this paper, we shall basically refine the sum in (1) by an integral, as we shall closely explain. In [4], a multiple kernel reformulation is modeled by semi-definite programming for selecting the optimum weights of corresponding kernels. This reformulation has some drawbacks in computation time because of semi-definite programming and this reformulation is developed in [21] by semi-infinite linear programming with the following optimization model:

$$\begin{aligned} &\max_{\theta,\boldsymbol{\beta}} \ \theta \quad (\theta \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^K) \\ &\text{such that} \quad \boldsymbol{\beta} \geqslant 0, \ \sum_{\kappa=1}^{K} \beta_\kappa = 1, \\ &\sum_{\kappa=1}^{K} \beta_\kappa S_\kappa(\alpha) \geqslant \theta \ \ \forall \boldsymbol{\alpha} \in \mathbb{R}^l, \\ &0 \leqslant \boldsymbol{\alpha} \leqslant C\mathbf{1} \quad \text{and} \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \end{aligned} \tag{2}$$

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

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 nonlinearly distributed and large-scaled ones. A finite combination may fail, here. In order to overcome this, with the motivation of previous studies [3,12,14,15], a combination of *infinitely* many kernels in Riemann–Stieltjes integral form is proposed to allow an infinite wealth of possible choices of kernels in the kernel space which is called *infinite kernel learning* (*IKL*) [12,13,16]. This makes the problem infinite in both its dimension and its number of constraints; which is so-called *infinite programming* (*IP*). The mathematical foundations of IKL is established by mathematical analysis and the theory of semi-infinite programming in [12,13,16]. An infinite combination is represented by the following formula:

$$k_\beta(\mathbf{x}_i, \mathbf{x}_j) := \int_\Omega k(\mathbf{x}_i, \mathbf{x}_j, \omega) d\beta(\omega), \tag{3}$$

where $\omega \in \Omega$ is a kernel parameter and $\beta$ is a monotonically increasing function of integral 1, or just a probability measure on $\Omega$. Furthermore, the kernel function $k(\mathbf{x}_i, \mathbf{x}_j, \omega)$ is assumed to be a twice continuously differentiable function with respect to $\omega$, i.e., $k(\mathbf{x}_i, \mathbf{x}_j, \cdot) \in C^2$. The infinite combination can be, e.g., a combination of Gaussian kernels with different widths from a set $\Omega$, i.e., $k_\beta(\mathbf{x}_i, \mathbf{x}_j) = \int_\Omega \exp(-\omega \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2) d\beta(\omega)$. It is obvious that the Gaussian kernel is from a family of twice continuously differentiable functions of the variable $\omega$. Hereby, the wealth of infinitely many kernels is used to overcome the limitation of the kernel combination given by finitely many pre-chosen kernels. The questions on *which* combination of kernels and on the *structure* of the mixture of kernels can be considered and optimized, and it may, e.g., be answered by *homotopies* [12,13,16].

With this new formulation, we have the opportunity of recording ("scanning") all possible choices of kernels from the kernel space and, hence, the uniformity is also preserved. Let us note that infinitely many kernels correspond to infinitely many coefficients. Kernel coefficients are defined through an *increasing monotonic function* by means of *positive measures* [13,16].

The IKL formulation is given in [12,13,16] by

$$\begin{aligned}
&\max_{\theta,\beta}\ \theta \quad (\theta \in \mathbb{R},\ \beta\text{: a positive measure on } \Omega)\\
&\text{such that } \theta - \int\limits_\Omega T(\omega, \boldsymbol{\alpha}) d\beta(\omega) \leqslant 0 \quad (\boldsymbol{\alpha} \in A), \qquad (4)\\
&\int\limits_\Omega d\beta(\omega) = 1,
\end{aligned}$$

where $T(\omega, \boldsymbol{\alpha}) := S(\omega, \boldsymbol{\alpha}) - \sum_{i=1}^l \alpha_i$, $S(\omega, \boldsymbol{\alpha}) := \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j, \omega)$ and $\Omega := [0, 1]$ and $A := \{\boldsymbol{\alpha} \in \mathbb{R}^l \mid 0 \leqslant \boldsymbol{\alpha} \leqslant C\mathbf{1}$ and $\sum_{i=1}^l \alpha_i y_i = 0\}$ are our index sets.

We note that there are infinitely many inequality constraints because of the inequality constraint which are uniform in $\alpha \in A$, and the state variable $\beta$ is from an infinite dimensional space. Thus, our problem is a one of *infinite programming* (*IP*) [1]. The dual of (4) can be written as

$$\begin{aligned}
&\min_{\sigma,\rho}\ \sigma \quad (\sigma \in \mathbb{R},\ \rho\text{: a positive measure on } A)\\
&\text{such that } \sigma - \int\limits_A T(\omega, \boldsymbol{\alpha}) d\rho(\boldsymbol{\alpha}) \geqslant 0 \ (\omega \in \Omega), \qquad (5)\\
&\int\limits_A d\rho(\boldsymbol{\alpha}) = 1.
\end{aligned}$$

Because of the conditions $\int_\Omega d\beta(\omega) = 1$ and $\int_A d\rho(\boldsymbol{\alpha}) = 1$, positive measures $\beta$ (or $\rho$) are probability measures and these measures are parametrized in this paper via the probability density functions as in [13,16].

We note that the primal IKL formulation (4) and the dual IKL formulation (5) are very similar, where minimization is swapped with maximization, the direction of inequalities in the constraints are changed in (5), the index sets $A$, $\Omega$ and the variables $\boldsymbol{\alpha}$, $\omega$ are replaced with each other such that both index sets are compact and the objective functions, $\theta$, $\sigma$, of both the dual and the primal are continuous. Thus, similar formulations, definitions and the theorems (covergence, discretizations etc.) written for the primal problem can be expressed for the dual problem in terms of the variables and the index set of the dual problem. However, we observe that the primal and the dual problem are different as far as the way how the sets of inequality constraints are defined. In fact, there can be problems with nondegeneracy (that can be related with instability) just for one of the problems, not for the other. In [11–13,16], it is explained that the LICQ condition is violated because of the linear dependency of the equality and the inequality constraints on the lower level problem of the primal problem. In order to overcome this degenerate case, we perturbed the equality constraint of (2) with a monotonically decreasing sequence $(\xi_\nu)_{\nu\in\mathbb{N}}$ such that $\xi_\nu \to 0$ $(\nu \to \infty)$, where, $\sum_{i=1}^{l} \alpha_i y_i = \xi_\nu$.

**Corollary 1** [13,16] *Let us assume that there exist* $(\beta, \theta)$ *and* $(\rho, \sigma)$ *which are feasible for their respective problems, and are complementary slack, i.e.,*

$$\sigma = \int\limits_{A} T(\omega, \boldsymbol{\alpha}) d\rho(\boldsymbol{\alpha}) \quad \text{and} \quad \theta = \int\limits_{A} T(\omega, \boldsymbol{\alpha}) d\beta(\omega).$$

*Then,* $\beta$ *has measure only where* $\sigma = \int_A T(\omega, \boldsymbol{\alpha}) d\rho(\boldsymbol{\alpha})$ *and* $\rho$ *has measure only where* $\theta = \int_\Omega T(\omega, \boldsymbol{\alpha}) d\beta(\omega)$ *which implies that both solutions are optimal for their respective problems.*[1]

In this study, we restrict ourselves to probability measures, which constitute our subspace of positive measures, and we use parametrized models of IKL given by probability density functions *(pdfs)* in [13,16]. Throughout this study, we assume that we are given pdf function $f^{\mathcal{P}}(\omega; \cdot)$ for our primal problem. We do not need to write the equality constraint $\int_\Omega d\beta(\omega) = 1$, since we assume that our measures are probability measures. Then, we parametrize these measures via pdfs $f^{\mathcal{P}} = f^{\mathcal{P}}(\omega; \boldsymbol{\wp}^{\mathcal{P}})$, taking the place of positive measures $\beta$. Let us denote the parameters of a pdf by $\boldsymbol{\wp}^{\mathcal{P}} = (\wp_1^{\mathcal{P}}, \wp_2^{\mathcal{P}}, \ldots, \wp_{\iota^{\mathcal{P}}}^{\mathcal{P}})^T$ for the primal problem. It is constrained and elements of a suitable parameter set can be written as follows:

$$P^{\mathcal{P}} := \{\boldsymbol{\wp}^{\mathcal{P}} \in \mathbb{R}^{\iota^{\mathcal{P}}} \mid \mathrm{u}_i^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}}) = 0 \ \ (i \in I^{\mathcal{P}}), \ \mathrm{v}_j^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}}) \geq 0 \ \ (j \in J^{\mathcal{P}})\}.$$

We note that after a parametrization, our primal (or dual) problem has variables in finite dimension since instead of optimizing with respect to measure $\beta$ which is in an infinite dimensional space, we minimize with respect to the pdf parameter vector $\boldsymbol{\wp}^{\mathcal{P}}$. This parametrization allows us to define our infinite programming problem by *semi-infinite programming* (*SIP*) since the variables are in a finite dimension and there are infinitely many inequality constraints. Hence, our primal problem turns into the following SIP with additional constraint functions $\mathrm{u}_i^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}})$ and $\mathrm{v}_j^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}})$, coming from the definition of the parameter sets related to the specific pdf function of the primal problem [13,16]:

$$\begin{aligned} (\text{Primal SIP}) \ \min_{\theta, \boldsymbol{\wp}^{\mathcal{P}}} \quad & -\theta \\ \text{such that} \ \int\limits_{\Omega} T(\omega, \boldsymbol{\alpha}) f^{\mathcal{P}}(\omega; \boldsymbol{\wp}^{\mathcal{P}}) d(\omega) - \theta \geq 0 \ \ & (\boldsymbol{\alpha} \in A), \\ \mathrm{u}_i^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}}) = 0 \ \ (i \in I^{\mathcal{P}}), & \\ \mathrm{v}_j^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}}) \geq 0 \ \ (j \in J^{\mathcal{P}}). & \end{aligned} \tag{6}$$

## 2 Numerical analysis of IKL

One of the early methods mostly used to solve SIP problems in practice, e.g., in engineering applications, is *discretization* [7]. It is based on a selection of finitely many points from the

[1] Communication with E. J. Anderson.

infinite index set of inequality constraints. In our study, these infinite index sets are $A$ and $\Omega$ for the primal and the dual problems, respectively.

