Regression tasks in machine learning via Fenchel duality

Radu Ioan Boţ * André Heinrich †

April 27, 2012

Abstract. Supervised learning methods are powerful techniques to learn a function from a given set of labeled data, the so-called training data. In this paper the support vector machines approach for regression is investigated under a theoretical point of view that makes use of convex analysis and Fenchel duality. Starting with the corresponding Tikhonov regularization problem, reformulated as a convex optimization problem, we introduce a conjugate dual problem to it and prove that, whenever strong duality holds, the function to be learned can be expressed via the optimal solutions of the dual problem. Corresponding dual problems are then derived for different loss functions. The theoretical results are applied by numerically solving the regression task for two data sets and the accuracy of the regression when choosing different loss functions is investigated.

Keywords. machine learning, Tikhonov regularization, conjugate duality, support vector regression

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

1 Introduction

Supervised learning methods such as Support Vector Machines for Classification and Regression belong to the class of kernel based methods that have become, especially in the last decade, a popular approach for learning functions from a given set of labeled data. They have wide fields of application such as image and text classification (cf. [6]), computational biology (cf. [8]) or time series forecasting and credit scoring (cf. [7, 15]) and have proven to be able to provide good results.

In this paper we deal in particular with the Support Vector Regression Problem. Starting with the general Tikhonov regularization problem (cf. [14]), to which the supervised learning problem gives rise, which turns out to be a convex (not necessarily differentiable) optimization problem, we construct a conjugate dual to it (see, for instance, [2]), prove under suitable qualification conditions the existence of strong duality and express the optimal solutions of the primal problem via the ones of the dual. This has as consequence the formulation of the regression function to be learned by means of the optimal solutions of the dual problem. Hence, for the specific learning task one

*Department of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: radu.bot@mathematik.tu-chemnitz.de.

†Department of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: andre.heinrich@mathematik.tu-chemnitz.de.

only has to numerically solve the dual problem, which, different to the primal one, is mainly a convex differentiable optimization problem. Our scope in this article is not only to provide the necessary theoretical background of this approach, but also to derive the corresponding dual problems for some popular loss functions and to compare the accuracy of the regression for two data sets widely used as benchmarks in the literature (cf. [4, 10, 11]).

The paper is organized as follows. In Section 2 the general regularization problem is introduced and it is stated as an equivalent convex optimization problem. A Fenchel-type dual problem to it is provided and, under a suitable weak qualification condition, the existence of strong duality for this primal-dual pair is proved, which gives rise to the formulation of necessary and sufficient optimality conditions. In Section 3 the general theory from the previous section is employed for several particular loss functions and the corresponding dual programs are calculated. In Section 4 the dual programs are solved numerically for two data sets. First, the dual problems resulting from the several choices of the corresponding loss function are transformed into equivalent representations, easier to handle when solving them numerically. Then, we compute the regression functions first on the basis of a toy data set generated from the sinc-function for a fixed set of parameters. After that we solve the regression problem based on the Boston Housing data set and compare the performances of the different resulting regression functions.

2 Supervised learning based on conjugate duality

Given a set of *training data* $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and the corresponding observed values $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, a common approach for learning a regression function based on the *Structural Risk Minimization Principle* is to apply *Support Vector Machines* (SVM) techniques for regression. These supervised learning methods were investigated in detail by Vapnik in [16]. Considering $\mathcal{D} = \{(x_i, y_i) : i = 1, \ldots, n\} \subset \mathbb{R}^d \times \mathbb{R}$ the *training set*, the aim of the SVM approach is to find a function f belonging to $\mathcal{F}$, a space of real valued functions defined on $\mathbb{R}^d$ enhanced with some a priori information, that best approximates the given data.

A so-called *loss function* $v : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, assumed to be proper and convex in its first variable, enables to impose a penalty for predicting $f(x_i)$ while the true, or observed, value is y_i, for $i = 1, \ldots, n$. One of the common assumptions on f is *smoothness*, which guarantees that two similar inputs correspond to two similar outputs. In order to control it, one needs to consider a *smoothness functional* $\Omega : \mathcal{F} \to \mathbb{R}$ (cf. [14]) having the desired characteristic of taking high values for non-smooth functions and low values for smooth ones.

Hence, the desired function f will be the optimal solution of the *Tikhonov regularization problem*

$$\inf_{f \in \mathcal{F}} \left\{ C \sum_{i=1}^{n} v(f(x_i), y_i) + \frac{1}{2}\Omega(f) \right\} \tag{1}$$

where $C > 0$ is the so-called *regularization parameter* controlling the tradeoff between the accuracy and the generalization ability of the learned regression function (see [3]). In the following the function f is assumed to be an element of the Reproducing Kernel

Hilbert Space (RKHS) $\mathcal{H}_k$ induced by a continuous *kernel function* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (cf. [1]) which we assume to be *symmetric* and *finitely positive semidefinite*. The kernel k is said to be symmetric if $k(x, y) = k(y, x)$ for all $x, y \in \mathbb{R}^d$. A symmetric kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which for all $m \geq 1$ and all finite sets $\{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ fulfills $\sum_{i,j=1}^{m} a_i a_j k(x_i, x_j) \geq 0$ for every arbitrary $a \in \mathbb{R}^d$ is called *finitely positive semidefinite* (cf. [12]).

Hence, the kernel function k can be decomposed as $k(x, y) = \langle \phi(x), \phi(y) \rangle_k$, where $\langle \cdot, \cdot \rangle_k$ denotes the *inner product* of $\mathcal{H}_k$ and $\phi : \mathbb{R}^d \to \mathcal{H}_k$ is a so-called *feature map*. The *representer theorem* (cf. [17]) ensures that for every minimizer f of (1) there exists a vector $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ such that

$$f(\cdot) = \sum_{i=1}^{n} c_i k(\cdot, x_i). \tag{2}$$

For $i = 1, ..., n$ a vector x_i with the property that the corresponding coefficient c_i is not equal to zero is a so-called *support vector*.

