Optimization problems in statistical learning: duality and optimality conditions

Nicole Lorenz *

Abstract. Regularization methods are techniques for learning functions from given data. We consider regularization problems that consist of a loss 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, whereas the regulizer avoids overfitting. In general, these are convex optimization problems, for which we construct conjugate duals, by means of which we derive necessary and sufficient optimality conditions. In the second part of the paper we consider some particular cases of the general problem, namely the Support Vector Machines problem and Support Vector Regression problem. Our approach allows to avoid the use of pseudo-inverse matrices in case of finitely positive semidefinite kernel functions.

Keywords. machine learning, regularization, convex analysis, duality

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.

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

They were first mentioned by Vapnik in [15]. The reader is also referred to the books of Vapnik [14] and [16] for a deeper insight into this field.

Evgeniou et al. ([7]) distinguish between two types of statistical learning problems: the *Support Vector Machines Regression* problem (SVMR) and the *Regularization Networks* (RN). The first type has 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).

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 looks for a function $f \in \mathfrak{F}$, such that for a previously unknown value X the function f predicts the value Y.

To this end one has to define a so-called *cost* or *loss function* $v : \mathbb{R}^2 \to \overline{\mathbb{R}}$ that indicates the penalty for predicting $f(X_i)$ while the true value is $Y_i, i = 1, \ldots, n$. The problem of finding an optimal function f is ill-posed since there are infinitely many solutions. In order to get a well-posed problem, and, consequently, to be able to choose a particular solution, we need some additional apriori information about f. A common one is the assumption that the function f is smooth, i.e. one controls the complexity of the function. Therefore one has to introduce a *regularization term* $\frac{\lambda}{2}\Omega(f)$ (cf. [2], [3], [13]), where the *regularization parameter* $\lambda > 0$ controls the effect of the regularization (cf. [17]). Ω is also called *smoothness functional* or *regulizer* and it is defined such that lower values of Ω correspond to smoother functions. The following *Tikhonov regularization problem* arises:

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

Here $\sum_{i=1}^{n} v(f(X_i), Y_i) + \frac{\lambda}{2}\Omega(f)$ is the so-called *regularization functional*. By $\mathfrak{H}_k$ we denote any Reproducing Kernel Hilbert Space (RKHS) introduced by a kernel function $k : \mathbb{R}^{k \times k} \to \mathbb{R}$ (cf. [1]) and, in the following, we ask f to be an element of $\mathfrak{H}_k$. Let us further assume that k is symmetric, i.e. $k(x, y) = k(y, x), \quad \forall x, y \in \mathbb{R}^k$.

We define a *kernel matrix* $K \in \mathbb{R}^{n\times n}$ by $k(X_i, X_j) = K_{ij},\ i, j = 1, \ldots, n$, which is called the *Gram matrix of* k *with respect to* $X_1, \ldots, X_n$. This matrix is symmetric and, in addition, it is positive semidefinite if we assume that the kernel k is a finitely positive semidefinite function (see for instance [12]):

Definition 1.1. *A symmetric function* $k : \mathbb{R}^{k\times k} \to \mathbb{R}$ *which for all finite sets* $\{X_1, \ldots, X_n\} \subset \mathbb{R}^k$ *gives rise to a positive semidefinite Gram matrix* K*, i.e.*

$$\sum_{i,j=1}^{n} a_i a_j k(X_i, X_j) = a^T K a \geq 0, \quad \forall n \in \mathbb{N}, \forall a \in \mathbb{R}^n,$$

is called finitely positive semidefinite kernel.

It is well-known that in case of having a positive definite kernel and a Gram matrix, respectively, one can find a RKHS $\mathfrak{H}_k$ induced by k such that the so-called *reproducing property*

$$\forall x \in \mathbb{R}^k : \ \ f(x) = \langle f(\cdot), k(x, \cdot)\rangle$$

is fulfilled (cf. [1]).

Shawe-Taylor and Cristianini (cf. [12]) have shown how one can construct a RKHS $\mathfrak{H}_k$ for any given kernel function (even for a finitely positive semidefinite one) such that the reproducing property is valid. To this aim they considered the following space of functions

$$\left\{ \sum_{j=1}^{n} c_j k(X_j, \cdot) : n \in \mathbb{N}, X_j \in \mathbb{R}^k, c_j \in \mathbb{R}, j = 1, \ldots, n \right\}$$

along with some operations for the elements of this space. They lead to the following finite representation for $f \in \mathfrak{H}_k$, where k is a finitely positive semidefinite kernel, which is important with regard to practical applications:

$$\forall f \in \mathfrak{H}_k,\ \exists c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n : \qquad f(x) = \sum_{j=1}^{n} c_j k(x, X_j),\ \forall x \in \mathbb{R}^k.$$

This follows from the reproducing property in the Hilbert space $\mathfrak{H}_k$ induced by the kernel function k.

