Optimization problems in statistical learning: duality and optimality conditions

Radu Ioan Boţ * Nicole Lorenz †

February 8, 2011

Abstract. Regularization methods are techniques for learning functions from given data. We consider regularization problems the objective function of which consisting of a cost function and a regularization term with the aim of selecting a prediction function f with a finite representation $f(\cdot) = \sum_{i=1}^{n} c_i k(\cdot, X_i)$ which minimizes the error of prediction. Here the role of the regularizer is to avoid overfitting. In general these are convex optimization problems with not necessarily differentiable objective functions. Thus in order to provide optimality conditions for this class of problems one needs to appeal on some specific techniques from the convex analysis. In this paper we provide a general approach for deriving necessary and sufficient optimality conditions for the regularized problem via the so-called conjugate duality theory. Afterwards we employ the obtained results to the Support Vector Machines problem and Support Vector Regression problem formulated for different cost functions.

Keywords. machine learning, Tikhonov regularization, convex duality theory, optimality conditions

AMS subject classification. 47A52, 90C25, 49N15

1 Some elements of statistical learning

Support Vector Machines are techniques for solving problems of learning from a given example data set based on the *Structural Risk Minimization Principle* and they were first mentioned by Vapnik in [22]. The reader is also referred to [21, 23] for a deeper insight into this field.

Evgeniou, Pontil and Poggio distinguish in [8] between two types of statistical learning problems: the *Support Vector Machines Regression* problem (SVMR) and the *Regularization Networks* (RN). The problems belonging to the first class have as possible application the approximation and determination of a function by means of a data set. We deal here with a particular case of this problem, the so-called *Support Vector Machines Classification* (SVMC).

*Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: radu.bot@mathematik.tu-chemnitz.de. Research partially supported by DFG (German Research Foundation), project WA 922/1–3.

†Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: nicole.lorenz@mathematik.tu-chemnitz.de.

Consider a given set with n training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i \in \mathbb{R}^k$ and $Y_i \in \mathbb{R}, i = 1, \ldots, n$, and let $\mathfrak{F}$ be a space of functions defined on $\mathbb{R}^k$ with real values. The SVMC problem looks for a function $f \in \mathfrak{F}$ such that for a previously unknown value X the function f predicts the value Y. The penalty for predicting $f(X_i)$ having as true value Y_i for $i = 1, \ldots, n$ is measured by a so-called *cost function* $v : \mathbb{R}^2 \to \overline{\mathbb{R}}$.

The problem of finding an optimal function f in $\mathfrak{F}$ is *ill-posed* since there are infinitely many solutions. In order to get a *well-posed* problem, and, consequently, to be able to provide a particular solution, we need some additional *a priori* information about f. A common one is the assumption that the function f is *smooth*, in other words, two similar inputs correspond to two similar outputs. In this way one is able to control the complexity of f. To this aim one has to introduce a *regularization term* $\frac{\lambda}{2}\Omega(f)$ (cf. [2, 3, 20]), where the *regularization parameter* $\lambda > 0$ controls the tradeoff between the cost function and the regularizer Ω (cf. [25]). In this context Ω is also called *smoothness functional* and has the desired characteristic of taking high values for non-smooth functions and low values for smooth functions. The following *Tikhonov regularization problem* arises

$$\inf_{f \in \mathfrak{F}} \left\{ \sum_{i=1}^{n} v(f(X_i), Y_i) + \frac{\lambda}{2}\Omega(f) \right\}, \tag{1}$$

the objective function of which being called *regularization functional*.

Further let $\mathfrak{H}_k$ be a *Reproducing Kernel Hilbert Space* (RKHS) introduced by a *kernel function* $k : \mathbb{R}^{k \times k} \to \mathbb{R}$ (cf. [1]). In the following we ask f to be an element of $\mathfrak{H}_k$. Moreover, we assume that k is *symmetric*, namely that $k(x, y) = k(y, x)$ for $x, y \in \mathbb{R}^k$. The kernel function k introduces a *kernel matrix* $K \in \mathbb{R}^{n \times n}$, where $k(X_i, X_j) = K_{ij}$ for $i, j = 1, \ldots, n$. In this context K, which is a symmetric matrix, is said to be the *Gram matrix of k with respect to $X_1, \ldots, X_n$*. A symmetric kernel function $k : \mathbb{R}^{k \times k} \to \mathbb{R}$ which for all $n \geq 1$ and all finite sets $\{X_1, \ldots, X_n\} \subset \mathbb{R}^k$ fulfills $\sum_{i,j=1}^{n} a_i a_j k(X_i, X_j) \geq 0$ for every arbitrary $a \in \mathbb{R}^n$ is called *finitely positive semidefinite kernel* (cf. [19]). One can easily see that such a kernel function gives rise to a positive semidefinite Gram matrix K. On the other hand, it is worth noticing that (see [19, Theorem 3.11]) a k which is either continuous or has a finite domain can be decomposed as $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$, where $\Phi : \mathbb{R}^k \to F$ is a *feature map* and F a Hilbert space, if and only if it is finitely positive semidefinite.

It is well-known that when having a symmetric finitely positive definite kernel k and a corresponding Gram matrix one can find a RKHS $\mathfrak{H}_k$ induced by it, such that the so-called *reproducing property*, namely that $f(x) = \langle f(\cdot), k(x, \cdot) \rangle$ for all $x \in \mathbb{R}^k$, is fulfilled (cf. [1]). Shawe-Taylor and Cristianini have shown in [19] that one can construct a RKHS $\mathfrak{H}_k$ even for a symmetric finitely positive semidefinite kernel function such that the reproducing property is valid. More than that, via the so-called *representer theorem* (cf. [25]) one has that for every minimizer f of (1) there exists $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ such that

$$f = \sum_{j=1}^{n} c_j k(\cdot, X_j). \tag{2}$$

This is the setting considered in this paper and in the following we additionally assume that for $f \in \mathfrak{H}_k$ the smoothness functional is defined as $\Omega(f) = \|f\|_k^2$, where

$||\cdot||_k$ is the norm in $\mathfrak{H}_k$. If for $f \in \mathfrak{H}_k$ the vector $c \in \mathbb{R}^n$ is the one that comes from the representation given in (2), then $\Omega(f) = ||f||_k^2 = c^T Kc$ and for all $i = 1, ..., n$ it holds $f(X_i) = \sum_{j=1}^n c_j K_{ij} = (Kc)_i$. Thus the optimization problem (1) can be equivalently written as

$$\inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^n v((Kc)_i, Y_i) + \frac{\lambda}{2} c^T Kc \right\}. \tag{3}$$

Unfortunately, the most popular and most efficient cost functions used in the literature on machine learning fail to be differentiable (see, for instance, [8, 16, 18]). This causes some difficulties when trying to furnish optimality conditions for the above problem. On the other hand, these functions turn out to be convex in the first variable and, consequently, problem (3) becomes a convex optimization problem. In the following section we provide a general approach for deriving optimality condition for problem (3) by means of the conjugate duality theory in convex optimization. The optimality conditions for (3) will be expressed as systems of nonlinear equations involving the conjugates of the cost functions or, alternatively, via convex subdifferential formulae. As a byproduct we extend in this way the approach presented in [14], where when dealing with problem (3) the authors impose invertibility for K. We show that, in spite of the fact that we avoid this assumption, one can deliver handleable optimality conditions for (3), only by exploiting the very strong results of the convex analysis.