The discretized primal SIP problem of (6) can be written by the following formulations:

$$P(A_k) \qquad \min_{\theta, \wp^{\mathcal{P}}} -\theta$$
$$\text{subject to} \quad g^{\mathcal{P}}((\theta, \wp^{\mathcal{P}}), \boldsymbol{\alpha}) := \int_{\Omega} T(\omega, \boldsymbol{\alpha}) f^{\mathcal{P}}(\omega; \wp^{\mathcal{P}}) d\omega - \theta \geqslant 0 \ \ (\boldsymbol{\alpha} \in A_k), \tag{7}$$
$$\mathrm{u}_i^{\mathcal{P}}(\wp^{\mathcal{P}}) = 0 \ \ (i \in I^{\mathcal{P}}),$$
$$\mathrm{v}_j^{\mathcal{P}}(\wp^{\mathcal{P}}) \geq 0 \ \ (j \in J^{\mathcal{P}}).$$

Here, by the symbol $P(\cdot)$ we denote the primal, $k$ is the iteration step, and the discretized set $A_k$ will be discussed within Strategies I and II in Sect. 3. It is obvious that $\Omega_k$ can be defined by a one-dimensional uniform grid.[2] Hereby, $k$ should not be confused with our kernel function.

Let $v^{\mathcal{P}}(A_k)$, $\mathcal{M}^{\mathcal{P}}(A_k)$ and $\mathcal{G}^{\mathcal{P}}(A_k)$ denote the minimal value, the feasible set and the set of (global) minimizers of our primal problem (6) with $A$ replaced by $A_k$. Under suitable regularity conditions (reduction ansatz) [7], the optimal solutions of the lower level problems depend locally on the parameters, i.e., measures. Furthermore, the relation with the pdfs has been established with a dual pairing and by the pdfs as test kind of functions from the dual space.

Let $d_1$ be the *Hausdorff* distance $d_1(A_k, A)$ between $A$ and $A_k$, which is given by

$$d_1(A_k, A) := \max_{\mathbf{y} \in A} \min_{\mathbf{y}' \in A_k} \left\| \mathbf{y} - \mathbf{y}' \right\|_2 .$$

Now, with the Hausdorff distance, we will introduce the discretizability notion for our problems based on the definitions in [22]. In these problems, $\mathbf{y} = \boldsymbol{\alpha}$ and $\mathbf{y}' = \boldsymbol{\alpha}'$ for the primal case. In the following definitions, the distance to the solution $(\theta^*, \wp^{\mathcal{P}*})$ of the primal SIP will be defined by the Hausdorff distance, too. We note that the optimal solution of the primal problem exists because of the continuity of the objective functions and inequality constraints, and compactness of the feasible sets which is proposed subsequently in Closer Explanation 1 [24]. Here, we employ Theorem of Weierstrass. We denote the distance functions $d_1$ for the primal problem as $d_1^{\mathcal{P}}$.

**Definition 1** The primal problem (6) is called **finitely reducible** if there is a finite set $A_{k^0} \subset A$ for some $k = k^0$ such that $v^{\mathcal{P}}(A_{k^0}) = v^{\mathcal{P}}(A)$, and $(A_k)_{k \in \mathbb{N}_0}$ strictly isotonically increases[3] as $k \to \infty$.

**Definition 2** The primal problem (6) is called *weakly discretizable* if there exists a sequence of discretizations $(A_k)_{k \in \mathbb{N}_0}$ such that $v^{\mathcal{P}}(A_k) \to v^{\mathcal{P}}(A)$ $(k \to \infty)$.

We note that we have $v^{\mathcal{P}}(A_{k^1}) \leq v^{\mathcal{P}}(A_{k^2})$ if $A_{k^1} \subset A_{k^2}$ for our primal problem. Let us recall that we consider the standard form of primal SIP problems, i.e., a minimization problems. In closer detail, as the infinite index set grows, the number of inequality constraints increases. This forces the feasible set to become smaller at each iteration $k$. Thus, the minimum of the objective function increases (see Fig. 1).

[2] A uniform grid is discretization of a considered set where all elements $\mathbf{x} = (x_1, x_2, \ldots, x_l)^T$ have same spacing with respect to their $i$th coordinate $(i = 1, 2, \ldots, l)$. For example in $\mathbb{R}^2$, all rows have the same spacing and all of the columns have the same spacing (but not necessarily the same as the row spacing).

[3] A sequence $(A_k)_{k \in \mathbb{N}_0}$ is called strictly isotonically increasing if $A_k \underset{\neq}{\subset} A_{k+1}$ $(k \in \mathbb{N}_0)$.

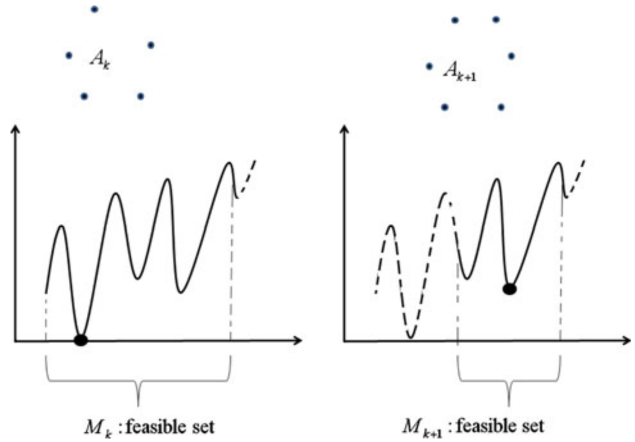


**Fig. 1** Symbolic illustration of the minimum values with respect to different feasible sets corresponding different discretizations; an example

**Definition 3** The primal problem, (6) is called *discretizable* if for each sequence of finite grids $A_k \subset A$ $(k \in \mathbb{N}_0)$ satisfying $d_1^{\mathcal{P}}(A_k, A) \to 0$ $(k \to \infty)$, where $d_1^{\mathcal{P}}(A_k, A) =: \max\limits_{\boldsymbol{\alpha} \in A} \min\limits_{\boldsymbol{\alpha}' \in A_k} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2$, there exist solutions $(\bar{\theta}_k, \bar{\boldsymbol{\wp}}_k^{\mathcal{P}})_{k \in \mathbb{N}_0}$ of the discretized primal problems (7) such that the following relations hold:

$$\min_{(\theta, \wp^{\mathcal{P}}) \in \mathcal{G}^{\mathcal{P}}(A)} \left\| (\bar{\theta}_k, \bar{\boldsymbol{\wp}}_k^{\mathcal{P}}) - (\theta, \wp^{\mathcal{P}}) \right\|_2 \to 0 \tag{8}$$

$$\text{and} \quad v^{\mathcal{P}}(A_k) \to v^{\mathcal{P}}(A) \;\; (k \to \infty).$$

**Corollary 2** *If the primal problem* (6) *is finitely reducible, then it is weakly discretizable.*

*Proof* Let us assume that (6) is finitely reducible. Then, by definition, there exist a $k_0 \in \mathbb{N}_0$ and finite sets $A_{k^0} \subset A$ such that $v^{\mathcal{P}}(A_{k^0}) = v^{\mathcal{P}}(A)$. Then, it is obvious that $v^{\mathcal{P}}(A_k) \to v^{\mathcal{P}}(A) \;\; (k \to \infty)$. □

Under the discretizability notion established above, we introduce the conceptual discretization algorithm in the following section.

## 3 Conceptual discretization method

The *conceptual discretization* method is based on an update of the discretization according to some stopping criterion for the convergence of the optimal solution. We adapt the conceptual discretization method [7,8,22] to our primal problem in Algorithm 1.

In Algorithm 1, the stopping criterion is *theoretically* established since one needs to check, e.g., $g^{\mathcal{P}}((\theta_k, \boldsymbol{\wp}_k^{\mathcal{P}}), \boldsymbol{\alpha}) \geq -\delta$ $(\alpha \in A)$. Alternatively, we introduce some stopping criterion based on the idea of a *Cauchy* sequence.

Generally speaking, in our problem and many real-world situations, an optimal solution is not known. In order to stop at a sufficiently close approximately optimal solution, the increment between the steps has to be small enough, i.e., $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 < \epsilon_0$ for a fixed $\epsilon_0 > 0$ which comes from the definition of "Cauchy sequence" evaluated at the $k$-th iteration

**Algorithm 1 Primal Conceptual Discretization Method (PCDM)**

```
Input:
    $\delta$          positive number, i.e., $\delta > 0$
    $f^{\mathcal{P}}$         probability density function
    $P^{\mathcal{P}}$         the set where pdf parameters lie
Output:
    $\theta$          unknown variable for minimization, to be evaluated
    $\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}$         the parameter vector of the pdf
PCDM$\left(\theta, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}, A, \delta, f^{\mathcal{P}}, P^{\mathcal{P}}\right)$
$k := 0$
Initialize a discretization $A_k \subset A$.
DO Compute a solution $(\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}})$ of
        $\min_{\theta \in \mathbb{R}, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}} \; (-\theta)$
        subject to $g^{\mathcal{P}}((\theta, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}) \geqslant 0 \;\; (\boldsymbol{\alpha} \in A_k)$,
        $\mathrm{U}_i^{\mathcal{P}}(\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) = 0 \;\; (i \in I^{\mathcal{P}})$,
        $\mathrm{V}_j^{\mathcal{P}}(\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) \geq 0 \;\; (j \in J^{\mathcal{P}})$.
if $g^{\mathcal{P}}((\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}) \geq -\delta \;\; (\boldsymbol{\alpha} \in A)$ then
    STOP
else
    $A_{k+1} := A_k \cup \{$any finitely many further points from $A\}$
    $k := k + 1$
end if
END DO
```

for a fixed $\epsilon_0 > 0$. A second alternative stopping criterion is based on the idea of a Cauchy sequence again, but on the value of the objective function $F$; it is determined by looking at the decrement of the objective function at iterations by $(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})) < \epsilon_1$ for a fixed $\epsilon_1 > 0$. As a third alternative, the first and the second criteria are both integrated in a single criterion by $(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})) \left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2^{-1} < \epsilon_2$ for a fixed $\epsilon_2 > 0$.

In our problems, the objective functions are $F^{\mathcal{P}}(\theta, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) := -\theta$ and $F^{\mathcal{D}}(\sigma, \boldsymbol{\wp}^{\boldsymbol{\mathcal{D}}}) := \sigma$ for the primal and the dual problems, respectively. With this notion, we establish our stopping criteria in different forms. In the following, we refer to one of the stopping criteria for the primal and the dual problems:

### 3.1 Stopping criteria for the primal problem

$$
\begin{aligned}
\left\|(\theta_{k+1}, \boldsymbol{\wp}_{\boldsymbol{k+1}}^{\boldsymbol{\mathcal{P}}}) - (\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}})\right\|_2 &< \epsilon_0 \quad \text{for a fixed } \epsilon_0 > 0, \\
\left|-\theta_k - (-\theta_{k+1})\right| &< \epsilon_1 \quad \text{for a fixed } \epsilon_1 > 0, \\
(-\theta_k - (-\theta_{k+1})) \left\|(\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}}) - (\theta_{k+1}, \boldsymbol{\wp}_{\boldsymbol{k+1}}^{\boldsymbol{\mathcal{P}}})\right\|_2^{-1} &< \epsilon_2 \quad \text{for a fixed } \epsilon_2 > 0.
\end{aligned}
\tag{9}
$$

Next, we will give an important assumption for the following Theorem 2.