In the following we assume for $f \in \mathfrak{H}_k$ that $\Omega(f)$ is the squared norm of the function f in $\mathfrak{H}_k$, $\Omega(f) = ||f||_k^2 = c^T K c$, where $c \in \mathbb{R}^n$ comes from the representation of f and $||\cdot||_k$ is the norm in $\mathfrak{H}_k$. From above we get

$$f(X_i) = \sum_{j=1}^{n} c_j K_{ij} = (Kc)_i, \quad \forall i = 1, \ldots, n. \tag{2}$$

Thus the above optimization problem (1) can be written as:

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

The described regularization framework includes many well-known learning methods. Depending on the application one can use different cost functions (see for instance [7] and [10] for several examples). We give here some typical examples. In case of the Support Vector Machine Classification problem the output Y takes values in $\{1, \ldots, m\}$. For $m = 2$, i.e. $Y \in \{+1, -1\}$, we speak about a *(binary) classification problem*, whereas for $m \in \mathbb{N}$ in general we have a *ranking problem*. One can consider as cost function the *hinge loss* or *soft margin* $v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (cf. [6], [15])

$$v(a, Y) = (1 - aY)_+,$$

but also the more theoretical *hard margin* $v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, given by

$$v(a, Y) = \begin{cases} 0, & 1 - aY \leq 0, \\ 1, & 1 - aY > 0. \end{cases}$$

For the Support Vector Regression the output Y may take arbitrary real values and an appropriate cost function $v : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is

$$v(a, Y) = \delta_{[-\varepsilon, \varepsilon]}(Y - a).$$

Evgeniou et al. [7] takes as possible cost function for Regularization Networks the following *quadratic loss*:

$$v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad v(a, Y) = (Y - a)^2.$$

Of course, one can also consider generalized Tikhonov regularization problems by choosing regularization functions different from $\Omega(f) = ||f||_k^2$.

This paper is organized as follows. In the following section we introduce some definitions and notations from the convex analysis we use within the paper. In Section 3 we construct a conjugate dual for a convex optimization problem (P) and prove the weak and strong duality theorems. By using the latter we derive necessary and sufficient optimality conditions. In the last section we present some special optimization problems which occur in statistical learning, namely the Support Vector Machines problem and the Support Vector Regression problem, respectively, for which we introduce their conjugate duals and derive necessary and sufficient optimality conditions by using the general results developed in Section 3.

2 Notations and Preliminaries

For two vectors $x, x^* \in \mathbb{R}^n$ we denote by $\langle x^*, x\rangle := (x^*)^T x$ its scalar product, where the upper index ^{T} transposes a column vector into a row one and viceversa. For a set $D \subseteq \mathbb{R}^n$ we denote by $\delta_D : \mathbb{R}^n \to \overline{\mathbb{R}}$ the *indicator function* of the set D, that is defined by

$$\delta_D(x) = \begin{cases} 0, & x \in D, \\ +\infty, & \text{otherwise}, \end{cases}$$

and by $\mathrm{ri}(D)$ we denote the *relative interior* of the set D. For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ we consider its *(Fenchel-Moreau) conjugate function*, $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$, defined by

$$f^*(x^*) = \sup_{x \in \mathbb{R}^n} \{x^T x^* - f(x)\}.$$

We have the following inequality known as the *Young-Fenchel inequality*:

$$f(x) + f^*(x^*) - x^T x^* \geq 0, \ \forall x, x^* \in \mathbb{R}^n.$$

The *effective domain* of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : \ f(x) < +\infty\}$ and we say that f is *proper* if $\mathrm{dom}(f) \neq \emptyset$ and $f(x) > -\infty$, $\forall x \in \mathbb{R}^n$.
The *sign-function* $\mathrm{sgn} : \mathbb{R} \to \{-1, 0, +1\}$ is defined as follows for $x \in \mathbb{R}$:

$$\mathrm{sgn}(x) = \begin{cases} +1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

For a linear mapping $K : \mathbb{R}^n \to \mathbb{R}^m$ we denote by $\mathrm{Im}(K)$ the *image* of K, i.e. $\mathrm{Im}(K) = \{Kx : \ x \in \mathbb{R}^n\}$.

For the optimization problem (P) we denote by $v(P)$ its optimal objective value and write min (max) instead of inf (sup) if the infimum (supremum) is attained.

Definition 2.1 (infimal convolution)**.** *For the proper functions $f_1, \ldots, f_k : \mathbb{R}^n \to \overline{\mathbb{R}}$, the function $f_1 \square \cdots \square f_k : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by*

$$(f_1 \square \cdots \square f_k)(p) = \inf \left\{ \sum_{i=1}^{k} f_i(p_i) : \ \sum_{i=1}^{k} p_i = p \right\}$$

is called the infimal convolution of $f_i, i = 1, \ldots, k$.

We recall the following result (cf. [11]):

Theorem 1. *Let $f_1, \ldots, f_k : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex functions such that $\bigcap_{i=1}^{k} \mathrm{ri}(\mathrm{dom}(f_i)) \neq \emptyset$. Then for each $p \in \mathbb{R}^n$ it holds*

$$\left(\sum_{i=1}^{k} f_i \right)^* (p) = (f_1^* \square \cdots \square f_k^*)(p) = \min \left\{ \sum_{i=1}^{k} f_i^*(p_i) : \ \sum_{i=1}^{k} p_i = p \right\}.$$

For $x \in \mathbb{R}$ we define $x_+ := \max(0, x)$ and denote by $e_i = (e_{i1}, \ldots, e_{in})^T$ the i-th unit-vector in $\mathbb{R}^n$, where

$$e_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

3 Duality and optimality conditions for a general convex optimization problem

Since we want to get for the problem (3) optimality conditions by means of duality, we will first have a look on the following convex 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

$$K^{-1}\left(\bigcap_{i=1}^{l}\mathrm{dom}(v_i)\right)\cap\mathrm{dom}(g)\neq\emptyset. \tag{4}$$

Here we denote by $K^{-1}(A)$ the set $\{c\in\mathbb{R}^n : Kc\in A\}$ and condition (4) guarantees that $v(P)<+\infty$.

As a conjugate dual problem to (P) we consider

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

We prove that for (P) and (D) weak duality always holds and give a regularity condition which ensures strong duality.