The described regularization framework includes many well-known learning methods. Depending on the application one can use different cost functions (see for instance [8, 14] for several examples). In section 3 we consider some particular instances of the *Support Vector Machines Classification* problem, namely when the output Y takes values in $\{+1, -1\}$. In this case we speak about a *(binary) classification problem*. In particular we deal with the *hinge loss* (or *soft margin*) (cf. [7, 22]) $v^{hl} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $v^{hl}(a, Y) = (1 - (a + b)Y)_+$, for $b \in \mathbb{R}$, but also with the *generalized hinge loss* (cf. [5]) $v^{ghl} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $v^{ghl}(a, Y) = (1 - (a + b)Y)_+^u$, where $u > 1$ is given.

In section 4 we turn our attention to the *Support Vector Regression* problem, which is characterized by the fact that the output Y may take arbitrary real values. In this context we deal with the following *extended loss* function $v^{el} : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, $v^{el}(a, Y) = \delta_{[-\varepsilon,\varepsilon]}(Y - a)$, where $\varepsilon > 0$, as well as with a *generalization of Vapnik's ε-insensitive loss* introduced by Smola, Schölkopf and Müller in [18], which we describe in detail in subsection 4.2. Especially by means of the extended loss we succeed in underlining the role of the regularity conditions when providing optimality conditions even in the context of machine learning. Obviously, via the general approach from section 2 one can consider also other cost functions suitable for the classification and regression problem.

It is worth to notice that in the investigations made in the sections 3 and 4 we take advantage of the convexity properties of cost functions involved. This fact allows us to employ the convex duality theory and to make use of the well-developed convex subdifferential calculus. On the other hand, this approach suggests the possibility to use nonsmooth and nonconvex cost functions in statistical learning. In order to provide optimality conditions for the optimization problems arising in this way, one could apply the calculus formulae which exist in the literature for different subdifferentials. In a first step one could consider locally Lipschitz cost functions in connection with the Clarke

subdifferential (cf. [6]), but also some more general classes of functions in connection with some appropriate subdifferential notions, as one can find in [10].

The paper is closed by a conclusive section.

2 Notation and preliminary results

For two vectors $x, y \in \mathbb{R}^n$ we denote by $x^T y$ their *scalar product*, where the upper index ^{T} transposes a column vector into a row one and viceversa. By e_i, $i = 1, ..., n$, we denote the *i-th unit-vector* in $\mathbb{R}^n$. For a nonempty set $D \subseteq \mathbb{R}^n$ we denote by $\delta_D : \mathbb{R}^n \to \overline{\mathbb{R}}$ the *indicator function* of D, which is defined by $\delta_D(x) = 0$ if $x \in D$, being equal to $+\infty$, otherwise. Further, by $\mathrm{ri}(D)$ we denote the *relative interior* of the set D, that is the interior of D relative to its affine hull. For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ we denote its *effective domain* by $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}$ and say that f is *proper* if $\mathrm{dom}(f) \neq \emptyset$ and $f > -\infty$. The *(Fenchel-Moreau) conjugate function* of f is $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$, defined by $f^*(p) = \sup_{x \in \mathbb{R}^n}\{p^T x - f(x)\}$. We have the following relation, known as the *Young-Fenchel inequality*, $f(x) + f^*(p) - p^T x \geq 0$ and this is true for all $x, p \in \mathbb{R}^n$. For $x \in \mathbb{R}^n$ with $f(x) \in \mathbb{R}$ we denote by $\partial f(x) := \{p \in \mathbb{R}^n : f(y) - f(x) \geq p^T(y - x)\ \forall y \in \mathbb{R}^n\}$ the *(convex) subdifferential of f at x*. Otherwise, we assume by convention that $\partial f(x) = \emptyset$. For $x \in \mathbb{R}^n$ with $f(x) \in \mathbb{R}$ one has that

$$p \in \partial f(x) \Leftrightarrow f(x) + f^*(p) = p^T x.$$

For a linear mapping $K : \mathbb{R}^n \to \mathbb{R}^m$ we denote by $\mathrm{Im}(K) := \{Kx : x \in \mathbb{R}^n\}$ the *image* of K. Further, for $x \in \mathbb{R}$ we define $x_+ := \max(0, x)$.

In order to develop a duality theory and to formulate necessary and sufficient optimality conditions for problem (3), we treat first, by means of some techniques from the convex analysis, the following optimization problem

$$(P) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{l} v_i(Kc) + g(c) \right\},$$

where $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $v_i : \mathbb{R}^m \to \overline{\mathbb{R}}, i = 1, \ldots, l$, are proper and convex functions and $K : \mathbb{R}^n \to \mathbb{R}^m$ is a linear mapping such that $K^{-1}\left(\bigcap_{i=1}^{l} \mathrm{dom}(v_i)\right) \cap \mathrm{dom}(g) \neq \emptyset$. The latter condition is called *feasibility condition* and guarantees that $v(P) < +\infty$, where by $v(P)$ we denote the optimal objective value of (P). Throughout the paper, for a given optimization problem, we write min (max) instead of inf (sup) if the infimum (supremum) is attained. Before stating optimality conditions for (P) we consider its following *Fenchel-type* conjugate dual problem

$$(D) \qquad \sup_{p^i \in \mathbb{R}^m, i=1,\ldots,l} \left\{ -\sum_{i=1}^{l} v_i^*(p^i) - g^*\left(-K^T\left(\sum_{i=1}^{l} p^i\right)\right) \right\}.$$

Next we show that for (P) and (D) weak duality always holds, namely that $v(P) \geq v(D)$, where by $v(D)$ we denote the optimal objective value of the dual (D).

Theorem 1. *(weak duality theorem) It holds $v(P) \geq v(D)$.*

Proof. Let be $c \in \mathbb{R}^n$ and $p^i \in \mathbb{R}^m, i = 1, \ldots, l$. Then, by the Young-Fenchel inequality, it holds

$$-\sum_{i=1}^{l} v_i^*(p^i) - g^*\left(-K^T\left(\sum_{i=1}^{l} p^i\right)\right) \leq \sum_{i=1}^{l} v_i(Kc) + g(c).$$

From here one automatically has that $v(D) \leq v(P)$. □

For strong duality, namely the situation when $v(P) = v(D)$ and the dual has an optimal solution, we need to impose the fulfillment of a so-called *regularity condition*. With this respect we use a weak interior-point regularity condition.

Theorem 2. *(strong duality theorem) Assume that the regularity condition*

$$(CQ) \qquad \exists c' \in \mathrm{ri}(\mathrm{dom}(g)) \text{ such that } Kc' \in \bigcap_{i=1}^{l} \mathrm{ri}(\mathrm{dom}(v_i))$$

is fulfilled. Then $v(P) = v(D)$ and the dual has an optimal solution.

