Mach Learn (2009) 74: 315–345
DOI 10.1007/s10994-008-5095-1

An algebraic characterization of the optimum of regularized kernel methods

Francesco Dinuzzo · Giuseppe De Nicolao

Received: 4 June 2007 / Revised: 6 November 2008 / Accepted: 24 November 2008 / Published online: 10 January 2009

Abstract The representer theorem for kernel methods states that the solution of the associated variational problem can be expressed as the linear combination of a finite number of kernel functions. However, for non-smooth loss functions, the analytic characterization of the coefficients poses nontrivial problems. Standard approaches resort to constrained optimization reformulations which, in general, lack a closed-form solution. Herein, by a proper change of variable, it is shown that, for any convex loss function, the coefficients satisfy a system of algebraic equations in a fixed-point form, which may be directly obtained from the primal formulation. The algebraic characterization is specialized to regression and classification methods and the fixed-point equations are explicitly characterized for many loss functions of practical interest. The consequences of the main result are then investigated along two directions. First, the existence of an unconstrained smooth reformulation of the original non-smooth problem is proven. Second, in the context of SURE (Stein's Unbiased Risk Estimation), a general formula for the degrees of freedom of kernel regression methods is derived.

Keywords Representer theorem · Regularization theory · Kernel methods · Support vector machines · Degrees of freedom

1 Introduction

Kernel methods are widely used to solve classification and regression problems (Schölkopf and Smola 2001). The core of such methods is the minimization of a cost functional consist-

Editor: Olivier Chapelle.

F. Dinuzzo (✉)
Department of Mathematics, University of Pavia, Via Ferrata, 1, 27100 Pavia, Italy
e-mail: francesco.dinuzzo@gmail.com

G. De Nicolao
Department of Computer Engineering and Systems Science, University of Pavia, Via Ferrata, 1, 27100 Pavia, Italy
e-mail: giuseppe.denicolao@unipv.it

ing of the sum of a loss term dependent on the experimental data and a penalty term given by an RKHS (Reproducing Kernel Hilbert Space) norm. The choice of the loss function enables one to obtain a variety of estimators, including smoothing splines (Wahba 1990), regularization networks (Poggio and Girosi 1992) and SVM (Support Vector Machines) for both classification and regression (Vapnik 1995).

The remarkable feature of these methods is that, even if the minimizer is searched within an infinite dimensional space, the solution can be written as the linear combination of a finite number of kernel functions. This fundamental property goes under the name of representer theorem. The result, which goes back to (Kimeldorf and Wahba 1971) for squared loss functions, extends to differentiable loss functions, see (Cox and O'Sullivan 1990) and (Poggio and Girosi 1992). More recently, it has been shown that the representer theorem holds in a very general setting (Schölkopf et al. 2001). For what concerns the characterization of the coefficients of the linear combination, if the loss function is differentiable, they are the solution of a system of algebraic equations (Wahba 1998). This result does not apply to SVM because the loss function (either the ϵ-insensitive or the "hinge") is not differentiable. It is worth noticing that the standard approach to SVM computation relies on quadratic programming and lacks a closed-form solution. When the loss function is non-smooth, resorting to sub-differential calculus, it has been shown that the coefficients can be characterized in terms of inclusions (Steinwart 2003; De Vito et al. 2004). For regression loss functions, including ϵ-insensitive SVR (Support Vector Regression), it has been recently proven that the inclusions can be converted into a set of algebraic equations by a proper change of variable (Dinuzzo et al. 2007). Such algebraic characterization was then used to evaluate the degrees of freedom of ϵ-insensitive SVR and develop a tuning procedure for the parameters ϵ and C.

The present paper provides three main results. First of all, we derive an algebraic characterization of the optimum for a general class of kernel methods including classification ones and allowing for the presence of the bias term. The algebraic characterization provides new insight into the structure of the solution of the variational problem associated with kernel methods. In particular, the solution coefficients are directly characterized by a system of algebraic equations, a major difference with respect to dual-problem approaches, involving both equalities and inequalities. A number of specific examples are considered, including Huber's kernel regression, the ϵ-insensitive SVM, the support vector classification with hinge loss, and some methods relying on smooth approximations of the previous non-differentiable loss functions. Although investigation of potential computational benefits is beyond the scopes of the present paper, our algebraic characterization has a fixed-point structure and, in principle, it may also be used for computational purposes. Recently, there has been a certain interest in training kernel methods in the primal (Perez-Cruz et al. 2005; Chapelle 2007; Shalev-Shwartz et al. 2007). In particular, connections and differences between primal and dual formulations are thoroughly analyzed in Chapelle (2007).

The algebraic characterization is then exploited to derive the second main result of the paper, namely that the original non-smooth variational problem can be exactly reformulated as an unconstrained smooth optimization problem. Recently, "reformulation" has become an important research topic in optimization, see the book by Fukushima and Qi (1999). Smooth minimization of non-smooth convex functions has been considered in Nesterov (2005) (see also references therein), where a smoothing method is proposed that guarantees better bounds on the computational complexity of a rather general class of convex problems. A smooth approximation is employed in Lee and Mangasarian (2001) to generate the so-called SSVM (Smooth Support Vector Machines), an unconstrained smooth reformulation of the support vector machine for pattern classification.