Theorem 2. *For (P) and (D) weak duality holds, i.e. $v(P)\geq v(D)$.*

Proof. Let be $p_i\in\mathbb{R}^m, i=1,\ldots,l$. It holds:

$$\begin{aligned}
&\sum_{i=1}^{l} v_i^*(p_i)+g^*\left(-K^T\left(\sum_{i=1}^{l}p_i\right)\right)\\
&=\sum_{i=1}^{l}\sup_{q_i\in\mathbb{R}^m}\{q_i^Tp_i-v_i(q_i)\}+\sup_{c\in\mathbb{R}^n}\left\{-c^T\left(K^T\sum_{i=1}^{l}p_i\right)-g(c)\right\}\\
&\geq\sum_{i=1}^{l}\sup_{c\in\mathbb{R}^n}\{p_i^T(Kc)-v_i(Kc)\}+\sup_{c\in\mathbb{R}^n}\left\{-\left(\sum_{i=1}^{l}p_i\right)^T(Kc)-g(c)\right\}\\
&\geq\sup_{c\in\mathbb{R}^n}\left\{-\sum_{i=1}^{l}v_i(Kc)-g(c)\right\}.
\end{aligned}$$

From here one automatically has that

$$\sup_{\substack{p_i\in\mathbb{R}^m,\\ i=1,\dots,l}}\left\{-\sum_{i=1}^{l}v_i^*(p_i)-g^*\left(-K^T\left(\sum_{i=1}^{l}p_i\right)\right)\right\}\leq\inf_{c\in\mathbb{R}^n}\left\{\sum_{i=1}^{l}v_i(Kc)+g(c)\right\},$$

and this concludes the proof. □

Now we state the strong duality result.

Theorem 3. *Assume that the regularity condition*

$$(CQ)\qquad \exists c'\in\mathrm{ri}(\mathrm{dom}(g))\ \textit{s.t.}\ \ Kc'\in\bigcap_{i=1}^{l}\mathrm{ri}(\mathrm{dom}(v_i))$$

is fulfilled. Then between (P) *and* (D) *strong duality holds, i.e.* $v(P)=v(D)$ *and* (D) *has an optimal solution.*

Proof. The regularity condition (CQ) implies that (cf. [11, Theorem 6.5]):

$$\exists c'\in\mathrm{ri}(\mathrm{dom}(g))\quad\text{s.t.}\quad Kc'\in\mathrm{ri}\left(\mathrm{dom}\left(\sum_{i=1}^{l}v_i\right)\right).$$

Thus, by [11, Corollary 31.2.1], there exists $\overline{p}\in\mathbb{R}^m$ such that

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

Using again (CQ), by Theorem 1, there exist $\overline{p}_1,\dots,\overline{p}_l\in\mathbb{R}^m$, $\sum\limits_{i=1}^{l}\overline{p}_i=\overline{p}$, 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 $(\overline{p}_1,\dots,\overline{p}_l)$ is an optimal solution for the dual (D). □

By means of the strong duality theorem one can derive necessary and sufficient optimality conditions for the primal-dual pair (P)-(D).

Theorem 4. *a) Let $\overline{c} \in \mathbb{R}^n$ be an optimal solution for (P) and assume that (CQ) is fulfilled. Then the problem (D) has an optimal solution $(\overline{p}_1, \ldots, \overline{p}_l)$, $\overline{p}_i \in \mathbb{R}^m, i = 1, \ldots, l$, and the following optimality conditions are satisfied:*

$$(i) \quad v_i(K\overline{c}) + v_i^*(\overline{p}_i) - \overline{p}_i^T(K\overline{c}) = 0, \qquad i = 1, \ldots, l,$$

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

b) If $\overline{c} \in \mathbb{R}^n$ and $(\overline{p}_1, \ldots, \overline{p}_l)$ is feasible to (D) fulfilling the optimality conditions (i) and (ii), then $v(P) = v(D)$ and the mentioned feasible points are optimal solutions of (P) and (D), respectively.

Proof. a) Assume that $\overline{c}$ is an optimal solution for (P). Since (CQ) is fulfilled, (D) has an optimal solution $(\overline{p}_1, \ldots, \overline{p}_l)$ (cf. Theorem 3) and we have the following relations fulfilled:

$$\begin{aligned}
& v(P) = v(D) \\
\Leftrightarrow \quad & \sum_{i=1}^{l} v_i(K\overline{c}) + g(\overline{c}) = -\sum_{i=1}^{l} v_i^*(\overline{p}_i) - g^*\left(-K^T\left(\sum_{i=1}^{l}\overline{p}_i\right)\right) \\
\Leftrightarrow \quad & \left[\sum_{i=1}^{l} v_i(K\overline{c}) + \sum_{i=1}^{l} v_i^*(\overline{p}_i) - \sum_{i=1}^{l}\overline{p}_i^T(K\overline{c})\right] + \sum_{i=1}^{l}\overline{p}_i^T(K\overline{c}) \\
& \quad + \left[g(\overline{c}) + g^*\left(-K^T\left(\sum_{i=1}^{l}\overline{p}_i\right)\right) + (K\overline{c})^T\left(\sum_{i=1}^{l}\overline{p}_i\right)\right] - (K\overline{c})^T\left(\sum_{i=1}^{l}\overline{p}_i\right) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{l}\left[v_i(K\overline{c}) + v_i^*(\overline{p}_i) - \overline{p}_i^T(K\overline{c})\right] \\
& \quad + \left[g(\overline{c}) + g^*\left(\sum_{i=1}^{l} K^T\overline{p}_i\right) + (K\overline{c})^T\left(\sum_{i=1}^{l}\overline{p}_i\right)\right] = 0.
\end{aligned}$$