Proof. Since (CQ) is fulfilled, by [15, Theorem 6.5], one has that there exists $\exists c' \in \mathrm{ri}(\mathrm{dom}(g))$ such that $Kc' \in \mathrm{ri}\left(\mathrm{dom}\left(\sum_{i=1}^{l} v_i\right)\right)$. Thus, by [15, Corollary 31.2.1], there exists $\overline{p} \in \mathbb{R}^m$ such that

$$v(P) = \max_{p \in \mathbb{R}^m} \left\{ -\left(\sum_{i=1}^{l} v_i\right)^*(p) - g^*(-K^T p) \right\} = -\left(\sum_{i=1}^{l} v_i\right)^*(\overline{p}) - g^*(-K^T \overline{p}).$$

Using again (CQ), from [15, Theorem 16.4] it follows that there exist $\overline{p}^1, \ldots, \overline{p}^l \in \mathbb{R}^m$, $\sum_{i=1}^{l} \overline{p}^i = \overline{p}$, such that

$$\left(\sum_{i=1}^{l} v_i\right)^*(\overline{p}) = \min\left\{ \sum_{i=1}^{l} v_i^*(p^i) : \quad \sum_{i=1}^{l} p^i = \overline{p} \right\} = \sum_{i=1}^{l} v_i^*(\overline{p}^i).$$

Thus we get $v(P) = -\sum_{i=1}^{l} v_i^*(\overline{p}^i) - g^*\left(-K^T(\sum_{i=1}^{l} \overline{p}^i)\right) = v(D)$ and $(\overline{p}^1, \ldots, \overline{p}^l)$ is an optimal solution to the dual (D). □

The strong duality theorem plays a determinant role when deriving necessary and sufficient optimality conditions for the primal-dual pair (P)-(D).

Theorem 3. *(optimality conditions) (a) Assume that (CQ) is fulfilled. If $\overline{c} \in \mathbb{R}^n$ is an optimal solution to (P), then there exists $(\overline{p}^1, \ldots, \overline{p}^l)$, $\overline{p}^i \in \mathbb{R}^m, i = 1, \ldots, l$, an optimal solution to (D), such that the following optimality conditions are satisfied:*

(i) $v_i(K\overline{c}) + v_i^*(\overline{p}^i) = \overline{p}^{iT}(K\overline{c}), i = 1, \ldots, l;$

(ii) $g(\overline{c}) + g^*\left(-\sum_{i=1}^{l} K^T \overline{p}^i\right) + (K\overline{c})^T\left(\sum_{i=1}^{l} \overline{p}^i\right) = 0.$

(b) If $\overline{c} \in \mathbb{R}^n$ and $(\overline{p}^1, \ldots, \overline{p}^l)$ fulfill the optimality conditions $(i)-(ii)$, then they are optimal solutions to (P) and (D), respectively, and $v(P) = v(D)$.

Proof. (a) If $\overline{c}$ is an optimal solution to (P), then, by Theorem 2, there exists $(\overline{p}^1, \ldots, \overline{p}^l)$, an optimal solution to (D), such that

$$\sum_{i=1}^{l} v_i(K\overline{c}) + g(\overline{c}) = -\sum_{i=1}^{l} v_i^*(\overline{p}^i) - g^*\left(-K^T\left(\sum_{i=1}^{l}\overline{p}^i\right)\right)$$

or, equivalently,

$$\sum_{i=1}^{l}\left[v_i(K\overline{c}) + v_i^*(\overline{p}^i) - \overline{p}^{iT}(K\overline{c})\right] + \left[g(\overline{c}) + g^*\left(\sum_{i=1}^{l}K^T\overline{p}^i\right) + (K\overline{c})^T\left(\sum_{i=1}^{l}\overline{p}^i\right)\right] = 0.$$

In this way we get a sum of $l+1$ nonnegative terms (cf. the Young-Fenchel inequality) which is zero. Thus equality in these inequalities must hold and $(i) - (ii)$ are valid.

(b) All calculations done within part (a) can be carried out in reverse direction, which concludes the proof. □

Remark 1. One can easily notice that the optimality conditions from Theorem 3 can be equivalently written as

(i) $\overline{p}^i \in \partial v_i(K\overline{c}), i = 1, \ldots, l;$

(ii) $K^T\left(-\sum\limits_{i=1}^{l}\overline{p}^i\right) \in \partial g(\overline{c}).$

In other words, providing that (CQ) is fulfilled, $\bar{c} \in \mathbb{R}^n$ is an optimal solution to (P) if and only if

$$0 \in K^T\left(\sum_{i=1}^{l}\partial v_i(K\overline{c})\right) + \partial g(\overline{c}).$$

The sufficiency in the above equivalence is always valid.

We come now to the optimization problem (3)

$$\inf_{c\in\mathbb{R}^n}\left\{\sum_{i=1}^{n} v((Kc)_i, Y_i) + \frac{\lambda}{2}c^TKc\right\},$$

where $\lambda > 0$, $K \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix and $v : \mathbb{R}\times\mathbb{R} \to \overline{\mathbb{R}}$ a given cost function. For the latter we assume that for all $Y_i \in \mathbb{R}$ the function $v(\cdot, Y_i) : \mathbb{R} \to \overline{\mathbb{R}}$, $i = 1, ..., n$, is convex. Moreover, we suppose that there exists $c' \in \mathbb{R}^n$ such that $(Kc')_i \in \mathrm{dom}(v(\cdot, Y_i))$ for all $i = 1, ..., n$, which is actually a natural *feasibility condition*. These assumptions are not restrictive at all, as they are fulfilled for the majority of the cost functions that appear in the literature of machine learning. Defining $g : \mathbb{R}^n \to \mathbb{R}$ by $g(c) := \frac{\lambda}{2}c^TKc$ and $v_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ by $v_i(c) := v(c_i, Y_i)$, for $i = 1, \ldots, n$, one can easily see that problem (3) is a particular instance of (P). Recall that in our context the *labels* $Y_i \in \mathbb{R}, i = 1, \ldots, n$, are given constants.

Let us notice that, by assuming invertibility for the matrix K, Rifkin and Lippert have investigated in [14] the problem (3) from the point of view of the optimality conditions, by equivalently rewriting it as being

$$\inf_{c\in\mathbb{R}^n}\left\{\sum_{i=1}^{n} v(c_i, Y_i) + \frac{\lambda}{2}c^TK^{-1}c\right\},$$

where K^{-1} is the inverse matrix of K. As follows from the investigations made above one can provide a dual problem to (3) and then derive optimality conditions for this primal-dual pair without making this assumption. More than that, different to [14], in the formulation of the optimality conditions the cost function and the regularization appear separately.