For what concerns the third main contribution of the paper, by making reference to the system of algebraic equations, it is relatively easy to obtain the derivatives of the coefficients with respect to the data and parameters. This sensitivity analysis can be used for many purposes. Among them, there is the evaluation of the degrees of freedom of kernel regression methods characterized by quadratic regularization terms. Sensitivity analysis and degrees of freedom of nonlinear estimators are interesting topics that have been the subject of several investigations, see e.g. Hastie et al. (2001) and references therein. Insights in sensitivity problems of SVMs and lasso can be found, for instance, in Pontil and Verri (1998) and Osborne (2001). A major motivation for studying degrees of freedom is that they enter the definition of many popular tuning criteria such as C_p (Mallows 1973), *AIC* (Akaike 1973), *BIC* (Schwarz 1978), *GCV* (Craven and Wahba 1979), and others. General references on the topic of degrees of freedom are Stein (1981), Ye (1998), Efron (2004). More recently, Gunter and Zhu (2007) and Dinuzzo et al. (2007) studied the degrees of freedom of support vector regression. The degrees of freedom of the lasso has been extensively studied in Zou et al. (2007), where an up-to-date list of references on the subject can be found. It is worth remarking that assessing the degrees of freedom for the lasso and kernel methods poses intrinsically different problems in that the regularization term in the lasso is not quadratic.

The paper is organized as follows. The general form of the new algebraic characterization is derived in Sect. 2. The unconstrained smooth reformulation of the original non-smooth variational problem is derived in Sect. 3. The specialization to regularized regression and classification is treated in Sect. 4. In Sect. 5, the application of the new result to a variety of kernel methods is discussed. Section 6 addresses the evaluation of the degrees of freedom for a generic kernel regression estimator. Finally, some conclusions end the paper.

Throughout the paper, vectors are denoted by boldface characters (e.g. $\mathbf{a}$), while matrices are denoted by sans serif uppercase letters (e.g. K).

2 Algebraic characterization for regularized kernel methods

Given an input set X and an output set Y, we consider the problem of estimating a functional relationship between inputs $x \in X$ and outputs $y \in Y$, given the set of ℓ training pairs

$$D := \{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}.$$

Let $\hat{g} : X \to Y$ denote the relationship that is learnt from the data. In order to cast the learning problem within a regularization approach, one may assume that $\hat{g}$ is the composition

$$\hat{g} = \sigma \circ \hat{f}$$

of a given output function $\sigma : \mathbb{R} \to Y$ and a real-valued function $\hat{f} : X \to \mathbb{R}$, estimated by solving

$$\hat{f} = \arg\min_{f \in \mathcal{S}} \left(C \sum_{i=1}^{\ell} V(y_i, f(x_i)) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right), \tag{1}$$

where V is a suitable measurable function, convex with respect to its second argument, $\mathcal{S}$ either coincides with $\mathcal{H}$, where $\mathcal{H}$ is an RKHS with kernel $K(\cdot,\cdot)$, or is given by the sum $\mathcal{S} = \mathcal{H} + \mathcal{B}$, with $\mathcal{B}$ a finite-dimensional parametric bias space spanned by functions $\{\psi_j\}_{j=1}^m$. The norm in $\mathcal{H}$ is denoted by $\|\cdot\|_{\mathcal{H}}$ and C is a positive real number that selects the

regularization intensity. Note that for regression, $Y = \mathbb{R}$ and $\sigma(t) = t$, whereas for binary classification, a possible choice is $Y = \{-1, 1\}$ and $\sigma(t) = \mathrm{sgn}(t)$.

By the representer theorem, the solution $\hat{f}$ can be expressed as a finite linear combination of kernel functions, namely, for $\mathcal{S} = \mathcal{H}$,

$$\hat{f}(x) = \sum_{i=1}^{\ell} a_i K(x_i, x),$$

or, in the bias case $\mathcal{S} = \mathcal{H} + \mathcal{B}$,

$$\hat{f}(x) = \sum_{i=1}^{\ell} a_i K(x_i, x) + \sum_{j=1}^{m} b_j \psi_j(x).$$

It has been recently shown that the coefficients a_i are optimal if and only if they satisfy the set of inclusions (Steinwart 2003; De Vito et al. 2004):

$$a_i \in -C\partial_2 V\big(y_i, \hat{f}(x_i)\big), \tag{2}$$

with the additional relationships

$$\sum_{i=1}^{\ell} \psi_j(x_i) a_i = 0, \quad j = 1, \ldots, m, \tag{3}$$

in the bias case. In (2), the symbol $\partial_2 V$ denotes the sub-differential of the loss function V with respect to the second argument. Steinwart (2003) developed his quantified representer theorem as a by-product of the analysis of convergence and stability of SVMs. Herein, we proceed further in that the inclusion (2) will be replaced by a system of algebraic equations. Note that the function $V(y_i, \cdot)$ is a convex real function of a real variable, so that the structure of the sub-differential $\partial_2 V$ is very simple: essentially, it contains either a single value or an interval, depending on the value of the second argument $\hat{f}(x_i)$. In the following, $D^-(\gamma)$ and $D^+(\gamma)$ will denote the left and right derivative of $V(y_i, \cdot)$ at γ:

$$D^-(\gamma) := \lim_{h \to 0^-} \frac{V(y_i, \gamma + h) - V(y_i, \gamma)}{h}, \qquad D^+(\gamma) := \lim_{h \to 0^+} \frac{V(y_i, \gamma + h) - V(y_i, \gamma)}{h}.$$