We get a sum of $l+1$ nonnegative terms (by the Young-Fenchel inequality) which is zero. Thus equality in these inequalities must hold and so we get (i) and (ii). (b) All calculations done within part (a) can be carried out in reverse direction and therefore the proof is complete. □

4 Application to statistical learning

In this section we consider as primal problems some optimization problems which are particular instances of

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

Here $v : \mathbb{R}\times\mathbb{R} \to \overline{\mathbb{R}}$ is a cost function, $K \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix and the function g is chosen due to Tikhonov regularization, namely as being $g : \mathbb{R}^n \to \mathbb{R}$, $g(c) = \frac{\lambda}{2}c^T Kc$, where $\lambda > 0$. In the following we fix the values $Y_i, i = 1,\ldots,n$, and define $v_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$v_i(Kc) := v((Kc)_i, Y_i), \quad i = 1,\ldots,n. \tag{5}$$

As one can see in Section 3 it is possible to derive a dual for the above problem and optimality conditions for the primal-dual pair without assuming that K is invertible as done in [10]. So we do not need any pseudo-inverse in the case of having a singular matrix K, which avoids additional expense in practical situations.

Further one can see that, using our approach, it is also possible to consider optimality conditions for the two involved functions (regularization and loss) separately. If needed, on can combine these afterwards (see also [10]).

4.1 The Support Vector Machines problem

The first particular instance that we consider is the Support Vector Machines problem. We consider the training data set

$$\{(X_1, Y_1),\ldots,(X_n, Y_n)\} \subseteq \mathbb{R}^k \times \{-1, +1\}.$$

Thus we get a problem from the class 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 $Y_i = -1$. Therefore the classification is realized by the

sign-function, i.e. for a given value X the predicted value is $Y = \operatorname{sgn}(f(X))$ 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 *decision boundary*. As cost function we consider first the *hinge loss function* $v(a, Y) = (1 - aY)_+$, which is one of the typical functions used in applications for Support Vector Machines Classification. Values for which $aY \leq 1$ are penalized linearly whereas the loss function is indifferent to $aY > 1$. Thus we get the following optimization problem:

$$(P_1) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} (1 - (Kc)_i Y_i)_+ + \frac{\lambda}{2} c^T K c \right\}.$$

Rifkin and Lippert have considered in [10] the same example, but they converted (P_1) by defining $y := Kc$ into

$$\inf_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} (1 - y_i Y_i)_+ + \frac{\lambda}{2} y^T K^{-1} y \right\},$$

under usage of the fact that K is invertible with K^{-1} its inverse matrix. As one can notice in the following one can derive a dual problem and optimality conditions for (P_1) by avoiding this restrictive assumption.

The conjugate function of g is (cf. [9]):

$$\begin{aligned} g^*(c^*) &= \sup_{c \in \mathbb{R}^n} \left\{ c^T c^* - \frac{\lambda}{2} c^T K c \right\} = \lambda \left(\frac{1}{2} \cdot^T K \cdot \right)^* \left(\frac{c^*}{\lambda} \right) \\ &= \begin{cases} \frac{1}{2\lambda} (c^*)^T K^- c^*, & c^* \in \operatorname{Im}(K), \\ +\infty, & \text{otherwise}, \end{cases} \end{aligned}$$

where K^- is the Moore-Penrose pseudo-inverse of K. On the other hand, we get for any fixed $p_i \in \mathbb{R}^n$ the conjugate function of $v_i : \mathbb{R}^n \to \mathbb{R}, \ \ v_i(p) = (1 - p_i Y_i)_+$

for $i = 1, \dots n$, by means of Lagrange duality:

$$
\begin{aligned}
-v_i^*(p_i) &= \inf_{w\in\mathbb{R}^n}\{-w^T p_i + (1 - w_i Y_i)_+\} = \inf_{\substack{w\in\mathbb{R}^n, z\in\mathbb{R},\\ z\geq 0, z\geq 1-w_i Y_i}} \{-w^T p_i + z\} \\
&= \sup_{q\geq 0, p\geq 0} \inf_{w\in\mathbb{R}^n, z\in\mathbb{R}} \{-w^T p_i + z - qz + p(1 - w_i Y_i - z)\} \\
&= \sup_{q\geq 0, p\geq 0} \left\{ \inf_{w\in\mathbb{R}^n} \{w^T(-p_i - p e_i Y_i)\} + \inf_{z\in\mathbb{R}} \{z(1 - q - p)\} + p \right\} \\
&= \sup_{\substack{p,q\geq 0, p+q=1,\\ -p_i - p e_i Y_i = 0}} p = \sup_{\substack{p\in[0,1],\\ -p_i - p e_i Y_i = 0}} p = \begin{cases} -\frac{p_{ii}}{Y_i}, & \text{if } -\frac{p_{ii}}{Y_i} \in [0,1],\ \ p_{ij} = 0, \forall j \neq i, \\ -\infty, & \text{otherwise.} \end{cases}
\end{aligned}
$$

These relations lead to the following dual problem to (P_1):

$$
\begin{aligned}
(D_1) \qquad & \sup_{p_i\in\mathbb{R}^n, i=1,\dots,n} \left\{ -\sum_{i=1}^n v_i^*(p_i) - g^*\left(-K\left(\sum_{i=1}^n p_i\right)\right)\right\} \\
&= \sup_{\substack{p_i\in\mathbb{R}^n, P_i\in\mathbb{R}, i=1,\dots,n,\\ p_i = e_i P_i, -\frac{p_{ii}}{Y_i}\in[0,1],\\ K\left(\sum_{i=1}^n p_i\right)\in \mathrm{Im}(K)}} \left\{ -\sum_{i=1}^n \frac{p_{ii}}{Y_i} - \frac{1}{2\lambda}\left(\sum_{i=1}^n p_i\right)^T K K^- K \left(\sum_{i=1}^n p_i\right)\right\}.
\end{aligned}
$$