To this aim we need the formula for the conjugate function of g, which looks like (cf. [9]):

$$g^*(p) = \begin{cases} \frac{1}{2\lambda} p^T K^- p, & \text{if } p \in \text{Im}(K), \\ +\infty, & \text{otherwise,} \end{cases} \quad \forall p \in \mathbb{R}^n,$$

where K^- is the *Moore-Penrose pseudo-inverse* of K. This leads to the following dual problem to (3)

$$\sup_{\substack{p^i \in \mathbb{R}^n, i=1,\ldots,n, \\ K\left(-\sum_{i=1}^n p^i\right) \in \text{Im}(K)}} \left\{ -\sum_{i=1}^n v_i^*(p^i) - \frac{1}{2\lambda} \left(\sum_{i=1}^n p^i\right)^T KK^-K \left(\sum_{i=1}^n p^i\right) \right\}.$$

Since, obviously, $K\left(-\sum_{i=1}^n p^i\right) \in \text{Im}(K)$, it holds

$$KK^- \left(K\left(\sum_{i=1}^n p^i\right)\right) = \text{Pr}_{\text{Im}(K)} \left(K\left(\sum_{i=1}^n p^i\right)\right) = K\left(\sum_{i=1}^n p^i\right),$$

where $\text{Pr}_{\text{Im}(K)}$ denotes the *orthogonal projection onto* $\text{Im}(K)$ and fulfills (cf. [9]) $\text{Pr}_{\text{Im}(K)}(x) = x$ for all $x \in \text{Im}(K)$. In this way we obtain the following dual problem to (3)

$$\sup_{p^i \in \mathbb{R}^n, i=1,\ldots,n} \left\{ -\sum_{i=1}^n v_i^*(p^i) - \frac{1}{2\lambda} \left(\sum_{i=1}^n p^i\right)^T K\left(\sum_{i=1}^n p^i\right) \right\}. \tag{4}$$

Remark 2. (a) In order to ensure the existence of strong duality for (3) and (4) one needs to assume that $\text{Im}(K) \cap \bigcap_{i=1}^n \text{ri}(\text{dom}(v_i)) \neq \emptyset$.

(b) In this particular instance we have $\partial g(c) = \{\lambda K c\}$ for all $c \in \mathbb{R}^n$. Thus, whenever the above regularity condition is valid and $\bar{c} \in \mathbb{R}^n$ is an optimal solution to (3), then there exists $(\overline{p}^1, \ldots, \overline{p}^n)$, $\overline{p}^i \in \mathbb{R}^n, i = 1, \ldots, n$, an optimal solution to (4), such that the following optimality conditions are satisfied:

(i) $\overline{p}^i \in \partial v_i(K\bar{c}), i = 1, \ldots, n$;

(ii) $K\left(\lambda \bar{c} + \sum_{i=1}^l \overline{p}^i\right) = 0.$

If $\bar{c} \in \mathbb{R}^n$ and $(\overline{p}^1, \ldots, \overline{p}^n)$ fulfill the optimality conditions $(i) - (ii)$ from above, then they are optimal solutions to (3) and (4), respectively, and the optimal objective values of the two problems coincide.

In other words, if $\text{Im}(K) \cap \bigcap_{i=1}^n \text{ri}(\text{dom}(v_i)) \neq \emptyset$, then $\bar{c} \in \mathbb{R}^n$ is an optimal solution to (3) if and only if

$$-\lambda K\bar{c} \in K\left(\sum_{i=1}^n \partial v_i(K\bar{c})\right).$$

The sufficiency in the above equivalence is always valid.

3 The Support Vector Machines problem

Let us consider as first particular instance of (3), the so-called Support Vector Machines problem. To this aim we assume that the training data set is given such that $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subseteq \mathbb{R}^k \times \{-1, +1\}$ and obtain, consequently, a problem from the family of *binary classification problems*. More precisely we are looking for a function $f : \mathbb{R}^k \to \mathbb{R}$ such that $f(X_i) > 0$ if $Y_i = +1$ and $f(X_i) < 0$ if $Y_i = -1$. This means that the classification is realized by the *sign-function*, i.e. for a given value X the predicted value is equal to the sign of $f(X)$ for $f(X) \neq 0$, whereas for $f(X) = 0$ we have to specify the allocation to one of the two classes. The set of points $\{X \in \mathbb{R}^k : \ f(X) = 0\}$ is called the *decision boundary*.

3.1 Hinge loss

As cost function we consider first the *hinge loss* function $v^{hl} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $v^{hl}(a, Y) = (1 - (a + b)Y)_+$, where $b \in \mathbb{R}$ is for the beginning a fixed bias term, which is one of the functions widely used in applications on Support Vector Machines Classification. Values for which $(a + b)Y \leq 1$ are penalized linearly whereas the cost function is indifferent to $(a + b)Y > 1$. Therefore problem (3) becomes the following optimization problem

$$(P^{hl}) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} (1 - ((Kc)_i + b)Y_i)_+ + \frac{\lambda}{2} c^T K c \right\}.$$

One can easily notice that for $Y_i \in \{-1, +1\}$ the function $v^{hl}(\cdot, Y_i)$ is convex and has as effective domain $\mathbb{R}$ for all $i = 1, ..., n$. Thus the feasibility condition imposed for the problem (3) is fulfilled.

Let be $i \in \{1, ..., n\}$ fixed. The conjugate function of $v_i^{hl} : \mathbb{R}^n \to \mathbb{R}$, $v_i^{hl}(c) = (1-(c_i+b)Y_i)_+$, can be calculated by employing the Lagrange duality. For $p = (p_1, ..., p_n)^T \in \mathbb{R}^n$ we have

$$-(v_i^{hl})^*(p) = \inf_{c \in \mathbb{R}^n} \{-p^T c + (1 - (c_i + b)Y_i)_+\} = \inf_{\substack{c \in \mathbb{R}^n, z \in \mathbb{R}, \\ z \geq 0, z \geq 1 - (c_i + b)Y_i}} \{-p^T c + z\} =$$

$$\sup_{q \geq 0, r \geq 0} \left\{ \inf_{c \in \mathbb{R}^n} \{(-p - re_i Y_i)^T c\} + \inf_{z \in \mathbb{R}} \{z(1 - q - r)\} + r(1 - bY_i) \right\} =$$

$$\sup_{\substack{q, r \geq 0, q + r = 1, \\ -p - re_i Y_i = 0}} r(1 - bY_i) = \sup_{\substack{r \in [0,1], \\ p + re_i Y_i = 0}} r(1 - bY_i) =$$

$$\begin{cases} -p_i(Y_i - b), & \text{if } p_i Y_i \in [-1, 0], \ \ p_j = 0, \forall j \neq i, \\ -\infty, & \text{otherwise.} \end{cases}$$

One can notice that in case $b = 0$ we rediscover the formula of the conjugate given in [14]. Now the problem (4) leads to the following dual problem to (P^{hl})

$$(D^{hl}) \qquad \sup_{\substack{p^i \in \mathbb{R}^n, P_i \in \mathbb{R}, p^i = e_i P_i, \\ p_i^i Y_i \in [-1,0], i = 1, ..., n,}} \left\{ -\sum_{i=1}^{n} p_i^i (Y_i - b) - \frac{1}{2\lambda} \left(\sum_{i=1}^{n} p^i \right)^T K \left(\sum_{i=1}^{n} p^i \right) \right\},$$

which can be equivalently written as

$$(D^{hl}) \qquad \sup_{\substack{P_i\in\mathbb{R},P_iY_i\in[-1,0],\\ i=1,\ldots,n}} \left\{-\sum_{i=1}^{n} P_i(Y_i-b)-\frac{1}{2\lambda}P^TKP\right\}.$$

That the regularity condition is fulfilled has to do with the fact that the function $v(\cdot,Y_i)$ has full effective domain for all $i=1,...,n$. Consequently, the strong duality is automatically guaranteed. In the following result we state necessary and sufficient optimality conditions for the primal-dual pair $(P^{hl})-(D^{hl})$ and these are derived via Theorem 3 and Remark 2.