In addition, let

$$\mathbf{a} := (a_1, \ldots, a_\ell)^T, \qquad \mathbf{b} := (b_1, \ldots, b_m)^T, \qquad \mathbf{y} := (y_1, \ldots, y_\ell)^T.$$

Notations like $\hat{f}(x_i; \mathbf{a}, \mathbf{b})$ will be used to denote dependence on coefficient vectors $\mathbf{a}$ and $\mathbf{b}$. The next theorem shows that the set of inclusions (2) can be replaced by a set of algebraic equations. The proof is constructive and the key idea is to introduce some new coefficients z_i.

Theorem 1 *For any $\delta > 0$, let*

$$z_i(\mathbf{a}, \mathbf{b}) := a_i K(x_i, x_i)\delta - \hat{f}(x_i; \mathbf{a}, \mathbf{b}).$$

Let $V(y_i, \cdot)$ denote a convex loss function that is twice differentiable everywhere except in a finite number of points $\gamma_k(y_i)$ $(k = 1, \ldots, n_\gamma^i)$. Then, coefficients a_i $(i = 1, \ldots, \ell)$, characterizing the solution of problem (1), are well defined by a system of algebraic equations of

the type

$$a_i = S_i(y_i, z_i(\mathbf{a}, \mathbf{b})), \tag{4}$$

where $S_i(y_i, \cdot)$ *are monotonic non-decreasing Lipschitz continuous functions. More precisely, when* $z_i \in I_k$, $k = 1, \ldots, n_\gamma^i$,

$$I_k := \left[-\big(\gamma_k(y_i) + CK(x_i, x_i)D^+(\gamma_k(y_i))\delta\big), -\big(\gamma_k(y_i) + CK(x_i, x_i)D^-(\gamma_k(y_i))\delta\big)\right], \tag{5}$$

we have

$$S_i(y_i, z_i) = \frac{1}{\delta}\frac{\gamma_k(y_i) + z_i}{K(x_i, x_i)}. \tag{6}$$

For all the other values of z_i, $S_i(y_i, z_i) = \overline{S}_i(y_i, z_i)$, *where* $\overline{S}_i$ *is a differentiable function implicitly defined by*

$$a_i = -C\partial_2 V(y_i, \hat{f}(x_i; \mathbf{a}, \mathbf{b})) = -C\partial_2 V(y_i, a_i K(x_i, x_i)\delta - z_i(\mathbf{a}, \mathbf{b})). \tag{7}$$

The proof of Theorem 1, which is similar to that of Theorem 1 in (Dinuzzo et al. 2007), is reported in the Appendix. There are three differences with respect to (Dinuzzo et al. 2007). First, a more general class of loss functions, not restricted to regression ones, is considered. The second difference is the introduction of the parameter δ, which will play a key role in the proof of the subsequent Theorem 2. Note also that, for $\delta = 1$, the variables z_i will enter the definition of pseudo-residuals and pseudo-margins given in Sect. 4. Finally, Theorem 1 applies to both the no-bias and bias case.

It is also useful to introduce a matrix form representation. Let $\mathbf{z} := (z_1, \ldots, z_\ell)^T$ and introduce the nonlinear operator

$$\mathbf{S}(\mathbf{y}, \mathbf{z}) := (S_1(y_1, z_1), \ldots, S_\ell(y_\ell, z_\ell))^T.$$

Let matrices K and Ψ be such that $\mathsf{K}_{ij} := K(x_i, x_j)$, $\Psi_{ij} = \psi_j(x_i)$ and let D denote the diagonal part of K:

$$\mathsf{D} := \mathrm{diag}(K(x_1, x_1), \ldots, K(x_\ell, x_\ell)).$$

From the definition of the coefficients z_i, it follows that

$$\mathbf{z} = (\delta\mathsf{D} - \mathsf{K})\mathbf{a} - \Psi\mathbf{b},$$

so that, in view of (4) and (3), the algebraic characterization can be given in the following compact form.