Since the condition $K\left(\sum_{i=1}^n p_i\right) \in \mathrm{Im}(K)$ is always fulfilled and it holds

$$
K K^- \left(K\left(\sum_{i=1}^n p_i\right)\right) = p_{\mathrm{Im}(K)}\left(K\left(\sum_{i=1}^n p_i\right)\right) = K\left(\sum_{i=1}^n p_i\right),
$$

where the operator $p_{\mathrm{Im}(K)}$ is the orthogonal projection onto $\mathrm{Im}(K)$ and fulfills (cf. [9])

$$
p_{\mathrm{Im}(K)}(x) = x, \qquad \forall x \in \mathrm{Im}(K),
$$

we get the following dual:

$$
(D_1) \qquad \sup_{\substack{p_i\in\mathbb{R}^n, P_i\in\mathbb{R}, i=1,\dots,n,\\ p_i = e_i P_i, -\frac{p_{ii}}{Y_i}\in[0,1]}} \left\{ -\sum_{i=1}^n \frac{p_{ii}}{Y_i} - \frac{1}{2\lambda}\left(\sum_{i=1}^n p_i\right)^T K \left(\sum_{i=1}^n p_i\right)\right\}.
$$

By defining $P := (P_1, \dots, P_n)^T \in \mathbb{R}^n$ we get $p_{ii} = P_i$ and $\sum_{i=1}^n p_i = P$ and thus the dual looks like

$$
(D_1) \qquad \sup_{\substack{P_i\in\mathbb{R}, -\frac{P_i}{Y_i}\in[0,1],\\ i=1,\dots,n}} \left\{ -\sum_{i=1}^n \frac{P_i}{Y_i} - \frac{1}{2\lambda} P^T K P\right\}.
$$

In this way one obtains the dual given for (P_1) in [10]. The regularity condition (CQ) is always fulfilled, since $\mathrm{ri}(\mathrm{dom}(g)) = \mathbb{R}^n$ and $\mathrm{ri}(\mathrm{dom}(v_i)) = \mathbb{R}^n, i = 1, \ldots, n$.

We can state now the following theorem, which gives necessary and sufficient optimality conditions for (P_1) and (D_1).

Theorem 5. *a) Let $\overline{c} \in \mathbb{R}^n$ be an optimal solution of (P_1). Then (D_1) has an optimal solution $\overline{P} = (\overline{P}_1, \ldots, \overline{P}_n)^T \in \mathbb{R}^n$ such that the following optimality conditions are satisfied:*

$$(i) \quad (1 - (K\overline{c})_i Y_i)_+ + \frac{\overline{P}_i}{Y_i} - \overline{P}_i e_i^T K\overline{c} = 0, \qquad i = 1, \ldots, n,$$

$$(ii) \quad 0 \leq -\frac{\overline{P}_i}{Y_i} \leq 1, \qquad i = 1, \ldots, n,$$

$$(iii) \quad \frac{\lambda}{2}\overline{c}^T K\overline{c} + \frac{1}{2\lambda}\overline{P}^T K\overline{P} + \overline{c}^T(K\overline{P}) = 0.$$

b) If $\overline{c} \in \mathbb{R}^n$ and $\overline{P} = (\overline{P}_1, \ldots, \overline{P}_n)^T$ is feasible to (D_1) fulfilling the optimality conditions $(i) - (iii)$, then $v(P_1) = v(D_1)$ and the mentioned feasible points are optimal solutions of (P_1) and (D_1), respectively.

As mentioned e.g. in [5] one can consider a more general loss function, the so-called *generalized hinge loss* $v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, given by

$$\widetilde{v}(a, Y) = (1 - aY)_+^u,$$

where $u > 1$, and consider the following primal problem (P_2):

$$(P_2) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} (1 - (Kc)_i Y_i)_+^u + \frac{\lambda}{2} c^T K c \right\}.$$

Above we have considered the case $u = 1$. Now we define $\widetilde{v_i}(p) = (1 - p_i Y_i)_+^u$, $\widetilde{v_i} : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, n$. This loss function can be written as $\widetilde{v_i} = h \circ v_i, i = 1, \ldots, n$, where v_i was given above and the function $h : \mathbb{R} \to \overline{\mathbb{R}}$ is defined by

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

The conjugate function of $v_i, i = 1, \ldots, n$, can be obtained by using the formula existing in the literature for the conjugate of a composed convex function.

Theorem 6. ([18]) *Let $\mathcal{Z}$ be a separated locally convex space and $f : \mathcal{Z} \to \mathbb{R}$, $h : \mathbb{R} \to \overline{\mathbb{R}}$ be convex functions such that h is increasing on $f(\mathcal{Z}) + [0, +\infty)$. We assume that there exists $x' \in \mathcal{Z}$ such that $f(x') \in \mathrm{dom}(h)$ and h is continuous at $f(x')$. Then for all $x^* \in \mathcal{Z}^*$ one has*

$$(h \circ f)^*(x^*) = \min_{\beta \geq 0} \{h^*(\beta) + (\beta f)^*(x^*)\}. \tag{6}$$