The existence of such a representation is essential for the purpose of this paper. Finally, we define the smoothness functional Ω to be $\Omega(f) = \|f\|_k^2$ for $f \in \mathcal{H}_k$, where $\|\cdot\|_k$ denotes the norm on $\mathcal{H}_k$. The *Gram matrix* of k with respect to the set $\{x_1, \ldots, x_n\}$ is denoted by $K \in \mathbb{R}^{n \times n}$, being the matrix with entries $K_{ij} := k(x_i, x_j)$, $i, j = 1, \ldots, n$. Obviously, K is symmetric and positive semidefinite. Taking $c \in \mathbb{R}^n$ to be the vector corresponding to representation (2), the smoothness functional becomes $\Omega(f) = \|f\|_k^2 = c^T K c$ and for $i = 1, ..., n$ it holds $f(x_i) = \sum_{j=1}^{n} c_j K_{ij} = (Kc)_i$. Thus we can rewrite optimization problem (1) equivalently as

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

Due to the nature of the loss function, this problem is mainly a convex and not necessarily differentiable optimization problem. In order to overcome this disadvantage, we provide a conjugate dual problem to it, prove the existence of strong duality and express the optimal solutions of (P_{gen}) via the ones of the dual. These considerations make sense, especially when the dual problem is easier to solve than the primal one, which is actually the case for the majority of the loss functions used for regression problems.

In order to make the paper self-contained, we introduce first some notions and results. On $\mathbb{R}^d$ we consider the Euclidian norm, while for two vectors $x, y \in \mathbb{R}^d$ we denote by $x^T y$ their inner product, where the upper index ^{T} transposes a column vector into a row one and viceversa. For a nonempty set $D \subseteq \mathbb{R}^n$ we denote by $\mathrm{ri}(D)$ the *relative interior* of the set D, that is the interior of D relative to its affine hull. The indicator function of D is defined as

$$\delta_D : \mathbb{R}^n \to \overline{\mathbb{R}}, \quad \delta_D(x) = \begin{cases} 0, & \text{if } x \in D, \\ +\infty, & \text{otherwise.} \end{cases}$$

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 $f^*(p) = \sup_{x \in \mathbb{R}^n}\{p^T x - f(x)\}$. For all $x, p \in \mathbb{R}^n$ we have the following relation, known as the *Young-Fenchel inequality*, $f(x) + f^*(p) - p^T x \geq 0$. 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

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

The *epigraph* of f is $\operatorname{epi} f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq r\}$ and f is said to be *convex*, if $\operatorname{epi} f$ is a convex set, while f is said to be *lower semicontinuous*, if $\operatorname{epi} f$ is a closed set. Having a convex set D and a function $f : D \to \mathbb{R}$, we say that f is *strictly convex on D*, if

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y)\ \forall x, y \in D, x \neq y,\ \forall \lambda \in (0, 1)$$

and that f is *strongly convex on D*, if there exists $\mu > 0$ such that

$$f(\lambda x + (1 - \lambda)y) + \lambda(1 - \lambda)\mu\|x - y\|^2 \leq \lambda f(x) + (1 - \lambda)f(y)\ \forall x, y \in D\ \forall \lambda \in (0, 1).$$

When $K \in \mathbb{R}^{n \times n}$ is a given matrix, we denote by $\operatorname{Im} K := \{Kx : x \in \mathbb{R}^n\}$. Further, for $x \in \mathbb{R}$ we define $x_+ := \max\{0, x\}$.

The dual problem to (P_{gen}) which we consider here is a Fenchel-type dual problem and it is formulated as

$$(D_{gen}) \qquad \sup_{\substack{P \in \mathbb{R}^n, \\ P = (P_1, \ldots, P_n)^T}} \left\{ -C \sum_{i=1}^{n} \Big(v(\cdot, y_i)\Big)^* \Big(-\frac{P_i}{C}\Big) - \frac{1}{2} P^T K P \right\} \tag{4}$$

in analogy with the duality concept discussed in [3]. However, the more detailed formulation of (D_{gen}) and the reduction of the dimension of the space of the dual variables compared to the one in [3] make it more suitable for calculations when dealing with concrete loss functions and consequently for numerical implementations. Let us denote by $v(P_{gen})$ the optimal objective value of the primal problem (P_{gen}) and by $v(D_{gen})$ the optimal objective value of its dual problem (D_{gen}). First of all, we show that for the minimization problem (P_{gen}) and its dual problem (D_{gen}) weak duality holds. The weak duality statement can be obtained as a particular instance of a more general result (taken, for instance, from [2]), nevertheless, we opt for providing it at this point for reader convenience.

Theorem 1. *For (P_{gen}) and (D_{gen}) weak duality holds, i. e. $v(P_{gen}) \geq v(D_{gen})$.*

Proof. Let be $c \in \mathbb{R}^n$ and $P = (P_1, \ldots, P_n)^T \in \mathbb{R}^n$. Then it holds, according to Young-

Fenchel inequality and due to the positive semidefiniteness of K, that

$$\begin{aligned}0 \leq C &\left[\sum_{i=1}^{n} v\big((Kc)_i, y_i\big) + \sum_{i=1}^{n} \big(v(\cdot, y_i)\big)^*\Big(-\frac{P_i}{C}\Big) + \sum_{i=1}^{n}(Kc)_i\frac{P_i}{C}\right] \\ &+ \frac{1}{2}(c-P)^T K(c-P) \\ = C&\sum_{i=1}^{n} v\big((Kc)_i, y_i\big) + C\sum_{i=1}^{n} \big(v(\cdot, y_i)\big)^*\Big(-\frac{P_i}{C}\Big) + P^T(Kc) \\ &+ \frac{1}{2}c^T Kc + \frac{1}{2}P^T KP - P^T(Kc) \\ = C&\sum_{i=1}^{n} v\big((Kc)_i, y_i\big) + \frac{1}{2}c^T Kc + C\sum_{i=1}^{n} \big(v(\cdot, y_i)\big)^*\Big(-\frac{P_i}{C}\Big) + \frac{1}{2}P^T KP\end{aligned}$$

and therefore

$$C\sum_{i=1}^{n} v\big((Kc)_i, y_i\big) + \frac{1}{2}c^T Kc \geq -C\sum_{i=1}^{n} \big(v(\cdot, y_i)\big)^*\Big(-\frac{P_i}{C}\Big) - \frac{1}{2}P^T KP,$$

i. e. $v(P_{gen}) \geq v(D_{gen})$. □

By introducing the functions $v_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, $v_i(z) = v(z_i, y_i)$, $i = 1, \ldots, n$, the problem (P_{gen}) can equivalently be written as

$$(P_{gen}) \qquad \inf_{c\in\mathbb{R}^n}\left\{C\sum_{i=1}^{n} v_i(Kc) + \frac{1}{2}c^T Kc\right\}. \tag{5}$$