Algebraic characterization of the optimum *Under the assumptions of Theorem 1, the following relationships hold*:

$$\begin{cases} \mathbf{a} = \mathbf{S}(\mathbf{y}, (\delta\mathsf{D} - \mathsf{K})\mathbf{a} - \Psi\mathbf{b}), \\ \Psi^T\mathbf{a} = 0. \end{cases} \tag{8}$$

or, equivalently,

$$\begin{cases} \mathbf{z} = (\delta\mathsf{D} - \mathsf{K})\mathbf{S}(\mathbf{y}, \mathbf{z}) - \Psi\mathbf{b}, \\ \Psi^T\mathbf{S}(\mathbf{y}, \mathbf{z}) = \mathbf{0}, \end{cases} \tag{9}$$

Theorem 1 above is the main result of the paper. In the next sections, we discuss its consequences. In particular, in Sect. 3, it is shown that the original non-smooth regularization problem admits an unconstrained smooth reformulation. In Sect. 4, the algebraic characterization of the optimum is specialized to regression and classification. Its application to a number of specific regression and classification methods is illustrated in the subsequent Sect. 5. Finally, the use of Theorem 1 to derive formulas for the effective degrees of freedom of kernel regression methods is the subject of Sect. 6.

3 A smooth reformulation

The next result is relative to problems with strictly positive kernel matrix K. It shows that, exploiting the freedom in the choice of δ, we can construct an equivalent unconstrained smooth optimization problem whose solution coincides with the coefficient vectors $\mathbf{z}$ and $\mathbf{b}$. Denote by $\|\cdot\|_{\mathsf{A}}$ the norm in $\mathbb{R}^\ell$ defined by $\|\mathbf{v}\|_{\mathsf{A}}^2 = \mathbf{v}^T\mathsf{A}\mathbf{v}$, where $\mathsf{A} \in \mathbb{R}^{\ell\times\ell}$ is a positive definite matrix.

Theorem 2 *Suppose that the matrix* K *is positive definite. Then, letting*

$$V_i^{eq}(y_i, z_i) := \frac{1}{C}\int_0^{z_i} S_i(y_i, s)ds,$$

there exists $\delta > 0$ *such that the functional*

$$\phi(\mathbf{u}, \mathbf{v}) := C\sum_{i=1}^{\ell} V_i^{eq}(y_i, u_i) + \frac{1}{2}\|\mathbf{u} + \Psi\mathbf{v}\|^2_{(\mathsf{K}-\delta\mathsf{D})^{-1}}$$

is well defined, convex, and has Lipschitz continuous gradient. Moreover,

$$(\mathbf{z}, \mathbf{b}) = \arg\min_{(\mathbf{u},\mathbf{v})\in\mathbb{R}^{\ell+m}} \phi(\mathbf{u}, \mathbf{v}).$$

Proof By Theorem 1, the function $S_i(y_i, z_i)$ is Lipschitz continuous with respect to z_i. Hence, $\sum_{i=1}^{\ell} V_i^{eq}(y_i, u_i)$ has Lipschitz continuous gradient with respect to $\mathbf{u}$. Now, we show that there exists $\delta > 0$ such that the functional $\phi(\mathbf{u}, \mathbf{v})$ is well defined. Indeed, letting $\{\lambda_i\}$ denote the eigenvalues of K, it suffices to take

$$0 < \delta < \frac{\min_i \lambda_i}{\max_i K(x_i, x_i)},$$

so that the matrix $\mathsf{K} - \delta\mathsf{D}$ has strictly positive eigenvalues. In fact, $\min_i \lambda_i > 0$ because K is positive definite. This means that the functional ϕ is well defined and has Lipschitz continuous gradient.

Since ϕ is smooth, each minimizer $(\mathbf{u}^*, \mathbf{v}^*)$ satisfies the necessary conditions for optimality:

$$\frac{\partial\phi(\mathbf{u}^*, \mathbf{v}^*)}{\partial\mathbf{u}} = \mathbf{S}(\mathbf{y}, \mathbf{u}^*) + (\mathsf{K} - \delta\mathsf{D})^{-1}(\mathbf{u}^* + \Psi\mathbf{v}^*) = \mathbf{0},$$
$$\frac{\partial\phi(\mathbf{u}^*, \mathbf{v}^*)}{\partial\mathbf{v}} = \Psi^T(\mathsf{K} - \delta\mathsf{D})^{-1}(\mathbf{u}^* + \Psi\mathbf{v}^*) = \mathbf{0}.$$

By substituting the first expression in the second and multiplying the first expression by $(\delta \mathsf{D} - \mathsf{K})$, from (9) it follows that $\mathbf{u}^* = \mathbf{z}$, $\mathbf{v}^* = \mathbf{b}$.

Now, we show that ϕ is strictly convex so that $(\mathbf{z}, \mathbf{b})$ is actually the only minimizer. Apart from a finite set of discontinuity points, there exists the Hessian of ϕ, given by

$$\nabla^2 \phi(\mathbf{u}, \mathbf{v}) = \begin{pmatrix} \frac{\partial \mathbf{S}(\mathbf{y}, \mathbf{u})}{\partial \mathbf{u}} & 0 \\ 0 & 0 \end{pmatrix} + (\mathsf{I} \quad \Psi^T)(\mathsf{K} - \delta \mathsf{D})^{-1} \begin{pmatrix} \mathsf{I} \\ \Psi \end{pmatrix}.$$

Since functions $S_i(y_i, \cdot)$ are monotone non-decreasing, the first term is positive semi-definite. In addition, $(\mathsf{K} - \delta \mathsf{D})^{-1}$ is positive definite, so that the Hessian $\nabla^2 \phi(\mathbf{u}, \mathbf{v})$ is positive definite whenever it exists. This proves that $\phi(\mathbf{u}, \mathbf{v})$ is strictly convex attaining its global minimum at $(\mathbf{z}, \mathbf{b})$. □