In our case we have $\mathcal{Z} = \mathbb{R}^n$, $f = v_i$ and h is increasing on $v_i(\mathcal{Z}) + [0, +\infty) = [0, +\infty)$. Further we assume that for all $i = 1, \ldots, n$ we have

$$\begin{aligned} & \exists x' \in \mathbb{R}^n : \quad v_i(x') \in \mathrm{dom}(h) = \mathbb{R}_+ \text{ and } h \text{ is continuous at } v_i(x') \\ \Leftrightarrow \quad & \exists x' \in \mathbb{R}^n : \quad v_i(x') > 0. \end{aligned} \tag{7}$$

We have (cf. [4]):

$$h^*(\beta) = \begin{cases} (u-1)\left(\frac{\beta}{u}\right)^{\frac{u}{u-1}}, & \beta \geq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Further we need $(\beta v_i)^*(p_i)$. For $\beta > 0$ we get the following:

$$(\beta v_i)^*(p_i) = \beta v_i^*\left(\frac{p_i}{\beta}\right) = \begin{cases} \frac{p_{ii}}{Y_i}, & \text{if } -\frac{p_{ii}}{Y_i} \in [0, \beta], \;\; p_{ij} = 0, \forall j \neq i, \\ +\infty, & \text{otherwise.} \end{cases} \tag{8}$$

For $\beta = 0$ it holds:

$$\begin{aligned} (-\beta v_i)^*(p_i) &= \inf_{w \in \mathbb{R}^n} \{-w^T p_i + \beta(1 - w_i Y_i)_+\} = \inf_{w \in \mathbb{R}^n} \{-w^T p_i\} \\ &= \begin{cases} 0, & p_i = 0, \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

In conclusion for all $\beta \geq 0$ formula (8) holds and by Theorem 6 and assumption

(7) we get for all $i = 1, \ldots, n$:

$$\begin{aligned}\widetilde{v_i}^*(p_i) = (h \circ v_i)^*(p_i) &= \min_{\beta \geq 0} \{h^*(\beta) + (\beta v_i)^*(p_i)\} \\ &= \min_{\substack{\beta \geq 0, \\ -\frac{p_{ii}}{Y_i} \in [0,\beta], \\ p_{ij}=0, \forall j \neq i}} \left\{ (u-1) \left(\frac{\beta}{u} \right)^{\frac{u}{u-1}} + \frac{p_{ii}}{Y_i} \right\}.\end{aligned}$$

Thus one can derive, by defining $P = (P_1, \ldots, P_n)^T \in \mathbb{R}^n$ as above, the following dual problem (D_2):

$$\begin{aligned}(D_2) \qquad &\sup_{P_i \in \mathbb{R}, i=1\ldots,n} \left\{ -\sum_{i=1}^{n} \min_{\substack{\beta_i \geq 0, \\ -\frac{P_i}{Y_i} \in [0,\beta_i]}} \left\{ (u-1) \left(\frac{\beta_i}{u} \right)^{\frac{u}{u-1}} + \frac{P_i}{Y_i} \right\} - \frac{1}{2\lambda} P^T K P \right\} \\ = &\sup_{\substack{P_i \in \mathbb{R}, \beta_i \geq 0, \\ -\frac{P_i}{Y_i} \in [0,\beta_i], \\ i=1\ldots,n}} \left\{ \sum_{i=1}^{n} \left\{ (1-u) \left(\frac{\beta_i}{u} \right)^{\frac{u}{u-1}} - \frac{P_i}{Y_i} \right\} - \frac{1}{2\lambda} P^T K P \right\}.\end{aligned}$$

Since the (generalized) hinge loss penalizes only values $aY \leq 1$ but is indifferent to values $aY > 1$ one can take as loss function also the *negative binomial log-likelihood deviance* (see [8]). In Figure 1 a comparison between this and the (generalized) hinge loss (for $u = 1$ and $u = 3$) is made. The negative binomial log-likelihood deviance is a convex function and is given by:

$$v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad v(a, Y) = \ln(1 + \exp(-aY)).$$

Again we use the notation $v((Kc)_i, Y_i) = v_i(Kc)$. For the calculation of the dual problem of

$$(P_3) \qquad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \ln(1 + \exp(-(Kc)_i Y_i)) + \frac{\lambda}{2} c^T K c \right\}$$

we need the conjugate function v_i^* of $v_i(p) = \ln(1 + \exp(-p_i Y_i))$, $v_i : \mathbb{R}^n \to \mathbb{R}$,

for fixed $p_i \in \mathbb{R}^n, i = 1, \ldots, n$. It holds:

$$
\begin{aligned}
-v_i^*(p_i) &= \inf_{w \in \mathbb{R}^n} \{-w^T p_i + \ln(1 + \exp(-w_i Y_i))\} \\
&= \begin{cases} \inf_{w \in \mathbb{R}} \{-w p_{ii} + \ln(1 + \exp(-w Y_i))\}, & p_{ij} = 0,\ \forall j \neq i, \\ -\infty, & \text{otherwise}, \end{cases} \\
&= \begin{cases} \frac{p_{ii}}{Y_i} \ln\left(-\frac{p_{ii}}{Y_i + p_{ii}}\right) + \ln\left(\frac{Y_i}{Y_i + p_{ii}}\right), & p_{ij} = 0,\ \forall j \neq i,\ \frac{p_{ii}}{p_{ii} + Y_i} < 0,\ p_{ii} \neq -Y_i, \forall i, \\ 0, & p_{ij} = 0,\ \forall j \neq i,\ p_{ii} = -Y_i,\ \forall i \\ -\infty, & \text{otherwise}. \end{cases}
\end{aligned}
$$