**Assumption 1** The feasible set $\mathcal{M}^{\mathcal{P}}(A)$ is compact.

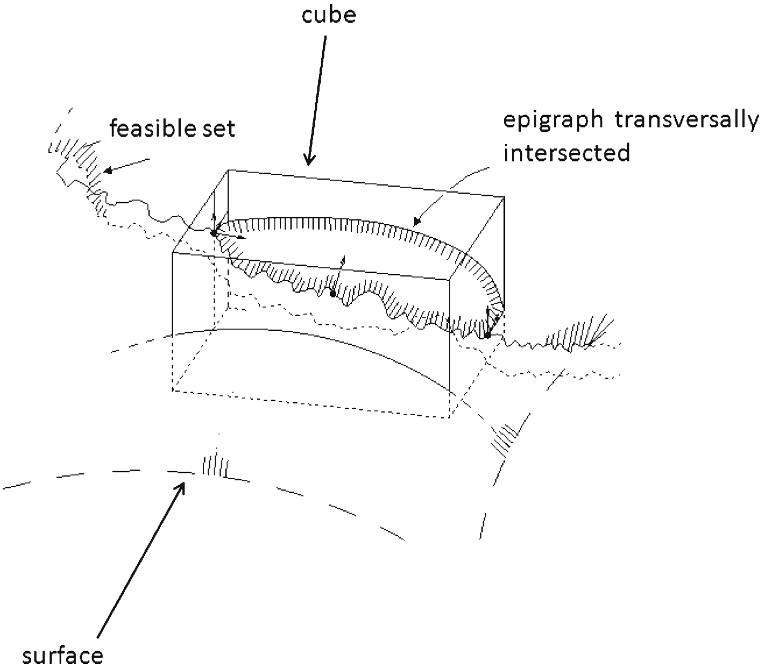


**Fig. 2** Transversal intersection (excision) of the feasible set with a box; an example taken from [18,24] (the surface may come from an equality constraint; the figure implies perturbational arguments of [24])

**Closer Explanation 1** *In fact, our (feasible) set satisfies compactness on the lower level but not on the upper level. Indeed, on the upper level, $\theta \in \mathbb{R}$ is unbounded for the primal problem* (6). *Let us recall that we parametrized $\beta$. We need a* **compact** *feasible set to have convergence of subsequences towards the optimal solution guaranteed, and also for the discretizability given in the following theorem. We encounter this problem by* **transversally** *intersecting the feasible set with sufficiently large transversal families of elementary geometrical sets (squares, boxes, cylinders or balls); this* **compactification** *is introduced in* [18,24].

*In an implicitly defined way, this corresponds to the following feasible subset of the primal SIP with some nonnegative (semi-continuous) functions* $\mathrm{G}^{\mathcal{P}}$:

$$\mathcal{M}^{\mathcal{P}}_{comp}(A) := \Big\{(\theta, \wp^{\boldsymbol{\mathcal{P}}}) \,\Big|\, \theta \in \mathbb{R},\ g^{\mathcal{P}}((\theta, \wp^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}) \geq 0 \ \ (\boldsymbol{\alpha} \in A),$$
$$(g^{\mathcal{P}} - \mathrm{G}^{\mathcal{P}})((\theta, \wp^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}) \leq 0 \ \ (\boldsymbol{\alpha} \in A)\Big\}, \tag{10}$$

*where $g^{\mathcal{P}}((\theta, \wp^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha})$ denotes the inequality constraint function of the primal problem. We note that the latter function may also be vector-valued.*

*Besides of this theoretical approach by function $\mathrm{G}^{\mathcal{P}}$, a more practical one consists of the idea of transversally cutting with a cube. This can be geometrically illustrated by the cube in Fig.* 2:

*Remark 1* When performing the transversal sections, it is important to take into account any *priori* information given about where a possible global solution, minimizer or maximizer, of our regarded optimization problem is located. Let us recall that we look at the primal and dual problems after parametrization, such that the parameters themselves become new decision

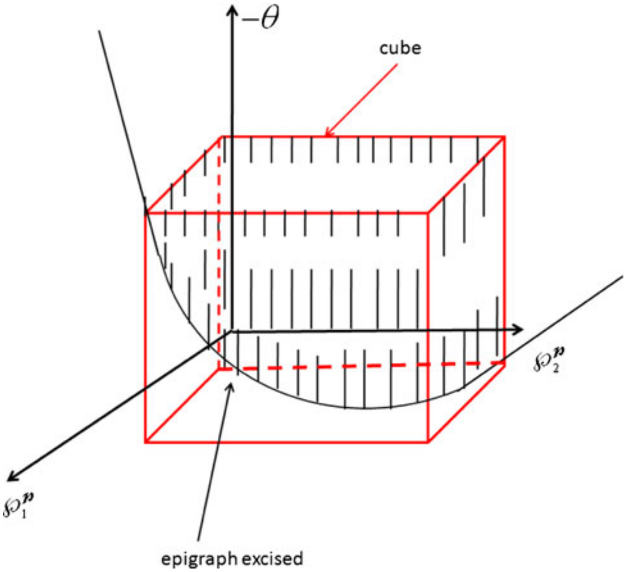


**Fig. 3** Illustration of the transversal cutting around the height function with a box; an example

variables. So we choose the intersecting parallelpipe large enough in order to include such an expected global solution. Of course, to gain that a priori knowledge, a careful analytical investigation may be helpful and should be done, e.g., in terms of growth behaviour and convexity kind of properties. In fact, for the ease of exposition, we just think of minimization rather than both minimization and maximization.

As a first, simple but important class of problems we mention such ones with a strictly convex graph (given by the constraints), i.e., an epigraph with the form of a potential, e.g., a paraboloid. In any such a case, we know that the lower level set with respect some arbitrary and sufficiently large level is nonempty and compact. Then, we can choose and raise our transversally cutting parallelpipe so that, in a projective sense, the lower level set and, hence, as an element, the global minimizer is contained in the parallelpipe and, therefore, in the excised subset of the epigraph.

This treatment and careful geometrical arrangement guarantees the equality of set of minimizers of the original problem, $\mathcal{G}^{\mathcal{P}}(A)$, and the set of minimizers after compactification, $\mathcal{G}^{\mathcal{P}}_{comp}(A)$, i.e., $\mathcal{G}^{\mathcal{P}}(A) = \mathcal{G}^{\mathcal{P}}_{comp}(A)$ which is illustrated in Fig. 3.

Let us underline that our strict convexity is not guaranteed in general. In fact, the fulfillment of this property on the one hand depends on how the kernel functions are chosen and how the kernel matrices are evaluated at the input data. On the other hand, it depends on how the parameters are involved into the density functions and how the possible nonlinearity can be characterized by convexity and the growth kinds of conditions, e.g., in terms of Morse indices [9].

We can adopt this idea to our problem in order to transversally cut around our height function on the boundary of the epigraph with a cube, as shown by Fig. 3.

Under our Closer Explanation 1, we obtain a general convergence result for this method based on Theorem 13 in [22].

**Theorem 2** *Let Assumption 1, after the compactification introduced in Closer Explanation 1 be satisfied, and let the primal problem and the sequences of discretizations* $(A_k)_{k\in\mathbb{N}_0}$ *and*

$(\Omega_k)_{k\in\mathbb{N}_0}$ *satisfy*

$$A_0 \subset A_k \ \ (k \in \mathbb{N}_0) \ \ \textit{and} \ \ d_1^{\mathcal{P}}(A_k, A) \to 0 \ \ \textit{for} \ \ k \to \infty.$$

*Based on possible compactifications, we may from now on suppose that* $\mathcal{M}(A_0)$ *and* $\mathcal{M}(\Omega_0)$ *are compact. Then, the primal problem,* (6) *is* **discretizable**, *i.e., the problem* $P(A_k)$ $(k \in \mathbb{N}_0)$ *has solutions* $(\theta_k, \boldsymbol{\wp}_k^{\mathcal{P}})$*, and such sequences of iterative solutions satisfy*

$$\min_{(\theta^*, \boldsymbol{\wp}^{\mathcal{P}*}) \in \mathcal{G}^{\mathcal{P}}(A)} \left\| (\theta_k, \boldsymbol{\wp}_k^{\mathcal{P}}) - (\theta^*, \boldsymbol{\wp}^{\mathcal{P}*}) \right\|_2 \to 0 \ \ (k \to \infty). \tag{11}$$

We refer to [22] for the proof of Theorem 2. By this theorem, we guarantee the convergence of approximate solutions to optimal solutions for sufficiently large $k$ with (11).

**Closer Explanation 3** *We note that the assumptions of Theorem* 2 *should be satisfied before we discretize our infinite index set. We know that our index set* $A$ *is compact, and we assume that our sequences of discretized set* $A_k$ $(k \in \mathbb{N}_0)$ *converge to* $A$*. Then, our semi-infinite problem is discretizable.*

*We also note that the minima which are stated in the theorem exist since the Euclidean norm is continuous and bounded from below, and, indeed, always nonnegative. Other properties used here are the existence of optimal solution* $(\theta^*, \boldsymbol{\wp}^{\mathcal{P}^*})$*, i.e., the set of minimizers* $\mathcal{G}^{\mathcal{P}}(A)$ *exists for the primal problem, since our feasible set is compact and the objective function is continuous, and we use that the set* $\mathcal{G}^{\mathcal{P}}$ *is compact, too, Theorem of Weierstrass* (*see* [2]).

Next, we give the definition for the *local* primal which is defined around some open neighbourhoods of the local minimizers.

**Definition 4** [22] Given a local minimizer $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})$ of the primal problem (6), the primal SIP is called **locally discretizable** at $(\bar{\theta}, \boldsymbol{\wp}^{\mathcal{P}})$ if the discretizability relation holds locally, i.e., if there exist neighbourhoods $U_{(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})}$ of $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})$ such that the locally discretized problem $P^{loc}(A)$ for the primal problem, namely,

$$\begin{aligned} P^{loc}(A): \quad & \min_{(\theta, \boldsymbol{\wp}^{\mathcal{P}}) \in U_{(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})}} \quad -\theta \\ & \text{subject to} \ \ \int_\Omega T(\omega, \boldsymbol{\alpha}) f^{\mathcal{P}}(\omega, \boldsymbol{\wp}^{\mathcal{P}}) d\omega - \theta \geqslant 0 \ \ (\boldsymbol{\alpha} \in A), \\ & \qquad\qquad\quad \mathrm{U}_i^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}}) = 0 \ \ (i \in I^{\mathcal{P}}), \\ & \qquad\qquad\quad \mathrm{V}_j^{\mathcal{P}}(\boldsymbol{\wp}^{\mathcal{P}}) \geq 0 \ \ (j \in J^{\mathcal{P}}) \end{aligned}$$

obtained as the restriction of $P(A)$ to open neighbourhood $U_{(\bar{\theta}, \bar{\wp}^{\mathcal{P}})}$, is discretizable.

The following Theorem 4 is based on Theorem 15 given in [22], and it gives a convergence result for the discretization method applied to our problems. Let us recall the definition of a local minimum of order $p$ before giving the result.