Theorem 4. *(a) If $\overline{c}\in\mathbb{R}^n$ is an optimal solution to (P^{hl}), then there exists $\overline{P}=(\overline{P}_1,\ldots,\overline{P}_n)^T\in\mathbb{R}^n$, an optimal solution to (D^{hl}), such that the following optimality conditions are satisfied:*

(i) $(1-((K\overline{c})_i+b)Y_i)_+ + \overline{P}_i(Y_i-b)=\overline{P}_i(K\overline{c})_i, i=1,\ldots,n;$

(ii) $-1\leq\overline{P}_iY_i\leq 0, i=1,\ldots,n;$

(iii) $K(\lambda\overline{c}+\overline{P})=0.$

(b) If $\overline{c}\in\mathbb{R}^n$ and $\overline{P}=(\overline{P}_1,\ldots,\overline{P}_n)^T$ fulfill the optimality conditions $(i)-(iii)$, then they are optimal solutions to (P^{hl}) and (D^{hl}), respectively, and $v(P^{hl})=v(D^{hl})$.

Remark 3. (a) One should notice that in case $b=0$ (D^{hl}) becomes the dual problem given for (P^{hl}) in [14] under the assumption that K is a symmetric and positive definite matrix.

(b) By making use of some slack variables the optimization problem (P^{hl}) can be equivalently written as

$$\begin{array}{rll}(P^{hl}) & \inf\limits_{c\in\mathbb{R}^n} & \sum\limits_{i=1}^{n}\xi+\frac{\lambda}{2}c^TKc.\\ & \text{s.t.} & ((Kc)_i+b)Y_i\geq 1-\xi_i, i=1,...,m\\ & & \xi_i\geq 0, i=1,...,m\end{array} \tag{5}$$

Consequently, we rediscovered above the dual problem and the optimality conditions for the Support Vector Machines Classification problem *with fixed (or without) bias term*, which has been investigated, for instance, in [8, 13, 24].

Remark 4. In the classical formulation of the Support Vector Machines Classification problem one minimizes over both $c\in\mathbb{R}^n$ and $b\in\mathbb{R}$ (see [7, 16, 19]), the primal optimization problem having the following formulation

$$\begin{array}{rll}\inf\limits_{c\in\mathbb{R}^n,b\in\mathbb{R}} & \sum\limits_{i=1}^{n}\xi+\frac{\lambda}{2}c^TKc\\ \text{s.t.} & ((Kc)_i+b)Y_i\geq 1-\xi_i, i=1,...,m\\ & \xi_i\geq 0, i=1,...,m\end{array} \tag{6}$$

or, equivalently,

$$\inf_{b\in\mathbb{R}}\inf_{c\in\mathbb{R}^n}\left\{\sum_{i=1}^{n}(1-((Kc)_i+b)Y_i)_+ +\frac{\lambda}{2}c^TKc\right\}.$$

By making use of the dual problem of the inner infimum problem, that we determined above, we further get the following formulation for (6)

$$\inf_{b\in\mathbb{R}} \max_{\substack{P_i\in\mathbb{R},P_iY_i\in[-1,0],\\ i=1,\dots,n}} \left\{ -\sum_{i=1}^{n} P_i(Y_i-b) - \frac{1}{2\lambda}P^TKP \right\},$$

where by writing "max" instead of "sup" we want to point out the fact that the supremum is attained. Consider the function $L : \mathbb{R} \times \{P = (P_1, ..., P_n) \in \mathbb{R}^n : P_iY_i \in [-1, 0], i = 1, \dots, n\} \to \mathbb{R}$, $L(b; P) = -\sum_{i=1}^{n} P_i(Y_i - b) - \frac{1}{2\lambda}P^TKP$. As L is convex in the first variable and concave and continuous in the second one, by the classical Ky Fan minmax theorem (see, for instance, [17, Theorem 3.2]), one has that

$$\inf_{b\in\mathbb{R}} \max_{\substack{P_i\in\mathbb{R},P_iY_i\in[-1,0],\\ i=1,\dots,n}} \left\{ -\sum_{i=1}^{n} P_i(Y_i-b) - \frac{1}{2\lambda}P^TKP \right\} =$$

$$\max_{\substack{P_i\in\mathbb{R},P_iY_i\in[-1,0],\\ i=1,\dots,n}} \inf_{b\in\mathbb{R}} \left\{ -\sum_{i=1}^{n} P_i(Y_i-b) - \frac{1}{2\lambda}P^TKP \right\} =$$

$$\max_{\substack{P_i\in\mathbb{R},P_iY_i\in[-1,0],i=1,\dots,n,\\ \sum_{i=1}^{n} P_i=0}} \left\{ -\sum_{i=1}^{n} P_iY_i - \frac{1}{2\lambda}P^TKP \right\}. \tag{7}$$

The problem (7) is the classical dual optimization problem to (6) as one can find it in the literature on Support Vector Machines. Via Theorem 4 one can show that if $(\overline{c}, \overline{b}) \in \mathbb{R}^n \times \mathbb{R}$ is an optimal solution to (6), then there exists $\overline{P} = (\overline{P}_1, \dots, \overline{P}_n)^T \in \mathbb{R}^n$, an optimal solution to (7), such that the following optimality conditions are satisfied:

(i) $(1 - ((K\overline{c})_i + \overline{b})Y_i)_+ + \overline{P}_i(Y_i - \overline{b}) = \overline{P}_i(K\overline{c})_i, i = 1, \dots, n;$

(ii) $-1 \leq \overline{P}_iY_i \leq 0, i = 1, \dots, n;$

(iii) $K(\lambda\overline{c} + \overline{P}) = 0;$

(iv) $\sum_{i=1}^{n} \bar{P}_i = 0.$

These are the optimality conditions for the primal-dual pair (6)-(7) as they can be found in the above mentioned literature.

3.2 Generalized hinge loss

Chapelle considered in [5] a more general cost function than v^{hl}, the so-called *generalized hinge loss*. We slightly modify it by inserting the fixed bias term $b \in \mathbb{R}$ and, consequently, work in this subsection with $v^{ghl} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, $v^{ghl}(a, Y) = (1-(a+b)Y)_+^u$, where $u > 1$ is a given constant. Also here, for $Y_i \in \{-1, +1\}$ the function $v^{ghl}(\cdot, Y_i)$ is convex and has as effective domain $\mathbb{R}$ for all $i = 1, ..., n$. Employing it as cost function for our learning problem, it leads to the following primal optimization problem

$$(P^{ghl}) \qquad \inf_{c\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} (1 - ((Kc)_i + b)Y_i)_+^u + \frac{\lambda}{2}c^TKc \right\}.$$