In order to ensure strong duality for the primal-dual pair $(P_{gen}) - (D_{gen})$, we impose the following *qualification condition*

$$(QC) \qquad \operatorname{Im} K \cap \prod_{i=1}^{n} \operatorname{ri}(\operatorname{dom} v(\cdot, y_i)) \neq \emptyset.$$

Theorem 2. *If (QC) is fulfilled, then it holds $v(P_{gen}) = v(D_{gen})$ and (D_{gen}) has an optimal solution.*

Proof. We notice first that

$$v(P_{gen}) = \inf_{c\in\mathbb{R}^n}\left\{\left(\sum_{i=1}^{n} Cv_i\right)(Kc) + \frac{1}{2}c^T Kc\right\}.$$

Denoting by $g : \mathbb{R}^n \to \mathbb{R}$, $g(c) = \frac{1}{2}c^T Kc$, and by taking into consideration that $\operatorname{dom}\left(\sum_{i=1}^{n} Cv_i\right) = \prod_{i=1}^{n} \operatorname{dom} v(\cdot, y_i)$, one has

$$K(\operatorname{ri}(\operatorname{dom} g)) \cap \operatorname{ri}\left(\sum_{i=1}^{n} Cv_i\right) = \operatorname{Im} K \cap \prod_{i=1}^{n} \operatorname{ri}(\operatorname{dom} v(\cdot, y_i)) \neq \emptyset.$$

This means that $v(P_{gen}) < +\infty$. Moreover, there exists a $\bar{P} \in \mathbb{R}^n$ such that (see [2, Theorem 2.1])

$$\begin{aligned} v(P_{gen}) &= \sup_{P \in \mathbb{R}^n} \Big\{ -\Big(\sum_{i=1}^n C v_i\Big)^*(-P) - g^*(KP) \Big\} \\ &= -\Big(C\sum_{i=1}^n v_i\Big)^*(-\bar{P}) - g^*(K\bar{P}). \end{aligned}$$

Since for $q \in \mathbb{R}^n$,

$$g^*(q) = \begin{cases} \frac{1}{2} q^T K^- q, & \text{if } q \in \operatorname{Im} K, \\ +\infty, & \text{otherwise,} \end{cases}$$

where K^- is the Moore-Penrose pseudo-inverse of K, it holds

$$g^*(K\bar{P}) = \frac{1}{2}(K\bar{P})^T K^-(K\bar{P}) = \frac{1}{2}\bar{P}^T K K^- K\bar{P} = \frac{1}{2}\bar{P}^T K\bar{P}$$

and, so,

$$v(P_{gen}) = -C\left(\sum_{i=1}^n v_i\right)^*\left(-\frac{1}{C}\bar{P}\right) - \frac{1}{2}\bar{P}^T K\bar{P}.$$

As from (QC) one has $\cap_{i=1}^n \operatorname{ri}(\operatorname{dom} v_i) = \prod_{i=1}^n \operatorname{ri}(\operatorname{dom} v(\cdot, y_i)) \neq \emptyset$, it follows (cf. [9]) that there exist $\bar{P}^i \in \mathbb{R}^n$, $i = 1, ..., n$, with $\sum_{i=1}^n \bar{P}^i = \bar{P}$, such that

$$\left(\sum_{i=1}^n v_i\right)^*\left(-\frac{1}{C}\bar{P}\right) = \sum_{i=1}^n v_i^*\left(-\frac{1}{C}\bar{P}^i\right)$$

and, therefore,

$$v(P_{gen}) = -C\sum_{i=1}^n v_i^*\left(-\frac{1}{C}\bar{P}^i\right) - \frac{1}{2}\left(\sum_{i=1}^n \bar{P}^i\right)^T K\left(\sum_{i=1}^n \bar{P}^i\right).$$

Further, for all $i = 1, \ldots, n$, it holds

$$v_i^*\left(-\frac{1}{C}\bar{P}^i\right) = \sup_{z \in \mathbb{R}^n}\left\{-\frac{1}{C}(\bar{P}^i)^T z - v(z_i, y_i)\right\} = \begin{cases} \Big(v(\cdot, y_i)\Big)^*(-\frac{\bar{P}_i^i}{C}), & \text{if } \bar{P}_j^i = 0, \forall j \neq i, \\ +\infty, & \text{otherwise.} \end{cases}$$

Since the optimal objective value of (P_{gen}) is finite, by defining $\bar{P}_i := \bar{P}_i^i$ for $i = 1, ..., n$, one has $\sum_{i=1}^n \bar{P}^i = (\bar{P}_1, \ldots, \bar{P}_n)^T \in \mathbb{R}^n$ and

$$v(P_{gen}) = -C\sum_{i=1}^n \Big(v(\cdot, y_i)\Big)^*\left(-\frac{\bar{P}_i}{C}\right) - \frac{1}{2}\bar{P}^T K\bar{P},$$

where $\bar{P} := (\bar{P}_1, \ldots, \bar{P}_n)^T$. This, along with the weak duality theorem, provides the desired result, $\bar{P}$ being an optimal solution to (D_{gen}). □

The next theorem furnishes the necessary and sufficient optimality conditions for the primal-dual pair $(P_{gen}) - (D_{gen})$.

Theorem 3. *Let (QC) be fulfilled. Then $\bar{c} \in \mathbb{R}^n$ is an optimal solution for (P_{gen}) if and only if there exists an optimal solution $\bar{P} \in \mathbb{R}^n$ to (D_{gen}) such that*

(i) $-\frac{\bar{P}_i}{C} \in \partial v(\cdot, y_i)((K\bar{c})_i)$, $i = 1, \ldots, n$;

(ii) $K(\bar{c} - \bar{P}) = 0$.