**Definition 5** A feasible point $\bar{\mathbf{x}}$ is called a *local minimizer of order* $p > 0$ of the problem to minimize $f(\mathbf{x})$ on a feasible set $\mathcal{M} \subseteq \mathbb{R}^n$ if with suitable constants $\epsilon > 0$ and $M > 0$, the following relation holds:

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) \geq M \left\| \mathbf{x} - \bar{\mathbf{x}} \right\|_2^p \ \ \text{for all } \mathbf{x} \in \mathcal{M} \text{ with } \left\| \mathbf{x} - \bar{\mathbf{x}} \right\|_2 < \epsilon.$$

**Theorem 4** *Let $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})$ be a local minimizer of the primal problem (6) of order $p$, and let sets $\mathcal{M}(A_k)$, $\mathcal{M}(A)$, be restricted to a compact subset $K \subset \mathbb{R}^n$. We further suppose that the Mangasarian–Fromovitz Constraint Qualification (MFCQ) (see [7]) holds at $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})$. Then, the problem (6) is locally discretizable at $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})$. In closer detail: There is some $\varsigma^{\mathcal{P}} > 0$ such that for any sequences of grids $(A_k)_{k \in \mathbb{N}_0} \in A^{\mathbb{N}_0}$ with $d_1^{\mathcal{P}}(A_k, A) \to 0$ $(k \to \infty)$ and for any sequences of solutions $(\theta_k, \boldsymbol{\wp}_k^{\mathcal{P}})_{k \in \mathbb{N}_0}$ of the locally restricted problem $P^{loc}(A)$, the following relation holds:*

$$\left\| (\theta_k, \boldsymbol{\wp}_k^{\mathcal{P}}) - (\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}}) \right\|_2 \leq \varsigma^{\mathcal{P}} d_1^{\mathcal{P}}(A_k)^{1/p} \quad (k \to \infty)$$

**Closer Explanation 5** *The result of Theorem 4 is true for the global minimization problem (6) since the sets $\mathcal{M}(A_k)$ and $\mathcal{M}(A)$ are restricted to a compact subset [22]. We note that after compactification by transversally intersecting the feasible set with sufficiently large transversal elementary geometrical sets (see Closer Explanation 1), we satisfy the compactness assumption for Theorem 4.*

*Let us observe that the sets $A$ and $\Omega$ are compact. We recall that the discretization of $\Omega$ may simply be a one-dimensional grid, and the elements of the discretized set of $A$ may consist of a combination of its corner points, which will be explained later in this section. All the discretized sets are further refined based on the previous sets, i.e., $A_k \subset A_{k+1}$ $(k \in \mathbb{N}_0)$. The refinement of the following iterations depends on the type and the dimension of the set. For example, if the index set $Y$ is an interval $\Omega := [a, b]$, then a one-dimensional grid $\hat{Y}$ can be chosen such that the distance between neighbouring grid points is defined by $\Delta y_i := \frac{b-a}{k_0}$ $(i = 0, 1, \ldots, k_0)$ for some $k_0 \in \mathbb{N}$, and with the grid $\hat{Y} := \{y_i \mid y_i = a + i\Delta y,\ i = 0, 1, \ldots, k_0\} \subset [a, b]$. We can refine $\hat{Y}$ by updating $k_0$ such that $k_1 = k_0 + 1$.*

Until now, we have provided theorems which guarantee convergence of the discretization method under some assumptions. If the dimension of the continuous index variable is larger than 2, then the computational complexity of the discretization grows exponentially. In fact, we need an $(l-1)$-dimensional grid of the index set. For example, we use a grid of $[0, C]^l$ for the vector $\boldsymbol{\alpha}$ in our primal problem for the discretization of the index set $A$. The size of the mesh grows fastly as the dimension $l$ increases. In closer detail: For our primal problem (6), the infinite index variable $\boldsymbol{\alpha}$ is lying in an $l$-dimensional underlying space. Moreover, the dimension of the elements in $A$ is the same as the number of the training points used in our SVM which forces the index variable to a high dimension as the number of the training points increases. This makes the discretization algorithmically more difficult. Let us observe that the set $A$ is an $(l-1)$-dimensional *polytope*, indeed, it is the intersection of the hyperplane $\sum_{i=1}^{l} \alpha_i y_i = 0$ with the box constraints $0 \leq \alpha_i \leq C$ $(i = 1, 2, \ldots, l)$, as we learn from the definition of $A$.

We propose two strategies to find a discretization of the set $A$. The first Strategy I is based on an interpretation of the set $A$ by the combination of its corner points. In this way, we can discretize the standard simplex instead of the set $A$ directly. The second Strategy II is based on the linearization of the set $A$, which is established on theoretical foundations [24].

*Strategy I* [10] (**Triangulation**):
In this first strategy, we use Lemma of *Carathedory* given to represent the elements of $A$ by its corner points. Furthermore, we apply a triangulation for some standard simplex $\Delta^N$ and, hence, a discretization of $A$ will be inherited via $\Delta^N$. To do this, we transform the polytope $A$ to the standard simplex and doing a normalization by representing the coordinates of $A$

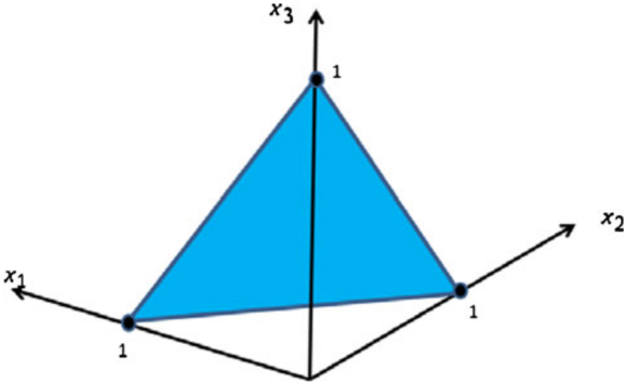


**Fig. 4** Illustration of the 2-simplex in $\mathbb{R}^3$

with its barycentric coordinates. After Example 1, we will explain how the triangulation is refined stepwise in an algorithmic way. Let us define the standard simplex and the relation with barycentric coordinates:

**Definition 6** For any $N \in \mathbb{N}_0$, let the **standard** $N$**-simplex** (or unit $N$-simplex) be given by

$$\Delta^N := \left\{ \boldsymbol{a} \in \mathbb{R}^{N+1} \;\middle|\; a_i \geq 0 \;\; (i = 1, 2, \ldots, N+1), \;\; \sum_{i=1}^{N+1} a_i = 1 \right\}.$$

The simplex $\Delta^N$ is lying in the affine hyperplane obtained by removing the restrictions $a_i \geq 0$ $(i = 1, 2, \ldots, N+1)$ in the above definition.

The vertices of the standard $N$-simplex are the standard unit-vectors (points)

$$\begin{aligned} \boldsymbol{e_0} &= (1, 0, 0, \ldots, 0)^T, \\ \boldsymbol{e_1} &= (0, 1, 0, \ldots, 0)^T, \\ &\vdots \\ \boldsymbol{e_N} &= (0, 0, 0, \ldots, 1)^T. \end{aligned}$$

There is a canonical map from the standard $N$-simplex to an arbitrary $N$-simplex (polytope) $\hat{\Delta}^N$ with vertices $\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_N}$, given by

$$\boldsymbol{a} \mapsto \hat{\boldsymbol{a}} := \sum_{i=1}^{N+1} a_i \boldsymbol{v_i} \quad (\boldsymbol{a} = (a_1, a_2, \ldots, a_{N+1})^T \in \Delta^N).$$

The coefficients $a_i$ are called the **barycentric coordinates** of a point $\hat{\boldsymbol{a}}$ in the $N$-simplex $\hat{\Delta}^N$ $(i = 1, 2, \ldots, N+1)$. The standard 2-simplex in $\mathbb{R}^3$ is illustrated in Fig. 4.

**Closer Explanation 6** *In order to apply the canonical mapping with barycentric coordinates, we assume $A = \hat{\Delta}^N$, $N+1$ is the number of vertices of $A$ and all vertices of $A$ have entries never different from 0 and $C$. Then, we can benefit from representing the points $\boldsymbol{\alpha} \in A$ by its barycentric coordinates and by the vertices of standard simplex or, as we will use below, we may assume all components $\alpha_i$ $(i = 1, 2, \ldots, l)$ to be* 0 *or $C$, respectively* [10].

Let us fix $y_i \in \{\pm 1\}$ $(i = 1, 2, \ldots, l)$ being the output data (labels) and recall the index set $A = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^l \,\middle|\, 0 \leq \alpha_i \leq C \;\; (i = 1, 2, \ldots, l) \text{ and } \sum_{i=1}^{l} \alpha_i y_i = 0 \right\}$.

Without loss of generality, we assume that there is some $i_0 \in \{1, 2, \ldots, l-1\}$ such that $y_1 = \cdots = y_{i_0} = 1$ and $y_{i_0+1} = \cdots = y_l = -1$. Furthermore, as prepared in our Closer Explanation 6 for simplicity, we take $C = 1$ for this strategy. (We could also choose $C$ different than 1; in fact, we can apply the same procedure below.) Since $\sum_{i=1}^{l} \alpha_i y_i = 0$, we have the following equation from the definition of the set $A$:

$$\alpha_1 + \cdots + \alpha_{i_0} = \alpha_{i_0+1} + \cdots + \alpha_l, \tag{12}$$

where $\alpha_i \in \{0, 1\}$ $(i \in \{1, 2, \ldots, l\})$. Specifically, the trivial solution to the Eq. 12 is a vertex of our polytope $A$. By this intuition, we will consider the elements of polytope $A$ by the combination of its binary vertices.

*Remark 2* The polytope $A$ has finitely many corner points. In particular, let $r := \min\{i_0, l - i_0\}$. Then, $A$ has $\sum_{i=0}^{r} \binom{i_0}{r}\binom{l-i_0}{r}$ corner points.

*Example 1* Let $l = 6$, $y_1 = y_2 = 1$, $y_3 = \cdots = y_6 = -1$. Then,

$$\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6. \tag{13}$$

There are 15 different corner points. The trivial one is $(0, 0, \ldots, 0)^T$, which corresponds to the number $\binom{2}{0}\binom{4}{0} = 1$.

We observe that we must have corner points with two nonzero elements or four nonzero elements to satisfy Eq. 13. Let us start with the corners having two nonzero elements:

$$(1,0,1,0,0,0),\;\; (1,0,0,1,0,0), \tag{14}$$

$$(1,0,0,0,1,0),\;\; (1,0,0,0,0,1), \tag{15}$$

$$(0,1,1,0,0,0), \ldots, (0,1,0,0,0,1), \tag{16}$$

$$(1,1,1,1,0,0), \ldots, (1,1,0,0,1,1), \tag{17}$$

where (14), (15) and (16) represent $\binom{2}{1}\binom{4}{1} = 8$ many points, and (17) corresponds to $\binom{2}{2}\binom{4}{2} = 6$ many ones. Then, the total number of corner points is $1 + 2 \cdot 4 + 1 \cdot 6 = 15$.