For all $c \in \mathbb{R}^n$ consider $v_i^{ghl} : \mathbb{R}^n \to \mathbb{R}$, $v_i^{ghl}(c) = (1-(c_i+b)Y_i)_+^u$. In order to calculate its conjugate we notice first that for all $i = 1, ..., n$ it holds $v_i^{ghl} = k \circ v_i^{hl}$, where $k : \mathbb{R} \to \overline{\mathbb{R}}$ is defined by

$$k(x) = \begin{cases} x^u, & x \geq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Now one can use the formula for the conjugate of a composed convex function. Let $i \in \{1, ..., n\}$ and $p = (p_1, ..., p_n)^T \in \mathbb{R}^n$ be fixed. Both functions v_i^{hl} and k are convex and the latter is increasing on the set $v_i^{hl}(\mathbb{R}^n) + \mathbb{R}_+ = \mathbb{R}_+$. Since there obviously exists $c' \in \mathbb{R}^n$ with $v_i^{hl}(c') > 0$, it holds $v_i^{hl}(c') \in \text{ri}(\text{dom}(k)) \cap \text{ri}(v_i^{hl}(\mathbb{R}))$ and so, by [4, relation (1)], one gets

$$(v_i^{ghl})^*(p) = \left(k \circ v_i^{hl}\right)^*(p) = \min_{q \geq 0} \{k^*(q) + (qv_i^{hl})^*(p)\}.$$

But for all $q \in \mathbb{R}_+$ we have $k^*(q) = (u-1)\left(\frac{q}{u}\right)^{\frac{u}{u-1}}$. Further we need $(qv_i)^*(p)$. For $q > 0$, by using the formula for the conjugate of v_i^{hl} from the previous subsection, we obtain that

$$(qv_i^{hl})^*(p) = q\left(v_i^{hl}\right)^*\left(\frac{1}{q}p\right) = \begin{cases} p_i(Y_i - b), & \text{if } p_iY_i \in [-q, 0], \ \ p_j = 0, j = 1, ..., n, j \neq i \\ +\infty, & \text{otherwise,} \end{cases}$$

while for $q = 0$ it holds $(qv_i)^*(p) = \delta_{\{0\}}(p)$. In conclusion we obtain that $(v_i^{ghl})^*(p) = \min_{q \geq 0, p_iY_i \in [-q,0]} \left[(u-1)\left(\frac{q}{u}\right)^{\frac{u}{u-1}} + p_i(Y_i - b)\right]$ in case $p_j = 0$ for $j = 1, ..., n, j \neq i$, being otherwise equal to $+\infty$. Alternatively, one can derive the same formula by using the second identity of Table 3 in [14]. Thus one can provide the following dual problem to (P^{ghl})

$$(D^{ghl}) \qquad \sup_{\substack{P_i \in \mathbb{R}, q_i \geq 0, \\ P_iY_i \in [-q_i, 0], i = 1..., n}} \left\{\sum_{i=1}^n \left[(1-u)\left(\frac{q_i}{u}\right)^{\frac{u}{u-1}} - P_i(Y_i - b)\right] - \frac{1}{2\lambda}P^TKP\right\}.$$

The cost function investigated in this subsection being one with full domain, the existence of strong duality is automatically guaranteed. Next we state the corresponding optimality conditions for the primal-dual pair $(P^{ghl}) - (D^{ghl})$.

Theorem 5. *(a) If $\overline{c} \in \mathbb{R}^n$ is an optimal solution to (P^{ghl}), then there exists $(\overline{P}, \overline{q}) \in \mathbb{R}^n \times \mathbb{R}_+, \overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T$, an optimal solution to (D^{ghl}), such that the following optimality conditions are satisfied:*

(i) $(1 - ((K\overline{c})_i + b)Y_i)_+^u + (u-1)\left(\frac{\overline{q}}{u}\right)^{\frac{u}{u-1}} + \overline{P}_i(Y_i - b) = \overline{P}_i(K\overline{c})_i, i = 1, \ldots, n;$

(ii) $-\overline{q} \leq \overline{P}_iY_i \leq 0, i = 1, \ldots, n;$

(iii) $K(\lambda\overline{c} + \overline{P}) = 0.$

(b) If $\overline{c} \in \mathbb{R}^n$ and $(\overline{P}, \overline{q}) \in \mathbb{R}^n \times \mathbb{R}_+, \overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T$, fulfill the optimality conditions (i) − (iii), then they are optimal solutions to (P^{ghl}) and (D^{ghl}), respectively, and $v(P^{ghl}) = v(D^{ghl})$.

Remark 5. By means of a minmax approach, similar to the one described in Remark 4, one can provide a dual problem and optimality conditions for the problem employing the generalized hinge loss as cost function, but when minimizing over both $c \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

4 The Support Vector Regression problem

The next particular instance of the general machine learning problem we treat in this paper is the problem of Support Vector Regression. This is a technique of predictive data analysis, where one tries to estimate the dependencies between the points $\{X_1, \ldots, X_n\} \subset \mathbb{R}^k$ and $\{Y_1, \ldots, Y_n\} \subset \mathbb{R}$ of the data set, represented by means of a function f. Thus for a given point X we predict Y by $Y = f(X)$.

Here we deal first with a general abstract cost function which gathers as special case some classical cost functions used in the literature on Support Vector Regression. To this aim we consider $\varepsilon > 0$ fixed. Let be $\beta : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and increasing function with $\beta(x) \geq 0$ for all $x \in \mathbb{R}$. Define the general cost function $v^{svr} : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}}$, $v^{svr}(a, Y) = \beta(|Y - a| - \varepsilon)$. Then $v(\cdot, Y_i) : \mathbb{R} \to \overline{\mathbb{R}}$ is convex and in the following we assume that there exists $c' \in \mathbb{R}^n$ such that $|Y_i - (Kc')_i| \in \mathrm{dom}(\beta) + \varepsilon$ for $i = 1, ..., n$. In this way the feasibility condition imposed in section 2 is verified. Suppose also that $\mathrm{ri}(\mathrm{dom}(\beta)) \cap (-\varepsilon, +\infty) \neq \emptyset$, a condition which is not too restrictive since, in the particular cases treated below, it will be automatically verified. The primal optimization problem looks in this case like

$$(P^{svr}) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \beta(|Y_i - (Kc)_i| - \varepsilon) + \frac{\lambda}{2} c^T K c \right\}.$$

Let be $i \in \{1, ..., n\}$ fixed and $v_i^{svr} : \mathbb{R}^n \to \overline{\mathbb{R}}$, $v_i^{svr}(c) = v^{svr}(c_i, Y_i)$. For its conjugate at $p = (p_1, ..., p_n)^T \in \mathbb{R}^n$ we have the following formulation

$$(v_i^{svr})^*(p) = \begin{cases} p_i Y_i + (\beta \circ (|\cdot| - \varepsilon))^*(-p_i), & \text{if } p_j = 0,\ j = 1, ..., n, j \neq i, \\ +\infty, & \text{otherwise.} \end{cases}$$

Again, by [4, relation (1)], it holds

$$(\beta \circ (|\cdot| - \varepsilon))^*(-p_i) = \min_{q \geq 0} \{\beta^*(q) + (q|\cdot| - q\varepsilon)^*(-p_i)\}.$$

For all $q \geq 0$ we have

$$(q|\cdot| - q\varepsilon)^*(-p_i) = \begin{cases} \varepsilon q, & \text{if } |p_i| \leq q, \\ +\infty, & \text{otherwise} \end{cases}$$

and therefore

$$(\beta \circ (|\cdot| - \varepsilon))^*(-p_i) = \min_{q \geq 0, |p_i| \leq q} \{\beta^*(q) + \varepsilon q\}.$$

Thus, for this special choice of the cost function, the dual problem (4) turns out to be

$$(D^{svr}) \qquad \sup_{\substack{P_i \in \mathbb{R}, q_i \geq 0, \\ |P_i| \leq q_i, i = 1, ..., n}} \left\{ -\sum_{i=1}^{n} (\beta^*(q_i) + P_i Y_i + \varepsilon q_i) - \frac{1}{2\lambda} P^T K P \right\}.$$

The optimality conditions for this primal-dual pair are consequences of Theorem 3 and Remark 2.