Proof. From Theorem 2 we get the existence of an optimal solution $\bar{P} \in \mathbb{R}^n$ to (D_{gen}) such that

$$C\left[\sum_{i=1}^{n} v\big((K\bar{c})_i, y_i\big) + \sum_{i=1}^{n} \Big(v(\cdot, y_i)\Big)^* \left(-\frac{\bar{P}_i}{C}\right) + \sum_{i=1}^{n} (K\bar{c})_i \frac{\bar{P}_i}{C}\right] + \frac{1}{2}\bar{c}^T K\bar{c} + \frac{1}{2}\bar{P}^T K\bar{P} - \bar{P}^T K\bar{c} = 0.$$

This is equivalent to

$$\begin{cases} v((K\bar{c})_i, y_i) + \big(v(\cdot, y_i)\big)^*\big(\frac{\bar{P}_i}{C}\big) = (K\bar{c})_i \frac{\bar{P}_i}{C} \ \forall i = 1, \ldots, n, \\ \frac{1}{2}(\bar{c} - \bar{P})^T K(\bar{c} - \bar{P}) = 0. \end{cases}$$

Thus $\bar{c} - \bar{P}$ is a global minimum of the convex function $p \mapsto 1/2p^T Kp$, which means that the second statement in the relations above is nothing else than $K(\bar{c} - \bar{P}) = 0$. □

Remark 1. If K is positive definite, then, due to the fact that $v(\cdot, y_i)$ is proper and convex for all $i = 1, ..., n$, the qualification condition (QC) is automatically fulfilled. Thus, according to Theorem 3, $\bar{c} \in \mathbb{R}^n$ is an optimal solution for (P_{gen}) if and only if there exists an optimal solution $\bar{P} \in \mathbb{R}^n$ to (D_{gen}) such that

(i) $-\frac{\bar{P}_i}{C} \in \partial v(\cdot, y_i)((K\bar{c})_i)$, $i = 1, \ldots, n$;

(ii) $\bar{c} = \bar{P}$.

Remark 2. If K is positive definite, then the function g is strongly convex (on $\mathbb{R}^n$). Consequently, if $v(\cdot, y_i)$, $i = 1, \ldots, n$, is, additionally, lower semicontinuous, the optimization problem (P_{gen}) has a *unique optimal solution* (see, for instance, [5, Satz 6.33]). Further, due to the fact that $P \mapsto \frac{1}{2}P^T KP$ is strictly convex (on $\mathbb{R}^n$), one can see that the dual problem (D_{gen}) has at most one optimal solution. Consequently, due to Remark 1, whenever K is positive definite and $v(\cdot, y_i)$ is lower semicontinuous, for $i = 1, ..., n$, then in order to solve (P_{gen}) one can equivalently solve (D_{gen}) which in this case has an unique optimal solution $\bar{P}$, this being also the unique optimal solution of (P_{gen}).

3 Dual programs for different loss functions

In this section we consider different loss functions for performing the regression task. For each of the regularization problems to which these loss functions give rise we derive the corresponding dual problem. Notice also that all considered loss functions in this section are proper, convex and lower semicontinuous in their first arguments.

3.1 The $\varepsilon-$insensitive loss function

The well known $\varepsilon-$*insensitive loss function* $v_\varepsilon : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is, for $\varepsilon > 0$, defined as

$$v_\varepsilon(a, y) = (|a - y| - \varepsilon)_+ = \begin{cases} 0, & |a - y| \leq \varepsilon, \\ |a - y| - \varepsilon, & \text{else.} \end{cases} \tag{6}$$

Thus the primal optimization problem (P_{gen}) becomes

$$(P_\varepsilon) \qquad \inf_{c \in \mathbb{R}^n} \left\{ C \sum_{i=1}^{n} (|(Kc)_i - y_i| - \varepsilon)_+ + \frac{1}{2} c^T K c \right\}. \tag{7}$$

To obtain its dual problem (D_ε) via (4), we use the Lagrange technique in order to calculate the conjugate function of $v_\varepsilon(\cdot, y_i)$, for $i = 1, ..., n$. For $z \in \mathbb{R}$ and $y \in \mathbb{R}$ we have

$$\begin{aligned} -\left(v_\varepsilon(\cdot, y)\right)^*(z) &= -\sup_{a \in \mathbb{R}} \{za - (|a - y| - \varepsilon)_+\} = \inf_{a \in \mathbb{R}} \{-za + (|a - y| - \varepsilon)_+\} \\ &= \inf_{\substack{a \in \mathbb{R}, \\ t \geq 0, t \geq |a-y| - \varepsilon}} \{-za + t\} \\ &= \sup_{\lambda \geq 0, \beta \geq 0} \left\{ \inf_{a \in \mathbb{R}, t \in \mathbb{R}} \{-za + t + \lambda|a - y| - \lambda\varepsilon - \lambda t - \beta t\} \right\} \\ &= \sup_{\lambda \geq 0, \beta \geq 0} \left\{ \inf_{a \in \mathbb{R}} \{-za + \lambda|a - y|\} + \inf_{t \in \mathbb{R}} \{t - \lambda t - \beta t\} - \lambda\varepsilon \right\}. \end{aligned}$$

Since

$$\inf_{a \in \mathbb{R}} \{-za + \lambda|a - y|\} = \begin{cases} -zy, & \lambda \geq |z|, \\ -\infty, & \text{else} \end{cases} \text{ and } \inf_{t \in \mathbb{R}} \{t - \lambda t - \beta t\} = \begin{cases} 0, & \lambda + \beta = 1, \\ -\infty, & \text{else,} \end{cases}$$

we get

$$-\left(v_\varepsilon(\cdot, y)\right)^*(z) = \begin{cases} -zy - \varepsilon|z|, & |z| \leq 1, \\ -\infty, & \text{else} \end{cases}$$

and the dual problem (D_ε) to the primal problem (P_ε) results in

$$(D_\varepsilon) \qquad \sup_{\substack{P = (P_1, ..., P_n)^T \in \mathbb{R}^n, \\ |P_i| \leq C, i = 1, ..., n}} \left\{ \sum_{i=1}^{n} P_i y_i - \varepsilon \sum_{i=1}^{n} |P_i| - \frac{1}{2} P^T K P \right\}. \tag{8}$$

3.2 The quadratic $\varepsilon-$insensitive loss function

The second loss function we consider here is the so-called *quadratic* $\varepsilon-$*insensitive loss function* $v_{\varepsilon^2} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which is defined, for $\varepsilon > 0$, by

$$v_{\varepsilon^2}(a, y) = (|a - y| - \varepsilon)_+^2 = \begin{cases} 0, & |a - y| \leq \varepsilon, \\ (|a - y| - \varepsilon)^2, & \text{else.} \end{cases} \tag{9}$$

The corresponding primal problem reads in this case

$$(P_{\varepsilon^2}) \qquad \inf_{c\in\mathbb{R}^n}\left\{C\sum_{i=1}^{n}(|(Kc)_i - y_i| - \varepsilon)_+^2 + \frac{1}{2}c^TKc\right\}. \tag{10}$$