The equivalent optimization problem, though very similar to the original one, is always differentiable even when the original loss function $V(y_i, \cdot)$ is not. The smooth loss functions $V_i^{eq}(y_i, z_i)$ that appear in Theorem 2 are easily obtained by integration (often in closed-form) given the functions $S_i(\cdot, \cdot)$. Once $\mathbf{z}$ is known, it is straightforward to obtain vector $\mathbf{a}$ by applying (4). Theorem 2 is remarkable because it provides the constructive equivalence with an unconstrained smooth problem. Proving its computational relevance is beyond the scopes of the present paper, and is left to future investigation.

4 Algebraic characterization for regularized regression and classification

Equations (8) and (9) apply to general convex loss functions and give an analytic characterization for the coefficients a_i of the solution of the regularized problem (1). Without loss of generality, hereafter it will be assumed that $\delta = 1$. Two important classes of loss functions V are:

1. Symmetric loss functions that depend directly on the *residual* $y_i - \hat{f}(x_i)$ (regression loss functions):

$$V(y_i, \hat{f}(x_i)) = \widetilde{V}(\hat{f}(x_i) - y_i) = \widetilde{V}(y_i - \hat{f}(x_i)) = \widetilde{V}(y_i + z_i - a_i K(x_i, x_i)).$$

2. Loss functions that depend directly on the *margin* $y_i \hat{f}(x_i)$ (classification loss functions):

$$V(y_i, \hat{f}(x_i)) = \widetilde{V}(y_i \hat{f}(x_i)) = \widetilde{V}(y_i(a_i K(x_i, x_i) - z_i)).$$

The following sub-sections specialize the new analytic characterization to these two classes.

Before proceeding, let us summarize briefly the procedure that will be used to obtain the functions S_i starting from the loss function V.

1. Determine the points γ_k where $V(y_i, \cdot)$ is not twice differentiable.
2. For each γ_k, evaluate left and right derivatives $D^-(\gamma_k)$, $D^+(\gamma_k)$ of $V(y_i, \cdot)$ and construct an interval I_k according to (5).
3. In each interval I_k, the linear relationship (6) holds.
4. In all the other points, we have $a_i = \overline{S}_i(y_i, z_i)$, where $\overline{S}_i(y_i, z_i)$ is implicitly defined by (7).

4.1 Algebraic characterization for regularized regression

In general, the functions $S_i(y_i, z_i)$ depend on y_i and z_i separately. This section shows how the analytic characterization of (Dinuzzo et al. 2007) follows from the more general result given in Theorem 1 above. In particular, in the case of regression loss functions, it is possible to introduce a single coefficient η_i that captures the dependence on both y_i and z_i, thus simplifying the representation.

First, note that, if the function $\widetilde{V}(\cdot)$ is convex, then the function $V(y_i, \cdot)$ is convex with respect to the second argument and, as such, it fits into the framework of Theorem 1. It is easy to see that the points where $V(y_i, \cdot)$ is not twice differentiable are $\gamma_k(y_i) = \gamma_k^0 + y_i$, where γ_k^0 are the points where the function $V(0, \cdot)$ is not twice differentiable.

The nonlinear equation (7) becomes:

$$a_i = C\widetilde{V}'(y_i + z_i - a_i K(x_i, x_i)). \tag{10}$$

Let $\widetilde{D}^-(\gamma)$ and $\widetilde{D}^+(\gamma)$ denote the left and right derivatives of $\widetilde{V}(\cdot)$ at γ.

Let us calculate the derivatives $D^-(\gamma_k(y_i))$ and $D^+(\gamma_k(y_i))$:

$$\begin{aligned} D^-(\gamma_k(y_i)) &= \lim_{h\to 0^-} \frac{V(y_i, \gamma_k + h) - V(y_i, \gamma_k)}{h}, \\ &= \lim_{h\to 0^-} \frac{V(y_i, \gamma_k^0 + y_i + h) - V(y_i, \gamma_k^0 + y_i)}{h}, \\ &= \lim_{h\to 0^-} \frac{\widetilde{V}(y_i - \gamma_k^0 - y_i - h) - \widetilde{V}(y_i - \gamma_k^0 - y_i)}{h} = \widetilde{D}^-(\gamma_k^0). \end{aligned}$$

This implies that $D^-(\gamma_k(y_i))$ is, in fact, independent of y_i. For $D^+(\gamma_k(y_i))$ we have a completely similar result:

$$D^+(\gamma_k(y_i)) = \lim_{h\to 0^+} \frac{\widetilde{V}(\gamma_k^0 + h) - \widetilde{V}(\gamma_k^0)}{h} = \widetilde{D}^+(\gamma_k^0).$$

Now, using (5), we can compute the intervals I_k where the function $S_i(y_i, z_i)$ is affine in z_i:

$$I_k = \left[-(\gamma_k^0 + y_i + CK(x_i, x_i)\widetilde{D}^+(\gamma_k^0)), -(\gamma_k^0 + y_i + CK(x_i, x_i)\widetilde{D}^-(\gamma_k^0))\right]. \tag{11}$$