### 3.2 Algorithmic way to find all vertices (or corner points) of $A$

Let $\boldsymbol{p} \in A$ be any point. Indeed, for the ease and completeness of explanation, we may assume that $\boldsymbol{p}$ is an interior point of $A$, especially, not a corner point. Now, we choose a line $d$ through $\boldsymbol{p}$ in $A$. We take two points $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$ on $d$ which lie on the opposite sides of $\boldsymbol{p}$ and maximize the distance to $\boldsymbol{p}$. Then, $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$ must be on same hypersurfaces (hyperfaces) bounding the convex region $A$. Next, choose a line $d_2$ through $\boldsymbol{q_1}$ which lies in the hypersurface containing $\boldsymbol{q_1}$. This line intersects that face into two parts. The face has one more *codimension* (one less dimension). The point $\boldsymbol{q_1}$ is a convex combination of the two new intersection points. Continuing this way finishes the construction principle.

We illustrate the intuition of this algorithmic way of finding corners of polytope $A$ with Fig. 5. Obviously, $\boldsymbol{p}$ is a convex combination of $\boldsymbol{q_3}$, $\boldsymbol{q_4}$ and $\boldsymbol{q_5}$, the vertices of $A$.
Now, let $N := \sum_{i=0}^{r} \binom{i_0}{r}\binom{l-i_0}{r}$. Then, we can discretize the *standard simplex* in $\mathbb{R}^{N+1}$ and finally map it onto $A$ to discretize $A$. More formally, we firstly recall Definition 6,

$$\Delta^N = \left\{ \boldsymbol{a} \in \mathbb{R}^{N+1} \,\middle|\, a_i \geq 0 \;\; (i = 1, 2, \ldots, N+1),\; \sum_{i=1}^{N+1} a_i = 1 \right\}. \tag{18}$$

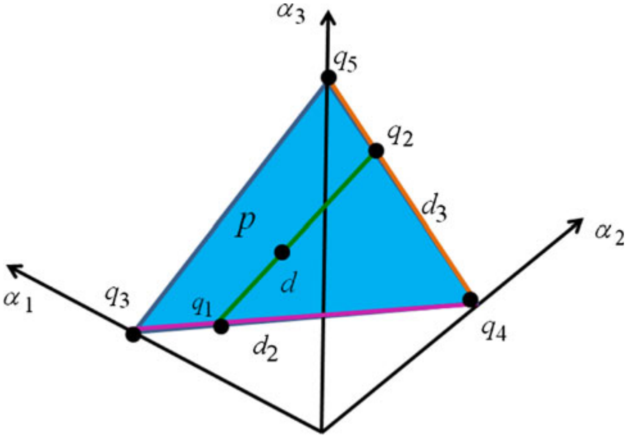


**Fig. 5** Illustration of the algorithmic way of finding corner points of $A$; an easy example for $l = 3$ (in $\mathbb{R}^3$)

Now, let us define a mapping

$$T\colon \Delta^N \longrightarrow A \text{ with } T(\boldsymbol{a}) := \sum_{i=1}^{N+1} a_i \boldsymbol{v_i} \in A \quad (\boldsymbol{a} = (a_1, a_2, \ldots, a_{N+1})^T \in \Delta^N),$$

where the set $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_{N+1}}\}$ consists of the vertices of $A$. By this methodology, we can find the elements of this discretization $A_k$ of $A$ which are represented by a combination of vertices of the simplex. This can be mathematically formulated as follows. Any point $\boldsymbol{p} \in A$ can be represented by

$$\boldsymbol{p} = \sum_{i=1}^{N+1} a_i \boldsymbol{v_i}, \tag{19}$$

where the set $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_{N+1}}\}$ is the collection of vertices of $A$ and $a_i$ $(i = 1, 2, \ldots, N+1)$ are the barycentric coordinates for $A$ (see Definition 6). To be able to write a point $\boldsymbol{p}$ from $A$ as in (19), we need to find the coordinates $a_i$ $(i = 1, 2, \ldots, N+1)$ from the standard $N$-simplex. Hence, the simplex $\Delta^N$ has to be discretized.

One of the main advantages of this strategy consists in working with the standard simplex and its vertices. However, the discretization of the simplex is **not** uniform because of the unsymmetries of the grid points. As it is clear from Fig. 6, the distances of the neighbouring mesh points are nonuniform, i.e., $\Delta_1 \neq \Delta_2 \neq \Delta_3 \neq \Delta_4$.

In order to overcome nonuniformity, we propose a method which transforms the barycentric coordinates of polytopes to a sphere as shown by Fig. 7 (for closer information, see [25]). Let us consider a particular face $F$ of some polytope and its corresponding spherical face $F'$ as shown in Fig. 7. Each point in $F$ can be described by barycentric coordinate systems induced by vertices of $F$ after the triangulation as given above. Let us assume that we create a distribution of points inside $F$. We can obtain each of the points in this distribution by a linear interpolation between the vertices of our barycentric coordinates system. Similarly, the distribution on $F'$ can be obtained through the same steps of interpolation between the vertices of barycentric coordinate systems on the sphere [25]. Since we have a uniform

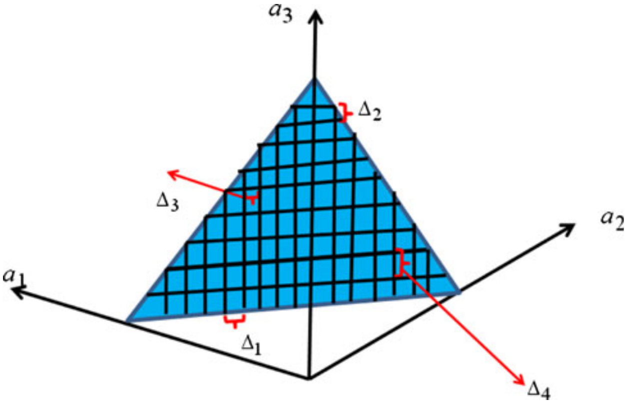


**Fig. 6** Nonuniform sampling of a standard simplex $\Delta^N$; an example in $\mathbb{R}^3$, $\Delta_1 \neq \Delta_2 \neq \Delta_3 \neq \Delta_4$

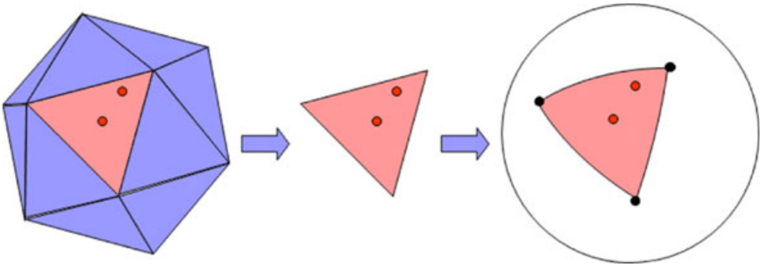
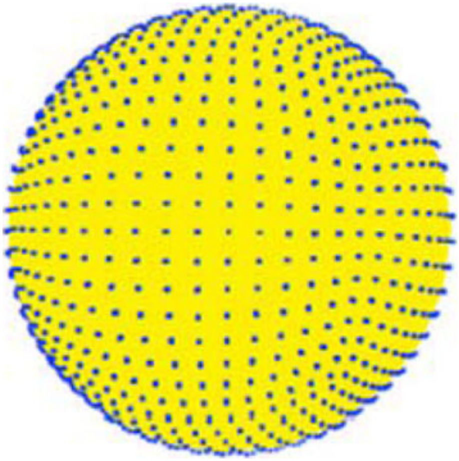


**Fig. 7** Transformation of the barycentric coordinates of a polytope to a sphere [25]

**Fig. 8** Discretization of the sphere; an example [25]



sampling over a sphere (see Fig. 8), we achieve a uniformly discretized set of points of our polytope $A$.

*Remark 3* It is important to observe the computational intractability of Strategy I because of the exponential growth of the corner points, as the dimension of $\boldsymbol{\alpha}$, i.e., the number of data points, increases. It is clear from the example that the number of binary vectors grows exponentially, namely, in the way of $2^l$, which makes the algorithm impractical. To make the algorithm practical, we do not search for all corners but we generate random corners in our implementations for exchange method which is introduced in Sect. 6.

We also propose a theoretically prepared second strategy which is based on a linearization procedure, and the implementation is left to future study.

Next, we propose a second strategy which is more theoretical.

*Strategy II* (**Linearization**):
The second strategy is based on the linearization of $A$ in some open neighbourhood $U_{(\bar{\theta},\bar{\boldsymbol{\wp}}^{\mathcal{P}})}$, locally around a given point $\bar{\boldsymbol{\alpha}} \in A$, e.g., a vertex of $A$. [24]. Mathematically, we define $\boldsymbol{z} = \hat{T}(\boldsymbol{\alpha})$ as follows:

$$\hat{T}\colon \begin{cases} z_1 & := u(\boldsymbol{\alpha}), \\ z_2 & := v_{\ell_1}(\boldsymbol{\alpha}), \\ & \vdots \\ z_{k+1} & := v_{\ell_k}(\boldsymbol{\alpha}), \\ z_{k+2} & := \boldsymbol{\zeta_1}^T(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}), \\ & \vdots \\ z_l & := \boldsymbol{\zeta_{l-1-k}}^T(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}), \end{cases} \tag{20}$$

where $k$ is the cardinality $|L_0(\bar{\boldsymbol{\alpha}})|$ of $L_0(\bar{\boldsymbol{\alpha}}) := \{\ell \in L \mid v_\ell(\bar{\boldsymbol{\alpha}}) = 0\}$, $L = \{1, 2, \ldots, 2l\}$. Let $L_0(\bar{\boldsymbol{\alpha}}) = \{\ell_1, \ell_2, \ldots, \ell_k\}$, $u(\boldsymbol{\alpha})$ and $v(\boldsymbol{\alpha})$ be defining equality and inequality constraints of the index set $A$ defined by $u(\boldsymbol{\alpha}) := \sum_{i=1}^{l} \alpha_i y_i$ and $v_r(\boldsymbol{\alpha}) := \alpha_r$, $v_s(\boldsymbol{\alpha}) := -\alpha_{l-s} + C$ ($r \in \{1, 2, \ldots, l\}$, $s \in \{l+1, l+2, \ldots, 2l\}$), and let the vectors $\boldsymbol{\zeta_\nu} \in \mathbb{R}^l$ ($\nu = 1, 2, \ldots, l-1-k$) complete the set $\{\nabla u(\bar{\boldsymbol{\alpha}})\} \cup \{\nabla v_\ell(\bar{\boldsymbol{\alpha}}) \mid \ell \in L_0(\bar{\boldsymbol{\alpha}})\}$ to a basis of $\mathbb{R}^l$.
Indeed, we assume that the *Linear Independent Constraint Qualification* (*LICQ*) condition which requires the linear independency of the equality and the inequality constraints, is satisfied for the lower level problem of (6). Here, we refer to our analysis from Sect. 5.3.3 in [11], including the perturbation theory (if needed) as being presented there. Then, by means of Inverse Function Theorem applied at $\bar{\boldsymbol{\alpha}}$ on $\hat{T}$, we conclude that there exist open and bounded neighbourhoods $U^1 \subseteq \mathbb{R}^\iota$, $U^2 \subseteq \mathbb{R}^l$ around $((\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}}), \bar{\boldsymbol{\alpha}})$ such that $T := \hat{T}_{|U^1 \times U^2} : U^1 \times U^2 \to \mathcal{W} := \hat{T}(U^1 \times U^2)$ is a $C^1$-diffeomorphism. Shrinking $U^1$, we can guarantee that $\mathcal{W}$ is an axis parallel open box around $((\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}}), \boldsymbol{0}_l) \in \mathbb{R}^\iota \times \mathbb{R}^l$. Then, for each $(\theta, \boldsymbol{\wp}^{\mathcal{P}}) \in U^1$, the mapping $\Phi_{(\theta,\boldsymbol{\wp}^{\mathcal{P}})} := \left(\hat{T}((\theta, \boldsymbol{\wp}^{\mathcal{P}}), \cdot)\right)_{|U^2} : U^2 \to S^2$ is a $C^1$-diffeomorphism which transforms the (relative) neighbourhood $A \cap U^2$ of $\bar{\boldsymbol{\alpha}}$ on the (relative) neighbourhood