Again, in order to derive its dual problem (D_{ε^2}), we need to calculate, for $y, z \in \mathbb{R}$, the following conjugate function

$$\begin{aligned}
-\left(v_{\varepsilon^2}(\cdot, y)\right)^*(z) &= -\sup_{a\in\mathbb{R}}\{za - (|a-y|-\varepsilon)_+^2\} = \inf_{\substack{a\in\mathbb{R},\\ t\geq 0,\, t\geq |a-y|-\varepsilon}}\{-za + t^2\} \\
&= \sup_{\lambda\geq 0,\,\beta\geq 0}\left\{\inf_{a\in\mathbb{R},\,t\in\mathbb{R}}\left\{-za + t^2 + \lambda(|a-y| - \varepsilon - t) - \beta t\right\}\right\} \\
&= \sup_{\lambda\geq 0,\,\beta\geq 0}\left\{\inf_{a\in\mathbb{R}}\{-za + \lambda|a-y|\} + \inf_{t\in\mathbb{R}}\left\{t^2 - \lambda t - \beta t\right\} - \lambda\varepsilon\right\}
\end{aligned}$$

The first inner infimum has been already calculated in the previous subsection, while for the second one we have

$$\inf_{t\in\mathbb{R}}\left\{t^2 - (\lambda+\beta)t\right\} = -\frac{1}{4}(\lambda+\beta)^2.$$

Hence, one the above conjugate becomes

$$-\left(v_{\varepsilon^2}(\cdot, y)\right)^*(z) = -zy - \frac{1}{4}z^2 - \varepsilon|z|$$

and gives rise to the following dual problem

$$(D_{\varepsilon^2}) \qquad \sup_{P=(P_1,\ldots,P_n)^T\in\mathbb{R}^n}\left\{\sum_{i=1}^{n}P_iy_i - \frac{1}{4C}\sum_{i=1}^{n}P_i^2 - \varepsilon\sum_{i=1}^{n}|P_i| - \frac{1}{2}P^TKP\right\}. \tag{11}$$

3.3 The Huber loss function

Another popular choice for the loss function in SVM regression tasks is the *Huber loss function* $v_{\mathrm{H}} : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ which is defined, for $\varepsilon > 0$, as

$$v_{\mathrm{H}}(a, y) = \begin{cases}\varepsilon|a-y| - \frac{\varepsilon^2}{2}, & |a-y| > \varepsilon,\\ \frac{1}{2}|a-y|^2, & |a-y|\leq\varepsilon.\end{cases} \tag{12}$$

The primal problem associated with the Huber loss function therefore becomes

$$(P_{\mathrm{H}}) \qquad \inf_{c\in\mathbb{R}^n}\left\{C\sum_{i=1}^{n}v_H((Kc)_i, y_i) + \frac{1}{2}c^TKc\right\}. \tag{13}$$

For all $z, y \in \mathbb{R}$ one has

$$\begin{aligned}
-\left(v_{\mathrm{H}}(\cdot, y)\right)^*(z) &= -\sup_{a\in\mathbb{R}}\left\{za - v_{\mathrm{H}}(a,y)\right\} = \inf_{a\in\mathbb{R}}\left\{-za + v_{\mathrm{H}}(a,y)\right\} \\
&= \min\left\{\inf_{\substack{a\in\mathbb{R},\\ |a-y|\le\varepsilon}}\left\{-za + \frac{1}{2}|a-y|^2\right\}, \inf_{\substack{a\in\mathbb{R},\\ |a-y|>\varepsilon}}\left\{-za + \varepsilon|a-y| - \frac{\varepsilon^2}{2}\right\}\right\} \\
&= \min\left\{\inf_{\substack{a\in\mathbb{R},\\ |a-y|\le\varepsilon}}\left\{-za + \frac{1}{2}(a-y)^2\right\}, \inf_{\substack{a\in\mathbb{R},\\ a>y+\varepsilon}}\left\{-za + \varepsilon(a-y) - \frac{\varepsilon^2}{2}\right\},\right. \\
&\qquad \left.\inf_{\substack{a\in\mathbb{R},\\ a<y-\varepsilon}}\left\{-za + \varepsilon(y-a) - \frac{\varepsilon^2}{2}\right\}\right\}.
\end{aligned}$$

For the first infimum we get

$$\inf_{\substack{a\in\mathbb{R},\\ |a-y|\le\varepsilon}}\left\{-za + \frac{1}{2}(a-y)^2\right\} = \begin{cases} \frac{\varepsilon^2}{2} - zy + z\varepsilon - \frac{y^2}{2}, & z < -\varepsilon, \\ -\frac{1}{2}z^2 - zy - \frac{y^2}{2}, & z \in [-\varepsilon, \varepsilon], \\ \frac{\varepsilon^2}{2} - zy - z\varepsilon - \frac{y^2}{2}, & z > \varepsilon, \end{cases} \tag{14}$$

while the second and third infima result in

$$\inf_{\substack{a\in\mathbb{R},\\ a>y+\varepsilon}}\left\{-za + \varepsilon(a-y) - \frac{\varepsilon^2}{2}\right\} = \begin{cases} \frac{\varepsilon^2}{2} - zy - z\varepsilon, & z \le \varepsilon, \\ -\infty, & \text{else} \end{cases} \tag{15}$$

and

$$\inf_{\substack{a\in\mathbb{R},\\ a<y-\varepsilon}}\left\{-za + \varepsilon(y-a) - \frac{\varepsilon^2}{2}\right\} = \begin{cases} \frac{1}{2}\varepsilon^2 - zy + z\varepsilon, & z \ge -\varepsilon, \\ -\infty, & \text{else,} \end{cases} \tag{16}$$

respectively. Putting (14), (15) and (16) together we obtain the following formula for the conjugate function

$$\begin{aligned}
-\left(v_{\mathrm{H}}(\cdot, y)\right)^*(z) &= \begin{cases} \min\left\{-\frac{1}{2}z^2 - zy, \frac{\varepsilon^2}{2} - zy - z\varepsilon, \frac{\varepsilon^2}{2} - zy + \varepsilon z\right\}, & z \in [-\varepsilon, \varepsilon], \\ -\infty, & \text{else} \end{cases} \\
&= \begin{cases} -\frac{1}{2}z^2 - zy, & z \in [-\varepsilon, \varepsilon], \\ -\infty, & \text{else.} \end{cases}
\end{aligned}$$