In these intervals, we have the following affine relationship between a_i and z_i:

$$a_i = \frac{\gamma_k^0 + y_i + z_i}{K(x_i, x_i)}. \tag{12}$$

From (10) , (11) and (12) one can conclude that the single variable

$$\eta_i := y_i + z_i = y_i + a_i K(x_i, x_i) - \hat{f}(x_i),$$

suffices to capture the combined effect of both y_i and z_i. Indeed, (10), (11) and (12) become, respectively,

$$a_i = C\widetilde{V}'(\eta_i - a_i K(x_i, x_i)), \tag{13}$$

$$\eta_i \in \widetilde{I}_k := \left[-(\gamma_k^0 + CK(x_i, x_i)\widetilde{D}^+(\gamma_k^0)), -(\gamma_k^0 + CK(x_i, x_i)\widetilde{D}^-(\gamma_k^0))\right], \tag{14}$$

$$a_i = \frac{\gamma_k^0 + \eta_i}{K(x_i, x_i)}.$$

In conclusion, the algebraic characterization for the coefficients a_i involves functions $\widetilde{S}_i$ depending on just one variable η_i:

$$a_i = S_i(y_i, z_i) = \widetilde{S}_i(y_i + z_i) = \widetilde{S}_i(\eta_i),$$

where

$$\widetilde{S}_i(\eta_i) = \begin{cases} \frac{\gamma_k^0 + \eta_i}{K(x_i, x_i)}, & \eta_i \in \widetilde{I}_k, \\ \overline{S}_i(\eta_i), & \text{otherwise} \end{cases} \tag{15}$$

and $\overline{S}_i(\eta_i)$ is the implicit function defined by (13) that ties a_i to η_i in the differentiability points for V. Note that the coefficients η_i are such that

$$\eta_i = y_i + z_i = \big(y_i - \hat{f}(x_i)\big) + a_i K(x_i, x_i),$$

and can be interpreted as *pseudo*-residuals (the residual plus the term $a_i K(x_i, x_i)$). Letting

$$\boldsymbol{\eta} := (\eta_1, \ldots, \eta_\ell)^T, \qquad \widetilde{\mathbf{S}}(\boldsymbol{\eta}) := (\widetilde{S}_1(\eta_1), \ldots, \widetilde{S}_\ell(\eta_\ell))^T,$$

we have the following two (equivalent) systems in matrix form

$$\begin{cases} \mathbf{a} = \widetilde{\mathbf{S}}(\mathbf{y} + (\mathsf{D} - \mathsf{K})\mathbf{a} - \Psi\mathbf{b}), \\ \Psi^T\mathbf{a} = \mathbf{0}, \end{cases} \qquad \begin{cases} \boldsymbol{\eta} = \mathbf{y} + (\mathsf{D} - \mathsf{K})\widetilde{\mathbf{S}}(\boldsymbol{\eta}) - \Psi\mathbf{b}, \\ \Psi^T\widetilde{\mathbf{S}}(\boldsymbol{\eta}) = \mathbf{0}. \end{cases} \tag{16}$$

4.2 Algebraic characterization for regularized classification

Let now consider margin-dependent loss functions:

$$V(y_i, \hat{f}(x_i)) = \widetilde{V}(y_i \hat{f}(x_i)) = \widetilde{V}(y_i(a_i K(x_i, x_i) - z_i)).$$

As in the case of regression loss functions, convexity of $\widetilde{V}(\cdot)$ implies convexity of $V(y_i, \cdot)$. In fact, the value y_i acts merely as a scale factor. Let γ_k^1 denote the points where $V(1, \cdot)$ is not twice differentiable. Then, for $y_i \neq 0$, the points $\gamma_k(y_i)$ where $V(y_i, \cdot)$ is not twice differentiable are given by

$$\gamma_k(y_i) = \frac{\gamma_k^1}{y_i}.$$

In the differentiability points, using the structure of the loss function V, we can simplify equation (7) to obtain:

$$a_i = -C y_i \widetilde{V}'(y_i a_i K(x_i, x_i) - y_i z_i). \tag{17}$$

Let us calculate $D^-(\gamma_k(y_i))$ and $D^+(\gamma_k(y_i))$.

$$\begin{aligned} D^-(\gamma_k(y_i)) &= \lim_{h \to 0^-} \frac{V(y_i, \gamma_k + h) - V(y_i, \gamma_k)}{h} = \lim_{h \to 0^-} \frac{\widetilde{V}(y_i(\frac{\gamma_k^1}{y_i} + h)) - \widetilde{V}(y_i(\frac{\gamma_k^1}{y_i}))}{h}, \\ &= \lim_{h \to 0^-} \frac{\widetilde{V}(\gamma_k^1 + y_i h) - \widetilde{V}(\gamma_k^1)}{h} = y_i \widetilde{D}^-(\gamma_k^1), \end{aligned}$$

where $\widetilde{D}^-(\gamma)$ denotes the left derivative of $\widetilde{V}(\cdot)$ evaluated at γ. In a similar way,

$$D^+(\gamma_k(y_i)) = y_i \widetilde{D}^+(\gamma_k^1).$$

Now, we can use (5) in order to compute intervals I_k:

$$I_k = \left[-\left(\frac{\gamma_k^1}{y_i} + y_i C K(x_i, x_i) \widetilde{D}^+(\gamma_k^1) \right), -\left(\frac{\gamma_k^1}{y_i} + y_i C K(x_i, x_i) \widetilde{D}^-(\gamma_k^1) \right) \right]. \tag{18}$$