$$(\{\boldsymbol{0}\} \times \mathbb{H}^k \times \mathbb{R}^{l-1-k}) \cap S^2 \subseteq \mathbb{R}^l$$

of $\boldsymbol{0}$, where $S^2 = S(\boldsymbol{0}, \delta)$ stands for the open square around $\boldsymbol{0} = \boldsymbol{0}_l$ with a half side of length $\delta$. Here, $\mathbb{H}^k$ denotes the nonnegative orthant of $\mathbb{R}^k$:

$$\mathbb{H}^k := \{z \in \mathbb{R}^k \mid \; z_\ell \geq 0 \;\; (\ell \in \{1, 2, \ldots, k\})\}.$$

We call $\Phi_{(\theta,\boldsymbol{\wp}^{\mathcal{P}})}$ a *canonical local change of coordinates* of $\boldsymbol{\alpha}$. By this strategy, we locally transform $A$ into a rectangular manifold with corners and edges where the discretization will takes place in. More generally, a discretization point $\boldsymbol{z}$ from the discretized set (*regular grid*) $\mathbb{H}^k$ corresponds to a discretization point

$$\boldsymbol{\alpha} = \hat{T}^{-1}(\boldsymbol{z}) \tag{21}$$

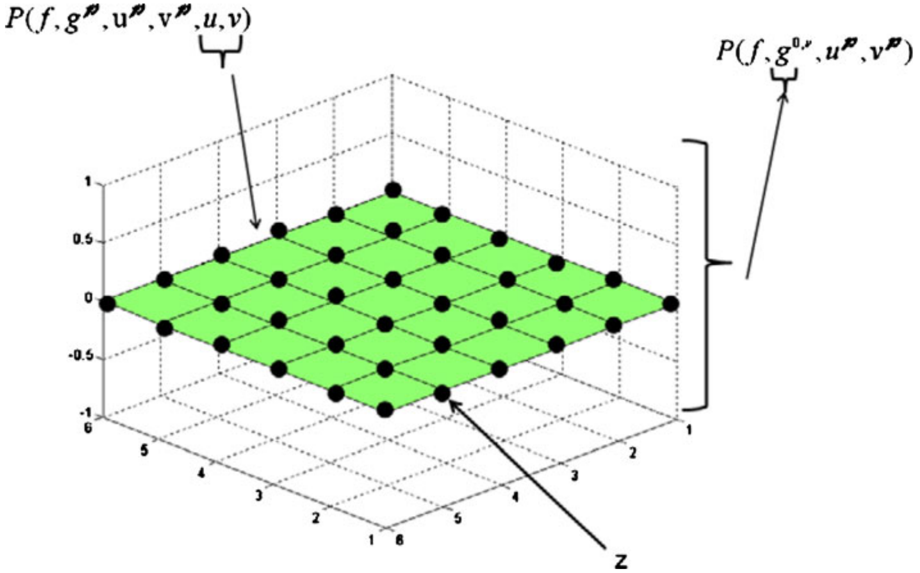


**Fig. 9** Illustration of the local discretization in $\mathbb{H}^k$, $P(f, g^{0,\nu}, u^{\mathcal{P}}, v^{\mathcal{P}})$ is the discretized problem and $P(f, g^{\mathcal{P}}, u^{\mathcal{P}}, v^{\mathcal{P}}, u, v)$ is the primal SIP problem, where $\nu$ is the number of grid points, $f$ is the objective function and $g^{\mathcal{P}}$ is the inequality constraint of the SIP problem; an example [24]

from the set $A$ by the back transformation $\hat{T}^{-1}$, implicitly represented by:

$$\hat{T}^{-1} : \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{k+1} \\ \alpha_{k+2} \\ \vdots \\ \alpha_l \end{bmatrix} := \hat{T}^{-1} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{k+1} \\ z_{k+2} \\ \vdots \\ z_l \end{bmatrix}. \tag{22}$$

The geometric illustration is shown in Fig. 9. The details of this method can be found in [24].

In the case of our problem, $A$ is already given by linear equalities and inequalities. For this reason, we can perform the linearization more easily. Indeed, we go from any vertex $\bar{\boldsymbol{\alpha}}$ to the neighbouring vertices and, by this, find a relative neighbourhood of $\boldsymbol{\alpha}$ in $A$ of "triangular" shape (cf. Fig. 10). Herewith, we obtain a linearization, but we do not guarantee $90^0$ inscribed at $\bar{\boldsymbol{\alpha}}$. However, it can be achieved by the transformation described above (if being wished).

*Note 1* Strategy II is more theoretical, but we can perform it more practically: it aims at finding how to compute "local" (neighbourhoods). In our problems, $u$ and $v$ are linear, so that the transformation $\hat{T}$ is linear and that inverse transformation, $\hat{T}^{-1}$, is linear, too. However, since $A$ has the special form of a polytope, one can use the neighbourhoods by the (relative) interiors of sub-polytopes (generated by neigbouring vertices), as being shown in Fig. 10. If we do this for all vertices $\bar{\boldsymbol{\alpha}}$, then only *interior* points remain, which constitute an **(interior sub-) polyhedron** that is often relatively small, especially, if the number of vertices is not too high.

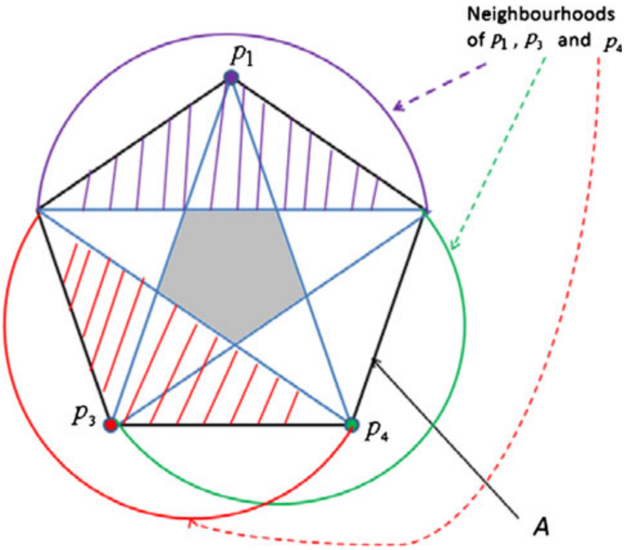


**Fig. 10** Illustration of the (local) linearization of $A$, with linear $u$ and $v$ [11]

This interior sub-polyhedron is shown by the shaded region in Fig. 10. With such a polyhedron we can proceed in our way again, and we continue, until the sub-polyhedron remaining is small enough, indeed. Now, all subdividing sets can be discretized by some scheme (e.g., by some canonical grids or by a uniform sampling on a sphere after transforming barycentric coordinates inside of the sub-polyhedron).

## 4 Exchange method

Another concept which is more powerful than discretization is the *exchange method* [6,7,22,23]. It is, in terms of refinement and complexity of the algorithm, located between discretization and the reduction ansatz. Given a discretization $A_k$, the discretized upper level problem $P(A_k)$ (7) is solved approximately, whereby the solution of the lower level problem

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & g((\theta, \beta), \boldsymbol{\alpha}) \\ \text{subject to} \quad & \boldsymbol{\alpha} \in A \end{aligned} \tag{23}$$

is obtained, firstly. In a next iteration, the discretization points of $A_k$ are updated, until the algorithm terminates according to some stopping criterion. The adapted exchange algorithm to our primal problem is given by Algorithm 2.

As it is discussed in Sect. 3, we can use anyone of the alternatives from (9) as a stopping criterion.

In this section, we apply the exchange algorithm to our SIP problem which is parametrized by a well known so-called *uniform continuous density* function from probabibilty theory [20]. Before constructing the algorithm, we find constraints of our density function in the following example.

*Example 2* We assume that the objective and the constraint functions, $f, h, g, u, v$, respectively, are two times continuously differentiable $C^2$ functions. Now, the global continuity

**Algorithm 2 Primal Exchange Method (PEM)**

```
Input:
  $\delta$          positive number, i.e., $\delta > 0$
  $f^{\mathcal{P}}$  probability density function
  $P^{\mathcal{P}}$  the set where pdf parameters lie
Output:
  $\theta_k$          unknown variable for minimization, to be evaluated
  $\boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}}$  the parameter vector of the pdf
  $\boldsymbol{\alpha_k}$        dual variable of SVM (support vectors)
PEM$\left(\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}}, \boldsymbol{\alpha_k}, A, \delta, f^{\mathcal{P}}, P^{\mathcal{P}}\right)$
1: $k := 0$
2: Initialize a discretization $A_k \subset \Omega$.
3: DO Compute a solution $(\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}})$ of
     $$\begin{array}{l} \min\limits_{\theta\in\mathbb{R}, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}} \quad -\theta \\ \text{subject to} \quad g^{\mathcal{P}}((\theta, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}) := \int_\Omega T(\omega, \boldsymbol{\alpha}) f^{\mathcal{P}}(\omega; \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) d\omega - \theta \geqslant 0 \ \ (\boldsymbol{\alpha} \in A_k), \\ \mathrm{U}_i^{\mathcal{P}}(\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) = 0 \ \ (i \in I^{\mathcal{P}}), \\ \mathrm{V}_j^{\mathcal{P}}(\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) \geq 0 \ \ (j \in J^{\mathcal{P}}). \end{array}$$
4: Compute local solutions $\boldsymbol{\alpha}_{\boldsymbol{k}}^{\boldsymbol{i}}$ $(i = 1, 2, \ldots, i_k)$ of the reduced problem such that one of them, say $\boldsymbol{\alpha}_{\boldsymbol{k}}^{\boldsymbol{i_0}}$, is a global solution, i.e.,
     $$g^{\mathcal{P}}((\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}_{\boldsymbol{k}}^{\boldsymbol{i_0}}) = \min_{\boldsymbol{\alpha}\in A} g^{\mathcal{P}}((\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}).$$
5: if $g^{\mathcal{P}}((\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}_{\boldsymbol{k}}^{\boldsymbol{i_0}}) \geq -\delta$ with a solution $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\boldsymbol{\mathcal{P}}}) \approx (\theta_k, \boldsymbol{\wp}_{\boldsymbol{k}}^{\boldsymbol{\mathcal{P}}})$, then
6:    STOP
7: else
8:    $A_{k+1} := A_k \cup \left\{\alpha_k^i \mid i = 1, 2, \ldots, i_k\right\}$
9:    $k := k + 1$
10: end if
11: END DO
```

can fail for our function $g$, depending on the parametrization of the corresponding pdf. As an example, we choose a uniform continuous density function with parameter vector $\boldsymbol{\wp}^{\mathcal{P}} = (a, b)$ $(a \leq b)$ [20]. Let us recall that the pdf of the uniform continuous density is

$$f^{\mathcal{P}}(\omega; (a, b)) = \begin{cases} \frac{1}{b-a}, & a \leqslant \omega \leqslant b, \\ 0, & \omega < a \text{ or } \omega > b. \end{cases} \tag{24}$$