Thus, the dual problem to (P_H) reads

$$(D_{\mathrm{H}}) \qquad \sup_{\substack{P=(P_1,\ldots,P_n)^T\in\mathbb{R}^n,\\ |P_i|\le\varepsilon C, i=1,\ldots,n}}\left\{\sum_{i=1}^n P_i y_i - \frac{1}{2C}\sum_{i=1}^n P_i^2 - \frac{1}{2}P^T K P\right\}$$

3.4 The extended loss function

Finally, we provide the resulting dual problem when using the *extended loss function* $v_{\text{ext}} : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}}$, which is defined, for $\varepsilon > 0$, as

$$v_{\text{ext}}(a, y) = \delta_{[-\varepsilon,\varepsilon]} = \begin{cases} 0, & |a - y| \leq \varepsilon, \\ +\infty, & \text{else} \end{cases} \tag{17}$$

This choice gives rise to the following primal problem

$$(P_{\text{ext}}) \qquad \inf_{\substack{c \in \mathbb{R}^n, \\ |(Kc)_i - y_i| \leq \varepsilon, i=1,\ldots,n}} \frac{1}{2} c^T K c. \tag{18}$$

By making again use of Lagrange duality, we get for all $y, z \in \mathbb{R}$

$$\begin{aligned} -\left(v_{\text{ext}}(\cdot, y)\right)^*(z) &= -\sup_{\substack{a \in \mathbb{R}, \\ |a-y| \leq \varepsilon}} \{za\} = \inf_{\substack{a \in \mathbb{R}, \\ |a-y| \leq \varepsilon}} \{-za\} = \inf_{\substack{a \in \mathbb{R}, \\ a-y-\varepsilon \leq 0, \\ y-a-\varepsilon \leq 0}} \{-za\} \\ &= \sup_{\lambda \geq 0, \beta \geq 0} \left\{ \inf_{a \in \mathbb{R}} \{(-z + \lambda - \beta)a - \lambda y - \lambda\varepsilon + \beta y - \beta\varepsilon\} \right\} \\ &= \sup_{\substack{\lambda \geq 0, \beta \geq 0, \\ \lambda - \beta = z}} \{-\lambda y - \lambda\varepsilon + \beta y - \beta\varepsilon\} = \sup_{\substack{\lambda \geq 0, \beta \geq 0, \\ \lambda - \beta = z}} \{-(\lambda - \beta)y - (\lambda + \beta)\varepsilon\} \\ &= -zy + \sup_{\substack{\lambda \geq 0, \beta \geq 0, \\ \lambda - \beta = z}} \{-\varepsilon(\lambda + \beta)\} = -zy - \varepsilon|z|. \end{aligned}$$

Consequently, the dual problem to (P_{ext}) has the following formulation

$$(D_{\text{ext}}) \qquad \sup_{P=(P_1,\ldots,P_n)^T \in \mathbb{R}^n} \left\{ \sum_{i=1}^n P_i y_i - \varepsilon \sum_{i=1}^n |P_i| - \frac{1}{2} P^T K P \right\}$$

4 Application

In this section we discuss two particular regression tasks in the light of the approach introduced in the previous sections and solve to this end the different dual optimization problems (D_ε), (D_{ε^2}), (D_{H}) and (D_{ext}) numerically. In a first step, we reformulate these optimization problems in order to get a representation of them that is suitable for standard optimization routines and therefore more easy to handle with. Having in mind the dual problem (D_ε), we note that for $z \in \mathbb{R}$ it holds

$$|z| = \inf_{\substack{\alpha \geq 0, \alpha^* \geq 0, \\ \alpha - \alpha^* = z}} \{\alpha + \alpha^*\} \tag{19}$$

for arbitrary $z \in \mathbb{R}$. If $z \geq 0$, then the optimal solution of this minimization problem is $(\alpha, \alpha^*) = (z, 0)$, while, when $z < 0$, the optimal solution is $(\alpha, \alpha^*) = (0, -z)$. This remark constitutes the starting point for giving an equivalent formulation of the dual

problem (D_ε) in terms of the variables α_i and α_i^*, $i = 1, \ldots, n$, which we will denote by (D_ε^α). For the problem

$$(D_\varepsilon) \qquad \sup_{\substack{P=(P_1,\ldots,P_n)^T\in\mathbb{R}^n,\\ |P_i|\leq C, i=1,\ldots,n}} \left\{ \sum_{i=1}^n P_i y_i - \varepsilon \sum_{i=1}^n |P_i| - \frac{1}{2} P^T K P \right\}$$

the equivalent formulation (D_ε^α) is

$$(D_\varepsilon^\alpha) \qquad \inf_{\substack{\alpha_i,\alpha_i^*\in[0,C],\\ i=1,\ldots,n}} \left\{ \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K_{ij} + \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) - \sum_{i=1}^n (\alpha_i - \alpha_i^*) y_i \right\}.$$

Using again (19) an equivalent formulation for the problem

$$(D_{\text{ext}}) \qquad \sup_{P=(P_1,\ldots,P_n)^T\in\mathbb{R}^n} \left\{ \sum_{i=1}^n P_i y_i - \varepsilon \sum_{i=1}^n |P_i| - \frac{1}{2} P^T K P \right\},$$

to which the use of extended loss gives rise, in terms of α_i and α_i^*, $i = 1, \ldots, n$, is

$$(D_{\text{ext}}^\alpha) \qquad \inf_{\substack{\alpha_i,\alpha_i^*\geq 0,\\ i=1,\ldots,n}} \left\{ \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K_{ij} + \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) - \sum_{i=1}^n (\alpha_i - \alpha_i^*) y_i \right\}$$

In order to obtain an equivalent formulation $(D_{\varepsilon^2}^\alpha)$ of the optimization problem (D_{ε^2}) we make use of the fact that

$$|z| = \inf_{\substack{\alpha,\alpha^*\geq 0,\\ \alpha-\alpha^*=z}} \left\{ \alpha + \alpha^* + \frac{\alpha\alpha^*}{2C\varepsilon} \right\}$$

for arbitrary $z \in \mathbb{R}$. Then the representation of

$$(D_{\varepsilon^2}) \qquad \sup_{P=(P_1,\ldots,P_n)^T\in\mathbb{R}^n} \left\{ \sum_{i=1}^n P_i y_i - \frac{1}{4C} \sum_{i=1}^n P_i^2 - \varepsilon \sum_{i=1}^n |P_i| - \frac{1}{2} P^T K P \right\}$$

is

$$(D_{\varepsilon^2}^\alpha) \qquad \inf_{\alpha_i,\,\alpha_i^*\geq 0} \left\{ \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K_{ij} + \frac{1}{4C} \sum_{i=1}^n (\alpha_i^2 + (\alpha_i^*)^2) \right.$$
$$\left. + \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) - \sum_{i=1}^n (\alpha_i - \alpha_i^*) y_i \right\}.$$