When $z_i \in I_k$, we can apply (6) to obtain

$$a_i = \frac{\frac{\gamma_k^1}{y_i} + z_i}{K(x_i, x_i)} = \frac{1}{y_i} \frac{\gamma_k^1 + y_i z_i}{K(x_i, x_i)}. \tag{19}$$

In the classification case, in view of (17), (18), (19), there does not exist a simple transformation, similar to the one used in the previous sub-section, able to capture the effect of both y_i and z_i with a single coefficient. However, in a classification context, the following assumption is natural: y_i can take only the values $+1$ and -1. Then, we have $y_i = \frac{1}{y_i}$ and this naturally leads to the definition of the *pseudo*-margins η_i and a new set of coefficients α_i:

$$\eta_i := y_i z_i = y_i (a_i K(x_i, x_i) - \hat{f}(x_i)), \quad \alpha_i := y_i a_i. \tag{20}$$

We can immediately obtain the system

$$\alpha_i = -C \widetilde{V}'(\alpha_i K(x_i, x_i) - \eta_i), \tag{21}$$

valid in the differentiability points. Moreover, when

$$\eta_i \in \widetilde{I}_k = \left[-(\gamma_k^1 + C K(x_i, x_i) \widetilde{D}^+(\gamma_k^1)), -(\gamma_k^1 + C K(x_i, x_i) \widetilde{D}^-(\gamma_k^1)) \right],$$

we have

$$\alpha_i = \frac{\gamma_k^1 + \eta_i}{K(x_i, x_i)}. \tag{22}$$

Hence, everything is formally similar to the regression case, the only difference being the modified definition of η_i and the introduction of the coefficients α_i replacing the original a_i. Thus, we have shown that there exist functions $\widetilde{S}_i$ such that

$$\alpha_i = S_i(y_i, z_i) = \widetilde{S}_i(y_i z_i) = \widetilde{S}_i(\eta_i), \tag{23}$$

$$\widetilde{S}_i(\eta_i) = \begin{cases} \frac{\gamma_k^1 + \eta_i}{K(x_i, x_i)}, & \eta_i \in \widetilde{I}_k, \\ \overline{S}_i(\eta_i), & \text{otherwise} \end{cases}$$

where $\overline{S}_i$ is a function implicitly defined by (21).

Observe that the usual way to express the solution for the Support Vector Classifier is a formula of the type:

$$\hat{f}(x) = \sum_{i=1}^{\ell} y_i \alpha_i K(x_i, x),$$

where α_i are Lagrange multipliers for the dual problem. This means that the coefficients α_i, here defined without passing through the dual problem, are precisely the Lagrange multipliers of the dual formulation. It is worth noting that the usual Kuhn-Tucker conditions contain also some inequalities and do not define the set of Lagrange multipliers by means of a system of algebraic equation like (23). To obtain a matrix form representation, let us introduce the matrix

$$\mathsf{Y}_D := \operatorname{diag}(y_1, \ldots, y_\ell),$$

and note that $\mathsf{Y}_D^2 = \mathsf{I}$. Then, from (20) and (23), we have

$$\begin{aligned}\boldsymbol{\eta} &= \mathsf{Y}_D\big((\mathsf{D}-\mathsf{K})\mathsf{Y}_D\widetilde{\mathbf{S}}(\boldsymbol{\eta}) - \Psi\mathbf{b}\big) = (\mathsf{Y}_D^2\mathsf{D} - \mathsf{Y}_D\mathsf{K}\mathsf{Y}_D)\widetilde{\mathbf{S}}(\boldsymbol{\eta}) - \mathsf{Y}_D\Psi\mathbf{b}\\ &= (\mathsf{D} - \mathsf{Y}_D\mathsf{K}\mathsf{Y}_D)\widetilde{\mathbf{S}}(\eta) - \mathsf{Y}_D\Psi\mathbf{b}.\end{aligned}$$

In conclusion,

$$\begin{cases}\alpha = \widetilde{\mathbf{S}}((\mathsf{D} - \mathsf{Y}_D\mathsf{K}\mathsf{Y}_D)\alpha - \Psi\mathbf{b}),\\ \Psi^T\mathsf{Y}_D\alpha = \mathbf{0},\\ \mathbf{a} = \mathsf{Y}_D\alpha.\end{cases}$$

5 Examples

The aim of this section is to illustrate by means of some examples the previously derived analytic characterization. We will be able to obtain algebraic characterization for the coefficients of many regularized kernel methods, including Support Vector Machines for both regression and classification.