We observe that the term $\frac{1}{b-a}$ makes the function $g$ (cf. (6)) discontinuous, actually, undefined at $a = b$. On the other hand, we need an inequality constraint, e.g., of the form "$\leq$", such as in $a \leq b$. To encounter this, let us introduce a sufficiently small positive number $\epsilon > 0$ such that the following relation is requested:

$$a + \epsilon \leq b.$$

Then, we prevent from equality of $a$ and $b$ with this small positive number and, hence, from discontinuity, by the additional inequality constraint functions. In the following, the algorithm of exchange method for solving our primal problem parametrized by a function (24), is presented.

**Algorithm 3 Primal Exchange Method (PEM) Parametrized by Uniform Continuous Density function**

```
Input:
  $A$   an infinite index set
  $\delta$   positive number, i.e., $\delta > 0$
  $\epsilon$   positive number, i.e., $\epsilon > 0$
Output:
  $\theta_k$   unknown variable for minimization, to be evaluated
  $a_k$   a parameter of the pdf
  $b_k$   a parameter of the pdf
  $\boldsymbol{\alpha_k}$   dual variable of SVM (support vectors)
PEM$\big(\theta_k, a_k, b_k, \boldsymbol{\alpha_k}, A, \delta, \epsilon\big)$
$k := 0$
Initialize a discretization $A_k \subset A$.
DO Compute a solution $(\theta_k, a_k, b_k)$ of

    $\min_{\theta\in\mathbb{R}, a\in\mathbb{R}, b\in\mathbb{R}} \quad -\theta$
    subject to $g^{\mathcal{P}}((\theta, a, b), \boldsymbol{\alpha}) := \int_{\Omega} T(\omega, \boldsymbol{\alpha}) f^{\mathcal{P}}(\omega; a, b) d\omega - \theta \geqslant 0 \;\; (\boldsymbol{\alpha} \in A_k),$
    $a + \epsilon \leq b.$

Compute local solutions $\boldsymbol{\alpha_k^i}$ $(i = 1, 2, \ldots, i_k)$ of the reduced problem such that one of them, say $\boldsymbol{\alpha_k^{i_0}}$, is
  global solution, i.e.,

    $g^{\mathcal{P}}((\theta_k, a_k, b_k), \boldsymbol{\alpha_k^{i_0}}) = \min_{\boldsymbol{\alpha}\in A} g^{\mathcal{P}}((\theta_k, a_k, b_k), \boldsymbol{\alpha}).$

if $g^{\mathcal{P}}((\theta_k, a_k, b_k), \boldsymbol{\alpha_k^{i_0}}) \geq -\delta$ with a solution $(\bar{\theta}, \bar{a}, \bar{b}) \approx (\theta_k, a_k, b_k)$, then
    STOP
else
    $A_{k+1} := A_k \cup \left\{\boldsymbol{\alpha_k^i} \mid i = 1, 2, \ldots, i_k\right\}.$
    $k := k + 1$
end if
END DO
```

The convergence of the exchange method applied on our primal problem by Algorithm 3 is presented through the following theorem [22].

**Theorem 7** [22]. *We refer to* $\mathcal{M}^{\mathcal{P}}_{comp}(A)$ *which is obtained by the compactification of feasible set* $\mathcal{M}^{\mathcal{P}}(A)$, *by transversally intersection of original feasible set with simple geometrical bodies* (*e.g., parallelpipes*) *provided by Closer Explanation 1. Then, the exchange method* (*with* $\delta = 0$) *either stops at some iteration* $k_0 \in \mathbb{N}_0$ *with a solution* $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\boldsymbol{\mathcal{P}}}) = (\theta_{k_0}, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k_0}})$ *of* (6) *or the sequence* $(\theta_k, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k}})_{k\in\mathbb{N}_0}$ *of solutions of* (7) *satisfies*

$$\min_{(\theta, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}})\in\mathcal{G}^{\mathcal{P}}(A)} \left\| (\theta_k, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k}}) - (\theta, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) \right\|_2 \to 0 \;\; (k \to \infty).$$

*Proof* We prove the theorem by contradiction. Let us assume that the algorithm does not stop with a minimizer of (6). As in the proof of Theorem 2 given in [22], by our assumptions, a solution $(\theta_k, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k}})$ of (6) exists, $(\bar{\theta}_k, \bar{\boldsymbol{\wp}}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k}}) \in \mathcal{M}^{\mathcal{P}}_{comp}(A_0)$, and with a suitable, existing subsequence $(\theta_{k_\nu}, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k_\nu}})_{\nu\in\mathbb{N}_0}$ and a vector $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\boldsymbol{\mathcal{P}}})$ such that $(\theta_{k_\nu}, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k_\nu}}) \to (\bar{\theta}, \bar{\boldsymbol{\wp}}^{\boldsymbol{\mathcal{P}}})$ $(\nu \to \infty)$, where the solution is in the compact elementary geometrical body (parallelpipe or so) $\mathcal{C}$ (see Closer Explanation 1), $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\boldsymbol{\mathcal{P}}}) \in \mathcal{C}$ and $\bar{\boldsymbol{\wp}}^{\boldsymbol{\mathcal{P}}} \in P^{\mathcal{P}}$, and we find

$$-\bar{\theta} \leq v(A).$$

Again, we must show $(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}}) \in \mathcal{M}^{\mathcal{P}}_{comp}(A)$ or, equivalently, $\varphi(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}}) \geq 0 \ \ (\boldsymbol{\alpha} \in A)$ for the **value function** $\varphi(\theta, \boldsymbol{\wp}^{\mathcal{P}})$ of lower level problem, i.e., $\varphi(\theta, \boldsymbol{\wp}^{\mathcal{P}}) = \min\limits_{\boldsymbol{\alpha} \in A} \ g((\theta, \boldsymbol{\wp}^{\mathcal{P}}), \boldsymbol{\alpha})$. In view of $\varphi(\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}}) = g((\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}}), \boldsymbol{\alpha}^{\mathbf{1}}_{\boldsymbol{k}})$, we can write

$$\varphi(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})=\varphi(\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}})+\varphi(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})-\varphi(\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}})=g((\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}}), \boldsymbol{\alpha}^{\mathbf{1}}_{\boldsymbol{k}})+\varphi(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})-\varphi(\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}}).$$

Since $\boldsymbol{\alpha}^{\mathbf{1}}_{\boldsymbol{k}} \in A_{k+1}$, we have $g((\theta_{k+1}, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k+1}}), \boldsymbol{\alpha}^{\mathbf{1}}_{\boldsymbol{k}}) \geq 0$ and by continuity of $g$ and $\varphi$, we find

$$\varphi(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}}) \geq (g((\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}}), \boldsymbol{\alpha}^{\mathbf{1}}_{\boldsymbol{k}})-g((\theta_{k+1}, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k+1}}), \boldsymbol{\alpha}^{\mathbf{1}}_{\boldsymbol{k+1}}))+ \Big(\varphi(\bar{\theta}, \bar{\boldsymbol{\wp}}^{\mathcal{P}})-\varphi(\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}})\Big) \to 0$$

for $k \to \infty$, which concludes the proof. We refer to [7] for detailed explanation. □

## 5 Conceptual reduction method

The *conceptual reduction method* is based on local reduction which starts with an arbitrary point $\mathbf{x}^*$ (not necessarily feasible) for the SIP problem and solves the lower level problem at that point, i.e., it solves $Q(\mathbf{x}^*)$ to find all the local minima $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^r$ of $Q(\mathbf{x}^*)$ (finiteness of local minnima is assumed):

$$Q(\bar{\mathbf{x}}) \qquad \begin{array}{l} \min\limits_{\mathbf{y}} \ g(\bar{\mathbf{x}}, \mathbf{y}) \\ \text{such that } \ u_k(\mathbf{y}) = 0 \ (k \in K) \text{ and } \ v_\ell(\mathbf{y}) \geq 0 \ \ (\ell \in L). \end{array} \tag{25}$$

We note that our infinite index sets are compact, and the differentiability, nondegeneracy and continuity assumptions of our model defining functions hold. Then, by Theorem of Heine-Borel there are finitely many local minima of the lower level problem $Q(x)$ indeed (cf. [24]).

It finds the optimal solution for the reduced finite problem which has $r$ many constraints, and the iteration continues until the stopping criterion given by line 4 of the Algorithm 4 is fulfilled. Alternatively, one can choose one of the stopping criteria from (9). In the following algorithms, we presented our conceptual reduction method, adapted to the primal problem (6) based on [7].

We observe that step 2 in Algorithm 4 is very costly as it requires a global search for minima of $g((\theta_k, \boldsymbol{\wp}^{\mathcal{P}}_{\boldsymbol{k}}), \boldsymbol{\alpha})$ on $A$. We must avoid an execution of this step in the overall process as much as possible. Step 2 assumes that there are only finitely many minima of the lower level problem for the primal (the dual case). If it does not hold, another method, e.g., discretization, should be used. Let us note that step 3 solves a finitely constrained optimization problem which requires only local searches and can be efficiently performed, e.g., by a Newton-type of method.

*Remark 4* The only difference between the exchange method and the conceptual reduction method is the starting point of the iteration. In the exchange method, we start with an initial feasible (discretized) index set. But, on the other hand, in the conceptual reduction method, we do not need to find a discretized set but an initial guess of the optimal solution of the upper level problem which does not need to be feasible. In our primal problem, as it is discussed in Sect. 2, we have difficulties in computing the discretization of the set $A$. We proposed different strategies to discretize the set $A$ by Strategies I and II. Alternatively, to solve our primal and the dual problems, we can use the conceptual reduction method without any need of a discretization step.