Theorem 6. *(a) Assume that the following regularity condition*

$$\exists c' \in \mathbb{R}^n : \quad |(Kc')_i - Y_i| \in \mathrm{ri}(\mathrm{dom}(\beta)) + \varepsilon, i = 1, ..., n,$$

is fulfilled. If $\overline{c} \in \mathbb{R}^n$ is an optimal solution to (P^{svr}), then there exists $(\overline{P}, \overline{q}) \in \mathbb{R}^n \times \mathbb{R}^n_+, \overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T, \overline{q} = (\overline{q}_1, ..., \overline{q}_n)^T$, an optimal solution to (D^{svr}), such that the following optimality conditions are satisfied:

(i) $\beta(|Y_i - (K\overline{c})_i| - \varepsilon) + \beta^*(\overline{q}_i) + \overline{P}_i Y_i + \varepsilon \overline{q}_i = \overline{P}_i (K\overline{c})_i, i = 1, \ldots, n;$

(ii) $|\overline{P}_i| \leq \overline{q}_i, i = 1, \ldots, n;$

(iii) $K(\lambda \overline{c} + \overline{P}) = 0.$

(b) If $\overline{c} \in \mathbb{R}^n$ and $(\overline{P}, \overline{q}) \in \mathbb{R}^n \times \mathbb{R}^n_+, \overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T, \overline{q} = (\overline{q}_1, ..., \overline{q}_n)^T$, fulfill the optimality conditions $(i) - (iii)$, then they are optimal solutions to (P^{svr}) and (D^{svr}), respectively, and $v(P^{svr}) = v(D^{svr})$.

4.1 Extended loss

When considering $\beta : \mathbb{R} \to \overline{\mathbb{R}}$, $\beta = \delta_{\mathbb{R}_-}$, which is a proper, convex and increasing function, one obtains as cost function for the regression problem

$$v^{el} : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}}, v^{el}(a, Y) = \begin{cases} 0, & \text{if } |Y - a| \leq \varepsilon, \\ +\infty, & \text{otherwise.} \end{cases}$$

The condition $\mathrm{ri}(\mathrm{dom}(\beta)) \cap (-\varepsilon, +\infty) \neq \emptyset$ is in this case fulfilled and in order to fit in the general framework one has to impose only the feasibility condition, namely that there exists $c' \in \mathbb{R}^n$ such that $|Y_i - (Kc')_i| \leq \varepsilon$ for $i = 1, ..., n$. Consequently, we obtain the following primal optimization problem

$$(P^{el}) \qquad \inf_{\substack{c \in \mathbb{R}^n, \\ |Y_i - (Kc)_i| \leq \varepsilon, i = 1, ..., n}} \left\{ \frac{\lambda}{2} c^T K c \right\}$$

and via (D^{svr}), using that $\beta^* = \delta_{\mathbb{R}_+}$, the corresponding dual problem

$$(D^{el}) \qquad \sup_{P_i \in \mathbb{R}, i = 1, ..., n} \left\{ -\sum_{i=1}^{n} (P_i Y_i + \varepsilon |P_i|) - \frac{1}{2\lambda} P^T K P \right\}.$$

The regularity condition which ensures strong duality and the corresponding optimality conditions for this primal-dual pair follow from Theorem 6.

Theorem 7. *(a) Assume that the following regularity condition*

$$\exists c' \in \mathbb{R}^n : \quad |(Kc')_i - Y_i| < \varepsilon, i = 1, ..., n,$$

is fulfilled. If $\overline{c} \in \mathbb{R}^n$ is an optimal solution to (P^{el}), then there exists $\overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T \in \mathbb{R}^n$, an optimal solution to (D^{el}), such that the following optimality conditions are satisfied:

(i) $\overline{P}_i Y_i + \varepsilon|\overline{P}_i| = \overline{P}_i(K\overline{c})_i, i = 1, \ldots, n;$

(ii) $|Y_i - (Kc)_i| \leq \varepsilon, i = 1, \ldots, n;$

(iii) $K(\lambda\overline{c} + \overline{P}) = 0.$

(b) If $\overline{c} \in \mathbb{R}^n$ *and* $\overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T \in \mathbb{R}^n$ *fulfill the optimality conditions* $(i) - (iii)$*, then they are optimal solutions to* (P^{el}) *and* (D^{el})*, respectively, and* $v(P^{el}) = v(D^{el})$*.*

Remark 6. The important role that is played in general by the regularity conditions in the duality theory, but also in some of its particular instances, is underlined by the investigations made in this subsection. Without having such a condition fulfilled one may have serious difficulties to provide optimality conditions for the solutions of the problem (P^{el}). This is another reason why we consider that the results we present in this paper decisively improve the ones in [14].

4.2 A generalization of Vapnik's ε-insensitive loss

Smola, Schölkopf and Müller considered in [18] a cost function for the Support Vector Regression problem which generalizes the celebrated Vapnik's *ε-insensitive loss* function. They derive optimality conditions for the primal problem treated in this setting by using Wolfe duality. We show in the following that using the general approach based on conjugate duality presented in this paper one may obtain a more handleable dual problem and corresponding optimality conditions than the ones in [18].

Let $\kappa : \mathbb{R} \to \mathbb{R}$ be a convex and increasing function with $\kappa(0) = 0$ and $\kappa(x) \geq 0$ for all $x \geq 0$. Taking $\beta : \mathbb{R} \to \mathbb{R}$, $\beta(x) = 0$ for $x < 0$ and $\beta(x) = \kappa(x)$, otherwise, notice that β is a proper, convex and increasing function and it gives rise to the general cost function considered in [18]

$$v^{gil} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, v^{gil}(a, Y) = \begin{cases} 0, & \text{if } |Y - a| \leq \varepsilon, \\ \kappa(|Y - a| - \varepsilon), & \text{otherwise.} \end{cases}$$

As β has full domain, the feasibility conditions imposed at the beginning of this section are fulfilled. The primal optimization problem (P^{svr}) looks like

$$(P^{gil}) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} v^{gil}((Kc)_i, Y_i) + \frac{\lambda}{2} c^T K c \right\}$$

and again, via (D^{svr}), the corresponding dual problem becomes

$$(D^{gil}) \qquad \sup_{\substack{P_i \in \mathbb{R}, q_i \geq 0, \\ |P_i| \leq q_i, i = 1, \ldots, n}} \left\{ -\sum_{i=1}^{n} ((\kappa + \delta_{\mathbb{R}_+})^*(q_i) + P_i Y_i + \varepsilon q_i) - \frac{1}{2\lambda} P^T K P \right\}.$$

We can state the following optimality conditions, by noting that the regularity condition is in this case automatically fulfilled.