For the dual problem of (P_3) we have to calculate $\sup_{\substack{p_i \in \mathbb{R}^n, \\ i=1,\ldots,n}} -v_i^*(p_i)$. Therefore we have to show, that under the condition

$$p_{ij} = 0,\ \forall j \neq i,\ \frac{p_{ii}}{p_{ii} + Y_i} < 0,\ p_{ii} \neq -Y_i, \forall i.$$

we have

$$f(p_{ii}) := \frac{p_{ii}}{Y_i} \ln\left(-\frac{p_{ii}}{Y_i + p_{ii}}\right) + \ln\left(\frac{Y_i}{Y_i + p_{ii}}\right) \geq 0.$$

Since $Y_i \in \{-1, +1\}$ we consider two cases. First we have $Y_i = +1$ and it follows

$$\frac{p_{ii}}{p_{ii} + Y_i} = \frac{p_{ii}}{p_{ii} + 1} < 0 \quad \Leftrightarrow \quad p_{ii} \in (-1, 0).$$

Thus we get $f(p_{ii}) > 0$. Second we consider $Y_i = -1$. There we have

$$\frac{p_{ii}}{p_{ii} + Y_i} = \frac{p_{ii}}{p_{ii} - 1} < 0 \quad \Leftrightarrow \quad p_{ii} \in (0, 1),$$

and we get $f(p_{ii}) > 0$.

We get the following dual (D_3):

$$
\begin{aligned}
(D_3) \qquad & \sup_{p_i\in\mathbb{R}^n, i=1,\ldots,n}\left\{-\sum_{i=1}^n v_i^*(p_i) - g^*\left(-K\left(\sum_{i=1}^n p_i\right)\right)\right\} \\
= & \sup_{\substack{p_i\in\mathbb{R}^n,\ \exists P_i\in\mathbb{R}: p_i=e_iP_i,\\ \frac{p_{ii}}{p_{ii}+Y_i}<0,\ p_{ii}\neq -Y_i, \forall i=1,\ldots,n,\\ K\left(\sum_{i=1}^n p_i\right)\in \mathrm{Im}(K)}} \left\{\sum_{i=1}^n\left(\frac{p_{ii}}{Y_i}\ln\left(-\frac{p_{ii}}{p_{ii}+Y_i}\right)+\ln\left(\frac{Y_i}{p_{ii}+Y_i}\right)\right)\right. \\
& \qquad \left. -\frac{1}{2\lambda}\left(\sum_{i=1}^n p_i\right)^T KK^-K\left(\sum_{i=1}^n p_i\right)\right\} \\
= & \sup_{\substack{p_i\in\mathbb{R}^n,\ \exists P_i\in\mathbb{R}: p_i=e_iP_i,\\ \frac{p_{ii}}{p_{ii}+Y_i}<0,\ p_{ii}\neq -Y_i, \forall i=1,\ldots,n}} \left\{\sum_{i=1}^n\left(\frac{p_{ii}}{Y_i}\ln\left(-\frac{p_{ii}}{p_{ii}+Y_i}\right)+\ln\left(\frac{Y_i}{p_{ii}+Y_i}\right)\right)\right. \\
& \qquad \left. -\frac{1}{2\lambda}\left(\sum_{i=1}^n p_i\right)^T K\left(\sum_{i=1}^n p_i\right)\right\}.
\end{aligned}
$$

Again, by defining $P := (P_1,\ldots,P_n)^T \in \mathbb{R}^n$ we get $p_{ii} = P_i$ and $\sum\limits_{i=1}^n p_i = P$ and thus the dual looks like

$$
(D_3) \qquad \sup_{\substack{P_i\in\mathbb{R}, \frac{P_i}{P_i+Y_i}<0,\\ P_i\neq -Y_i,\, i=1,\ldots,n}} \left\{\sum_{i=1}^n\left(\frac{P_i}{Y_i}\ln\left(-\frac{P_i}{P_i+Y_i}\right)+\ln\left(\frac{Y_i}{P_i+Y_i}\right)\right) - \frac{1}{2\lambda}P^TKP\right\}.
$$

When formulating the optimality conditions for the primal-dual pair $(P_3)-(D_3)$, instead of the conditions (i) and (ii) in Theorem 5, we have the following ones:

$$
(i) \quad \ln(1+\exp(-(K\overline{c})_iY_i)) - \frac{\overline{P}_i}{Y_i}\ln\left(-\frac{\overline{P}_i}{\overline{P}_i+Y_i}\right) - \ln\left(\frac{Y_i}{\overline{P}_i+Y_i}\right) - \overline{P}_ie_i^TK\overline{c} = 0, \quad i=1,\ldots,n,
$$

$$
(ii) \quad \frac{P_i}{P_i+Y_i} < 0, \quad i=1,\ldots,n,
$$

$$
(iii) \quad P_i \neq -Y_i, \quad i=1,\ldots,n.
$$

4.2 The Support Vector Regression problem

The next particular instance we treat 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 a function f. Later for some given point X we predict Y by $Y = f(X)$.

We consider a loss function typical for Support Vector Regression problems, which was also mentioned in the introduction, $v : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}}$,

$$v(a, Y) = \delta_{[-\varepsilon,\varepsilon]}(Y - a).$$

We use again the representation (2) and get the following primal problem (P_4):

$$(P_4) \quad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \delta_{[-\varepsilon,\varepsilon]}(Y_i - (Kc)_i) + \frac{\lambda}{2} c^T K c \right\}.$$

In order to get strong duality and derive necessary and sufficient optimality conditions one has to derive a regularity condition from the general problem in Theorem 3. Therefore we have $w \in \mathrm{dom}(v_i) \Leftrightarrow w_i \in [Y_i - \varepsilon, Y_i + \varepsilon]$ and get

$$\mathrm{ri}(\mathrm{dom}(v_i)) = \mathbb{R} \times \ldots \times \mathbb{R} \times (Y_i - \varepsilon, Y_i + \varepsilon) \times \mathbb{R} \times \ldots \times \mathbb{R}, \ i = 1, \ldots, n,$$

and

$$\bigcap_{i=1}^{n} \mathrm{ri}(\mathrm{dom}(v_i)) = \prod_{i=1}^{n} (Y_i - \varepsilon, Y_i + \varepsilon),$$

which becomes part of the regularity condition.