Finally, for arbitrary $z \in \mathbb{R}$ it holds

$$z^2 = \inf_{\substack{\alpha,\alpha^*\geq 0,\\ \alpha-\alpha^*=z}} \{\alpha^2 + (\alpha^*)^2\}$$

and therefore, an equivalent formulation of

$$(D_{\mathrm{H}}) \qquad \sup_{\substack{P=(P_1,\ldots,P_n)^T\in\mathbb{R}^n,\\ |P_i|\leq\varepsilon C, i=1,\ldots,n}} \left\{\sum_{i=1}^n P_i y_i - \frac{1}{2C}\sum_{i=1}^n P_i^2 - \frac{1}{2}P^T K P\right\}$$

in terms of α_i and α_i^*, $i=1,\ldots,n$, is

$$(D_{\mathrm{H}}^{\alpha}) \inf_{\substack{\alpha_i,\alpha_i^*\in[0,\varepsilon C],\\ i=1,\ldots,n}} \left\{\frac{1}{2}\sum_{i,j=1}^n(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K_{ij} + \frac{1}{2C}\sum_{i=1}^n\left(\alpha_i^2-(\alpha_i^*)^2\right) - \sum_{i=1}^n(\alpha_i-\alpha_i^*)y_i\right\}.$$

Remark 3. While the corresponding primal problems are either unconstrained nondifferentiable convex optimization problems or reformulations of constrained optimization problems with differentiable objective functions and not easily handleable inequality constraints, the duals $(D_{\varepsilon}^{\alpha})$, $(D_{\varepsilon^2}^{\alpha})$, $(D_{\mathrm{ext}}^{\alpha})$ and $(D_{\mathrm{H}}^{\alpha})$ assume the minimization of a convex quadratic objective function over some feasible sets expressed via box constraints or nonnegative orthants. This makes them easier solvable via some standard algorithms designed for these classes of optimization problems than their corresponding primal problems. Moreover, if $(\bar{\alpha}_1,\bar{\alpha}_1^*,...,\bar{\alpha}_n,\bar{\alpha}_n^*)$ represents an optimal solution of each of the reformulated dual problems, then $\bar{P}:=(\bar{P}_1,...,\bar{P}_n)^T, \bar{P}_i=\bar{\alpha}_i-\bar{\alpha}_i^*$, $i=1,\ldots,n$, represents an optimal solution of the corresponding initial dual.

The two particular regression tasks which we consider in this section involve a toy data set (cf. 4.1) and the popular Boston Housing data set (cf. 4.2). In both situations we use the Gaussian RBF kernel

$$k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{20}$$

with kernel parameter $\sigma>0$. This gives rise to a positive definite Gram matrix K and, therefore, according to Remark 2, an optimal solution $\bar{P}:=(\bar{P}_1,...,\bar{P}_n)^T$ of the dual will be an optimal solution of the primal, too. Thus, the components of this vector will provide the decision function one looks for when considering the regression task.

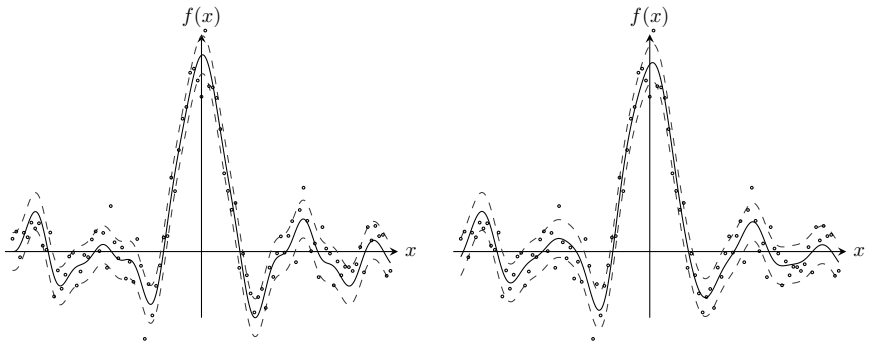
4.1 A toy data set

In this subsection we numerically solve a regression task where the data has been sampled from the function $f:\mathbb{R}\to\mathbb{R}$,

$$f(x)=\begin{cases}\frac{\sin(x)}{x}, & x\neq 0,\\ 1, & x=0.\end{cases}$$

The function values for all $x\in X=\{-5.0,-4.9,\ldots,4.9,5.0\}$ resulting in a total of 101 points were sampled. The values $f(x)$, $x\in X$, were perturbed by adding a random value drawn from the normal distribution $\mathcal{N}(0,0.1)$. In this way a training set $\mathcal{D}=\{(x_i,y_i):i=1,\ldots,101\}$ was obtained and used for training. On the basis of this set we solved the dual problems $(D_{\varepsilon}^{\alpha})$, $(D_{\varepsilon^2}^{\alpha})$, $(D_{\mathrm{H}}^{\alpha})$ and $(D_{\mathrm{ext}}^{\alpha})$ numerically, while Figure 4.1

shows the shapes of the resulting regression functions when choosing the corresponding loss function.



(a) $\varepsilon-$insensitive loss (b) quadratic $\varepsilon-$insensitive loss

(c) Huber loss (d) extended loss

Figure 4.1: Illustrations of the four resulting regression functions (solid lines) for the corresponding loss function and the $\varepsilon-$tube (dashed lines, where appropriate) based on the generated training set (dots). (a) $C = 100$, $\sigma = 0.5$, $\varepsilon = 0.1$ (b) $C = 100$, $\sigma = 0.5$, $\varepsilon = 0.1$ (c) $C = 100$, $\sigma = 0.5$, $\varepsilon = 0.1$ (d) $\sigma = 0.2$, $\varepsilon = 0.1$

Table 4.1 shows the corresponding mean squared errors. With respect to this special setting the use of the $\varepsilon-$insensitive loss function and of the quadratic $\varepsilon-$insensitive loss function produce similar mean squared errors, while the use of the extended loss function provides the lowest mean squared error, as expected.

loss function	$\varepsilon-$insensitive	ε^2-insensitive	Huber	extended
mean squared error	0.008192	0.008193	0.007566	0.006188

Table 4.1: The mean squared error for the four different loss functions obtained by applying the parameter settings described in the caption of Figure 4.1.