Theorem 8. *(a) If* $\overline{c} \in \mathbb{R}^n$ *is an optimal solution to* (P^{gil})*, then there exists* $(\overline{P}, \overline{q}) \in \mathbb{R}^n \times \mathbb{R}^n_+, \overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T, \overline{q} = (\overline{q}_1, ..., \overline{q}_n)^T$*, an optimal solution to* (D^{gil})*, such that the following optimality conditions are satisfied:*

(i) $v^{gil}((Kc)_i, Y_i) + (\kappa + \delta_{\mathbb{R}_+})^*(q_i) + \overline{P}_i Y_i + \varepsilon \overline{q}_i = \overline{P}_i (K\overline{c})_i, i = 1, \ldots, n;$

(ii) $|\overline{P}_i| \leq \overline{q}_i, i = 1, \ldots, n;$

(iii) $K(\lambda \overline{c} + \overline{P}) = 0.$

(b) If $\overline{c} \in \mathbb{R}^n$ *and* $(\overline{P}, \overline{q}) \in \mathbb{R}^n \times \mathbb{R}^n_+, \overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T, \overline{q} = (\overline{q}_1, ..., \overline{q}_n)^T$*, fulfill the optimality conditions* $(i) - (iii)$*, then they are optimal solutions to* (P^{gil}) *and* (D^{gil})*, respectively, and* $v(P^{gil}) = v(D^{gil})$*.*

Vapnik's ε*-insensitive loss*

$$v^{il} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, v^{il}(a, Y) = \begin{cases} 0, & \text{if } |Y - a| \leq \varepsilon, \\ |Y - a| - \varepsilon, & \text{otherwise} \end{cases}$$

arises when κ is the identity on $\mathbb{R}$. The primal problem we get in this setting is

$$(P^{il}) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} v^{il}((Kc)_i, Y_i) + \frac{\lambda}{2} c^T K c \right\}$$

and since, $(\kappa + \delta_{\mathbb{R}_+})^*(r) = \delta_{(-\infty, 1]}(r)$ for $r \in \mathbb{R}$, we obtain as dual problem to it

$$(D^{il}) \qquad \sup_{\substack{P_i \in \mathbb{R}, \\ |P_i| \leq 1, i = 1, \ldots, n}} \left\{ -\sum_{i=1}^{n} (P_i Y_i + \varepsilon |P_i|) - \frac{1}{2\lambda} P^T K P \right\}.$$

We have the following optimality conditions for the primal-dual pair $(P^{il}) - (D^{il})$.

Theorem 9. *(a) If* $\overline{c} \in \mathbb{R}^n$ *is an optimal solution to* (P^{il})*, then there exists* $\overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T \in \mathbb{R}^n$*, an optimal solution to* (D^{il})*, such that the following optimality conditions are satisfied:*

(i) $v^{il}((Kc)_i, Y_i) + \overline{P}_i Y_i + \varepsilon |P_i| = \overline{P}_i (K\overline{c})_i, i = 1, \ldots, n;$

(ii) $|\overline{P}_i| \leq 1, i = 1, \ldots, n;$

(iii) $K(\lambda \overline{c} + \overline{P}) = 0.$

(b) If $\overline{c} \in \mathbb{R}^n$ *and* $\overline{P} = (\overline{P}_1, ..., \overline{P}_n)^T \in \mathbb{R}^n$*, fulfill the optimality conditions* $(i) - (iii)$*, then they are optimal solutions to* (P^{il}) *and* (D^{il})*, respectively, and* $v(P^{il}) = v(D^{il})$*.*

Remark 7. Investigations regarding duality and optimality conditions for the Support Vector Regression problem with the ε-insensitive loss as cost functions have been previously made in [8, 16, 18].

5 Conclusions

In this paper we give optimality conditions for regularization problems, the objective function of which consists of a cost function and a regularization term, with the aim of selecting a prediction function f with a finite representation $f(\cdot) = \sum_{i=1}^{n} c_i k(\cdot, X_i)$ which

minimizes the error of prediction. The problems that arise in this context are convex optimization problems with not necessarily differentiable objective functions. Therefore, in order to provide optimality conditions for this class of problems we introduce first a dual problem, guarantee the existence of strong duality and derive, finally, the desired optimality conditions. The obtained results are employed to the Support Vector Machines problem and Support Vector Regression problem formulated for different cost functions.

We are confident that one can take advantage of the theoretical fundamentals presented in this paper for providing via the conjugate duality theory algorithmic and numerical implementations for statistical learning problems. The employment of the Fenchel duality furnishes the framework for successfully using smoothing techniques for solving the convex optimization problems which occur, in the lines of the ones developed by Nesterov in several works (see [11, 12]). This is topic of our current and future research.

Acknowledgements. The authors are thankful to anonymous reviewers for their comments which improved the quality of the paper.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 686:337–404, 1950.

[2] M. Bertero. Regularization methods for linear inverse problems. In C.G. Talenti, editor, *Inverse Problems*, volume 1225, pages 52–112. Springer-Verlag, Berlin, 1986.

[3] M. Bertero, T.A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, 1988.

[4] R.I. Boţ, S.M. Grad, and G. Wanka. New constraint qualification and conjugate duality for composed convex optimization problems. *Journal of Optimization Theory and Applications* 135:241–255, 2007.

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

[6] F.H. Clarke. *Optimization and Nonsmooth Analysis*. Canadian Mathematical Society Series of Monographs and Advanced Texts, New York, 1983.

[7] C. Cortes and V.N. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

[8] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 1999.

[9] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer-Verlag Berlin Heidelberg, 2004.

[10] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation, I. Basic Theory and II. Applications*. Series of Comprehensive Studies in Mathematics, Vol. 330, Springer-Verlag Berlin Heidelberg, 2006.

[11] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[12] Y. Nesterov. Primal-dual subgradient methods for convex problems . *Mathematical Programming*, 120(1):221–259, 2009

[13] R.M. Rifkin. *Everything Old is New Again : A Fresh Look at Historical Approaches in Machine Learning*. PhD thesis, Massachusetts Institute of Technology, 2002.

[14] R.M. Rifkin and R.A. Lippert. Value regularization and Fenchel duality. *Journal of Machine Learning Research*, 8:441–479, 2007.

[15] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[16] B. Schölkopf and A. Smola. *Learning with Kernels*. The MIT Press, Cambridge, 2002.

[17] S. Simons. *From Hahn-Banach to Monotonicity*. Lecture Notes in Mathematics, Vol. 1693, Springer-Verlag, Berlin Heidelberg, 2008.

[18] A. Smola, B. Schölkopf and K.-R. Müller. General cost functions for support vector regression. In T. Downs, M. Frean and M. Gallagher, editors, *Proceedings of the Ninth Australian Conference on Neural Networks Series*, pages 79–83. Brisbane, Australia, 1998.

[19] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[20] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-posed Problems*. W.H. Winston, Washington, D.C., 1977.

[21] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

[22] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[23] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[24] M. Vogt. SMO algorithms for Support Vector Machines without bias. *Institute Report, Institute of Automatic Control, TU Darmstadt, Darmstadt, Germany,*, 2002.

[25] G. Wahba. *Spline Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, 1990.