In order to calculate the dual problem we need the conjugate function of v_i for $i = 1, \ldots, n$:

$$\begin{aligned}
v_i^*(p_i) &= \sup_{w \in \mathbb{R}^n} \{w^T p_i - \delta_{[-\varepsilon,\varepsilon]}(Y_i - w_i)\} = \sup_{w \in \mathbb{R}^n} \{w^T p_i - \delta_{[Y_i - \varepsilon, Y_i + \varepsilon]}(w_i)\} \\
&= \begin{cases} \max\{p_{ii}(Y_i - \varepsilon),\, p_{ii}(Y_i + \varepsilon)\}, & p_{ij} = 0, j \neq i, \\ +\infty, & \text{otherwise}, \end{cases} \\
&= \begin{cases} p_{ii} Y_i - \varepsilon \max\{p_{ii}, -p_{ii}\}, & p_{ij} = 0, j \neq i, \\ +\infty, & \text{otherwise}, \end{cases} \\
&= \begin{cases} p_{ii} Y_i - \varepsilon |p_{ii}|, & \text{if } \exists P_i \in \mathbb{R} : \ p_i = e_i P_i, \\ +\infty, & \text{otherwise}. \end{cases}
\end{aligned}$$

Taking $P = (P_1, \ldots, P_n)^T$ leads to $\sum\limits_{i=1}^{n} p_i = P$, and we get the following dual problem to (P_4):

$$(D_4) \qquad \sup_{p_i \in \mathbb{R}^n, i=1,\ldots,n} \left\{ -\sum_{i=1}^{n} v_i^*(p_i) - g^* \left(-K \left(\sum_{i=1}^{n} p_i \right) \right) \right\}$$
$$= \sup_{P_i \in \mathbb{R}, i=1,\ldots,n} \left\{ \sum_{i=1}^{n} (\varepsilon |P_i| - P_i Y_i) - \frac{1}{2\lambda} P^T K P \right\}.$$

By Theorem 3 we obtain the following result:

Theorem 7. *a) Let $\overline{c} \in \mathbb{R}^n$ be an optimal solution of (P_4) and assume that the regularity condition*

$$(CQ) \qquad \exists c' \in \mathbb{R}^n : \quad Kc' \in \prod_{i=1}^{n} (Y_i - \varepsilon, Y_i + \varepsilon)$$

is fulfilled. Then the problem (D_4) has an optimal solution $\overline{P} = (\overline{P}_1, \ldots, \overline{P}_n)^T \in \mathbb{R}^n$ and the following optimality conditions are satisfied:

$$(i) \quad \overline{P}_i Y_i - \varepsilon |\overline{P}_i| - \overline{P}_i e_i^T K \overline{c} = 0, \qquad i = 1, \ldots, n,$$
$$(ii) \quad |Y_i - (K\overline{c})_i| \leq \varepsilon, \qquad i = 1, \ldots, n,$$
$$(iii) \quad \frac{\lambda}{2} \overline{c}^T K \overline{c} + \frac{1}{2\lambda} \overline{P}^T K \overline{P} + \overline{c}^T (K\overline{P}) = 0.$$

b) If $\overline{c}$ is feasible to (P_4) and $(\overline{P}_1, \ldots, \overline{P}_n)^T$ is feasible to (D_4) fulfilling the optimality conditions (i) - (iii), then $v(P_4) = v(D_4)$ and the mentioned feasible points are optimal solutions of (P_4) and (D_4), respectively.

Remark 4.1. *As one can see, the regularity condition (CQ) is not always fulfilled. The given example is a good sample to realize the importance of the given condition for having strong duality. This is a decisive improvement towards the paper of Rifkin and Lippert [10].*

Acknowlegdements. The author would like to thank Dr. R.-I. Boţ for his valueable and helpful suggestions.

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, N. Lorenz, and G. Wanka. Some formulas for the conjugate of convex risk measures. *Preprint 2006-19*, Faculty of Mathematics, Chemnitz University of Technology, 2006.

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

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

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

[8] T. Hasti, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

[9] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer, 2004.

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

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

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

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

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

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

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

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

[18] C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific Publishing, 2002.

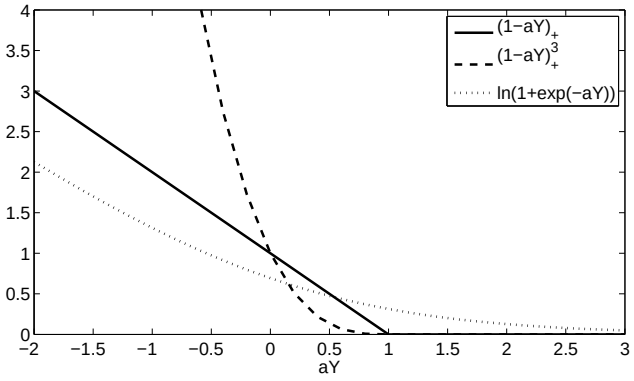

Figure 1: Comparison of the loss functions $v(a, Y) = (1 - aY)_+$, $v(a, Y) = (1 - aY)_+^3$ and $v(a, Y) = \ln(1 + \exp(-aY))$.