4.2 Boston Housing data set

In this section we solve the dual problems $(D_{\varepsilon}^{\alpha})$, $(D_{\varepsilon^2}^{\alpha})$, $(D_{\mathrm{H}}^{\alpha})$ and $(D_{\mathrm{ext}}^{\alpha})$ for the the well known Boston Housing data set. This data set consists of 506 instances each of them

described by 13 attributes.

		ε		
C	σ	0.01	0.15	1.0
10	0.1	39.60	41.69	61.78
	0.5	10.51	9.21	34.37
	1.0	11.83	11.55	26.50
100	0.1	39.60	41.69	61.78
	0.5	12.58	10.54	34.13
	1.0	10.37	9.46	26.93
1000	0.1	39.60	41.69	61.78
	0.5	26.48	14.66	34.13
	1.0	15.45	10.16	26.93

(a) $\varepsilon-$insensitive loss

		ε		
C	σ	0.01	0.15	1.0
10	0.1	40.14	42.39	62.42
	0.5	8.63	9.30	36.52
	1.0	9.79	10.50	31.92
100	0.1	39.64	41.76	61.85
	0.5	10.37	9.77	34.39
	1.0	8.33	8.85	27.51
1000	0.1	39.60	41.69	61.79
	0.5	17.03	11.96	34.16
	1.0	10.49	9.85	26.99

(b) quadratic $\varepsilon-$insensitive loss

		ε		
C	σ	0.01	0.15	1.0
10	0.1	72.96	43.66	40.03
	0.5	33.79	13.72	8.95
	1.0	34.19	15.85	10.89
100	0.1	47.27	39.55	39.55
	0.5	15.61	10.18	10.02
	1.0	17.32	10.62	8.67
1000	0.1	39.52	39.52	39.52
	0.5	10.61	13.38	17.56
	1.0	11.89	10.24	10.13

(c) Huber loss

	ε			
σ	0.01	0.1	0.15	0.25
0.1	39.60	40.84	41.69	43.60
0.2	17.59	16.81	17.15	18.65
0.3	19.15	14.00	13.07	13.14
0.5	48.88	22.47	17.38	11.64
1.0	151.33	79.68	48.66	20.22
2.0	530.87	254.01	147.94	39.83

(d) extended loss

Figure 4.2: Four tables representing the average mean squared error over ten test folds for the resulting regression functions w. r. t. the corresponding loss functions and different parameter combinations.

For a detailed description of the data set we refer to [18]. In order to determine good parameter choices for the kernel parameter σ, the regularization parameter C and the loss function parameter ε, we performed a 10-fold cross validation. In tables 2(a), 2(b), 2(c) and 2(d) the mean test errors over 10 folds for all four loss functions are shown for a part of the whole tested parameter values, where we choose the mean squared error for evaluation. As in [13], we scaled the data before solving the problems numerically. As one can notice, the best result, i. e. the lowest mean squared error over 10 test folds, is obtained for the quadratic $\varepsilon-$insensitive loss function followed by the $\varepsilon-$insensitive loss function and the Huber loss function.

5 Conclusions

In this paper we solved the Support Vector Regression problem by making use of convex analysis specific techniques. The dual problems, to which the use of different loss functions for regression in the primal gave rise, were determined and numerically solved on the basis of two data sets. One can notice that, when considering the Boston Housing data set, the quadratic $\varepsilon-$insensitive loss and the Huber loss perform a slightly better regression, at least for our allowed parameter choices, than the standard $\varepsilon-$insensitive loss.

Acknowledgements

The authors are thankful to two anonymous reviewers for remarks which improved the quality of the paper. The research of R.I. Boţ was partially supported by DFG (German Research Foundation), projects BO 2516/4-1 and WA 922/1-3.
The research of A. Heinrich was supported by the European Union, the European Social Fund (ESF) and prudsys AG in Chemnitz.

References

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

[2] R.I. Boţ. *Conjugate Duality in Convex Optimization.* Lecture Notes in Economics and Mathematical Systems, Vol. 637, Springer-Verlag, Berlin Heidelberg, 2010.

[3] R.I. Boţ and N. Lorenz. *Optimization problems in statistical learning: Duality and optimality conditions.* European Journal of Operational Research, 213(2):395–404, 2011.

[4] O. Chapelle, V.N. Vapnik and J. Weston. *Transductive Inference for Estimating Values of Functions.* 1999.

[5] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben.* Springer-Verlag, Berlin Heidelberg New York, 2002.

[6] T. Joachims. *Learning to Classify Text using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers, Boston Dordrecht London, 2002.

[7] K. Kim. *Financial time series forecasting using support vector machines.* Neurocomputing 55(1-2):307–319, 2003.

[8] W.S. Noble. *Support vector machine application in computational biology.* In: B. Schölkopf, K. Tsuda, J.-P. Vert (Eds.), Kernel Methods in Computational Biology, MIT Press, Cambridge, pp. 71-92, 2004.

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

[10] C. Saunders, A. Gammerman and V. Vovk. *Ridge Regression Learning Algorithm in Dual Variables.* In. J.W. Shavlik (Ed.), Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, pp. 515–521, 1998.

[11] B. Schölkopf, P. Bartlett, A. Smola and R. Williamson. *Shrinking the Tube: A New Support Vector Regression Algorithm.* In: M.S. Kearns, S.A. Solla, D.A. Cohn (Eds.), Advances in Neural Information Processing Systems 11, MIT Press, Cambridge, pp. 330–336, 1999.

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

[13] M.O. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, J. Weston and Surrey Tw Ex. *Support Vector Regression with ANOVA Decompsition Kernels.* In: B. Schölkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, pp. 285–292, 1997.

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

[15] T. Van Gestel, B. Baesens, J. Garcia and P. Van Dijcke. *A support vector machine approach to credit scoring.* Bank en Financiewezen, 2:73–82, 2003.

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

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

[18] Zhong Yi, Zhou Chunguang, Huang Lan, Wang Yan and Yang Bin. *Support Vector Regression for Prediction of Housing Values.* Proceedings of the International Conference on Computational Intelligence and Security CIS'09, pp. 61–65, 2009.