**Algorithm 4 Primal Conceptual Reduction Method (PCRM)**

```
Input:
  $(\theta_0, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\mathbf{0}})$   initial guess for the optimal solution which is not necessarily feasible
  $\epsilon$   sufficiently small positive number to be used for one of the stopping criteria given by (9)
  $f^{\mathcal{P}}$   probability density function
  $P^{\mathcal{P}}$   the set where the pdf parameters lie
Output:
  $\theta_k$   unknown variable for minimization, to be evaluated
  $\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k}}$   the parameter vector of the pdf
  $\boldsymbol{\alpha_k}$   dual variable of SVM (support vectors) $(i = 1, 2, \ldots, r)$

PCRM$\left(\theta_k, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k}}, \boldsymbol{\alpha_k}, \theta_0, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\mathbf{0}}, \delta, f^{\mathcal{P}}, P^{\mathcal{P}}\right)$
$k := 0$
Determine all local minima $\boldsymbol{\alpha}^{\mathbf{1}}_{\boldsymbol{k}}, \boldsymbol{\alpha}^{\mathbf{2}}_{\boldsymbol{k}}, \ldots, \boldsymbol{\alpha}^{\boldsymbol{r}}_{\boldsymbol{k}}$ of
    $\min_{\boldsymbol{\alpha} \in A} g^{\mathcal{P}}((\theta_k, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k}}), \boldsymbol{\alpha}).$
DO Compute a solution $(\theta^*, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}*}})$ of
    $\min_{\theta \in \mathbb{R}, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}} \quad -\theta$
    subject to $g^{\mathcal{P}}((\theta, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}), \boldsymbol{\alpha}^{\boldsymbol{l}}_{\boldsymbol{k}}) := \int_{\Omega} T(\omega, \boldsymbol{\alpha}) f^{\mathcal{P}}(\omega; \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) d\omega - \theta \geqslant 0 \ (l = 1, 2, \ldots, r),$
    $\mathrm{u}^{\mathcal{P}}_i(\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) = 0 \ (i \in I^{\mathcal{P}}),$
    $\mathrm{v}^{\mathcal{P}}_j(\boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}) \geq 0 \ (j \in J^{\mathcal{P}}).$
if One of the stopping criteria given by (9) is satisfied, then
    STOP
else
    $(\theta_{k+1}, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}}}_{\boldsymbol{k+1}}) := (\theta^*, \boldsymbol{\wp}^{\boldsymbol{\mathcal{P}*}})$
    $k := k + 1$
end if
END DO
```

**Table 1** Data set description

| Data set | # Instances | # Attributes | Attribute characteristics |
|---|---|---|---|
| Votes | 52 | 16 | Categorical |
| Bupa | 345 | 6 | Integer, real and categorical |

## 6 Implemantation of IKL

We tested our IKL problem based on Algorithms 3 and 4 for two kinds of data sets. The discretization of the index set is obtained by random search of corner points.

We first implemented our method on *votes* data which is a homogeneous data set and secondly, we tested our method on *bupa* data set[4] which is a heterogeneous data set. Data descriptions are given in Table 1.

We compared our results with the *single kernel SVM* model introduced in Sect. 1. Furthermore, we considered a single kernel SVM for the comparison with our IKL because of

[4] Available from http://archive.ics.uci.edu/ml/.

**Table 2** Percentage of the accuracy

| Data set | Single kernel SVM (%) | IKL by exchange method (%) | IKL by PCRM (%) |
|---|---|---|---|
| Votes | 92 | 75 | 91 |
| Bupa | 73 | 99 | 99 |

its simplicity. Let us note that *single kernel* is a special case of multiple kernel learning, i.e., the upper limit of multiple combination of kernel is taken as $K = 1$.

For both single kernel SVM and IKL, we selected the regularization constant of SVM, $C$, by 5-fold cross validation from the search set

$$\{2^{-5}, 2^{-3}, 2^{-1}, 2, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}\}.$$

The results are interpreted by the accuracy percentage which is computed by the ratio of correct classification to total number of test points. Further, we used LibSVM package [5] for single kernel SVM and the parameter of Gaussian kernel is chosen by 5 fold cross validation over a search set of $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. We implemented our IKL model by Algorithm 4 with a Gaussian kernel, and the search space for a Gaussian kernel width is chosen as $\Omega = [0, 1]$. All the implementations are done in Matlab 7.0.4. In order to have finitely many local minima in the lower level problem and to use Heine Borel Theorem, the domain of our variable $\theta$ is restricted to a compact set, e.g., $[-100, 100]$. We used the *fmincon* function of Matlab Optimization toolbox to implement both the exchange and the conceptual reduction method.

As our IKL aims to help for the classification of **heterogeneous data**, the results given by Table 2 show that IKL increased the accuracy from 73% to 99% for bupa set whereas it could not be successful for homogeneous data, votes. The results for homogeneous data can be improved if different kernels are chosen and different numerical methods are used. This will be a subject of our future study.

## 7 Conclusion

By means of new ideas, we developed well-known numerical methods of semi-infinite programming for our new kernel machine. We improved the discretization method for our specific model and proposed two new algorithms (see Strategies I and II). The advantage of these methods were discussed and the intuition behind these algorithms were visualized by figures and examples. We stated the convergence of the numerical methods with theorems and we analyzed the conditions and assumptions of these theorems such as optimality and convergence. We implemented our novel kernel learning algorithm called IKL by two well-known numerical methods for SIP, i.e., exchange method and conceptual discretization method. We achieved very satisfactory accuracy for heterogeneous data and we also got promising accuracy for homogeneous data. As it was claimed by us that IKL was developed to help classification of heterogeneous data, these results validated our proposal. The accuracy results for exchange method on votes data is not promising due to the discretization step at the exchange method.

In addition, we intend to study infinite programming and investigate primal-dual methods instead of reducing the infinite problem into semi-infinite programming. Furthermore, we will investigate our works with MFCQ and strong stability of all *Karush–Kuhn–Tucker (KKT)* points [7] in the reduction ansatz.

**Acknowledgements** The authors cordially thank to three anonymous referees for their constructive critisim, and to the professors E. Anderson, U. Çapar, M. Goberna, and Y. Ozan for their valuable advices.

## References

1. Anderson, E.J., Nash, P.: Linear Programming in Infinite-Dimensional Spaces. Wiley, New York (1987)
2. Apostol, T.M.: Mathematical Analysis: A Modern Approach to Advanced Calculus. Addison Wesley, Reading (1974)
3. Argyriou, A., Hauser, R., Micchelli, C., Pontil, M.: A dc-programming algorithm for kernel selection. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA (2006)
4. Bach, F.R., Lanckriet, G.R.G.: Multiple kernel learning, conic duality, and the smo algorithm. In: Proceedings of the 21st International Conference on Machine Learning (2004)
5. Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/cjlin/libsvm (2001)
6. Hettich, R., Jongen, H.Th.: Semi-infinite programming: conditions of optimality and applications. In: Stoer, J. (ed.) Optimization Techniques 2. Lecture notes in control and Information Science, Springer, Berlin, Heidelberg, New York (1978)
7. Hettich, R., Kortanek, O.: Semi-infinite programming: theory, methods and applications. SIAM Rev. 35, **3**, 380–429 (1993)
8. Hettich, R., Zencke, P.: Numerische Methoden der Approximation und semi-infiniten Optimierung. Tuebner, Stuttgart (1982)
9. Jongen, H.Th., Jonker, P., Twilt, F.: Nonlinear Optimization in Finite Dimensions—Morse Theory, Chebyshev Approximation, Transversality, Flows, Parametric Aspects. Springer, New York (2000)
10. Ozan, Y.: Scientific discussion with Yıldıray Ozan. Department of Mathematics, METU, Ankara (2008)
11. Özöğür-Akyüz, S.: A Mathematical Contribution of Statistical Learning and Continuous Optimization Using Infinite and Semi-Infinite Programming, to Computational Statistics (submitted). PhD thesis, Middle East Technical University, Insitiute of Applied Mathematics, Department of Scientific Computing, February (2009)
12. Özöğür-Akyüz, S., Weber, G.W.: Modelling of kernel machines by infinite and semi-infinite programming. In: Hakim, A.H., Vasant, P., Barsoum, N. (eds.) Proceedings of the Second Global Conference on Power Control and Optimization, AIP Conference Proceedings, Bali, Indonesia. Mathematical and Statistical Physics, pp. 306–313. ISBN 978-0-7354-0696-4 (August 2009), 1–3 June (2009)
13. Özöğür-Akyüz, S., Weber, G.-W.: Learning with infinitely many kernels via infinite and semi-infinite programming. Accepted to Optimization Methods and Software
14. Özöğür-Akyüz, S., Hussain, Z., Shawe-Taylor, J.: Prediction with the SVM using test point margins. In: R. Sharda, S.V.B (eds) Special issue on Data Mining of Journal of (German B) In Annals of Information Systems. Springer Book Series (2009) (in press)
15. Özöğür-Akyüz, S., Shawe-Taylor, J., Weber, G.-W., Ögel, Z.B.: Pattern analysis for the prediction of eukoryatic pro-peptide cleavage sites. Discret. Appl. Math. **157**(10), 2388–2394 (2009)
16. Özöğür-Akyüz, S., Weber, G.-W.: Learning with infinitely many kernels via semi-infinite programming. In: Continuous Optimization and Knowledge Based Technologies, 20th EURO Mini conference, Lithuania, pp. 342–348 (2008)
17. Pardalos, P.M., Hansen, P. (eds.): Data Mining and Mathematical Programming, volume CRM 45. American Mathematical Society, Providence (2008)
18. Rückmann, J.-J., Weber, G.-W.: Semi-infinite optimization: Excisional stability of noncompact feasible sets. Sibir. Math. Z. **39**, 129–145 (1998)
19. Seref, O., Kundakcioglu, O.E., Pardalos, P.M. (eds.): Data Mining, Systems Analysis, and Optimization in Biomedicine. Springer, New York (2007)
20. Shiryaev, A.N.: Probability. Springer, New York (1995)
21. Sonnenburg, S., Räetsch, G., Schafer, C., Schölkopf, B.: Large scale multiple kernel learning. J. Mach. Learn. Res. **7**, 1531–1565 (2006)
22. Still, G.: Semi-infinite programming: An introduction, preliminary version. Technical report, University of Twente Department of Applied Mathematics, Enschede, The Netherlands (2004)
23. Vaz, A.I.F., Fernandes, E.M.G.P., Gomes, M.P.S.F.: Discretization methods for semi-infinite programming. Investigaç ão Operacional **21**(1), (2001)

24. Weber, G.-W.: Generalized Semi-Infinite Optimization and Related Topics, volume 29 of Research and Exposition in Mathematics. Heldermann, Berlin (2003)
25. Yershova, A., LaVelle, S.M.: Deterministic sampling methods for spheres and $so(3)$. In: IEEE International Conference on Robotics and Automation (ICRA) (2004)