It goes without saying that Theorem 1 is of interest especially for non-smooth loss functions. Nevertheless, for the sake of comparison, also some smooth losses will be discussed. It will be seen that the function $\widetilde{S}_i(\eta_i)$ provides a rather immediate visualization of some important properties, such as sparsity and robustness. In fact, sparsity is associated with a dead zone of the function $\widetilde{S}_i(\eta_i)$, whereas robustness corresponds to saturation at the extremes.

Example 1 (Quadratic loss) Consider the quadratic loss function

$$V(y_i, \hat{f}(x_i)) = \widetilde{V}(y_i - \hat{f}(x_i)) = \frac{1}{2}(y_i - \hat{f}(x_i))^2.$$

Then, problem (1) boils down to regularized kernel least-squares whose solution is characterized by a linear system for the coefficients a_i.

Note that the quadratic loss is of the regression type (residual-dependent) and is everywhere differentiable. Then, there are no intervals $\widetilde{I}_k$ to compute and it is sufficient to write (13) to obtain

$$a_i = C(\eta_i - a_i K(x_i, x_i)).$$

Solving for a_i, the following fixed-point equation easily follows:

$$a_i = \widetilde{S}_i(\eta_i) = \frac{C}{1 + CK(x_i, x_i)}\eta_i(\mathbf{a}, \mathbf{b}).$$

This means that the function $\widetilde{S}_i(\eta_i)$ is everywhere linear (see Fig. 3, top left).

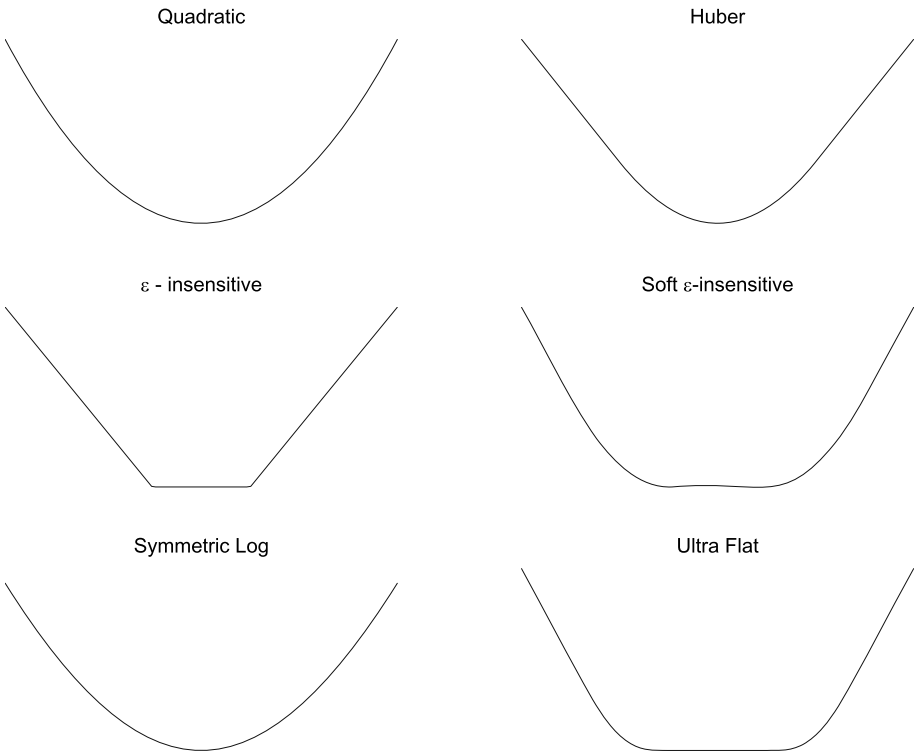

Fig. 1 Some common regression loss function

Let us write the matrix form (16) for the no-bias case:

$$\mathbf{a} = \widetilde{\mathbf{S}}(\boldsymbol{\eta}) = \left(\frac{\mathsf{I}}{C} + \mathsf{D}\right)^{-1} \boldsymbol{\eta} = \left(\frac{\mathsf{I}}{C} + \mathsf{D}\right)^{-1} (\mathbf{y} + (\mathsf{D} - \mathsf{K})\mathbf{a}).$$

From this expression we see that, in the case of regularized least-squares, the pseudo-residual characterization is equivalent to use the splitting:

$$\left(\frac{\mathsf{I}}{C} + \mathsf{K}\right) = \left(\frac{\mathsf{I}}{C} + \mathsf{D}\right) + (\mathsf{K} - \mathsf{D}),$$

to express the linear system

$$\left(\frac{\mathsf{I}}{C} + \mathsf{K}\right)\mathbf{a} = \mathbf{y}.$$

Example 2 (Huber's loss) Let now consider Huber's loss function

$$\widetilde{V}(y_i - \hat{f}(x_i)) = \begin{cases} \frac{1}{2}(y_i - \hat{f}(x_i))^2, & |y_i - \hat{f}(x_i)| \le d, \\ d|y_i - \hat{f}(x_i)| - \frac{d^2}{2}, & |y_i - \hat{f}(x_i)| > d. \end{cases}$$

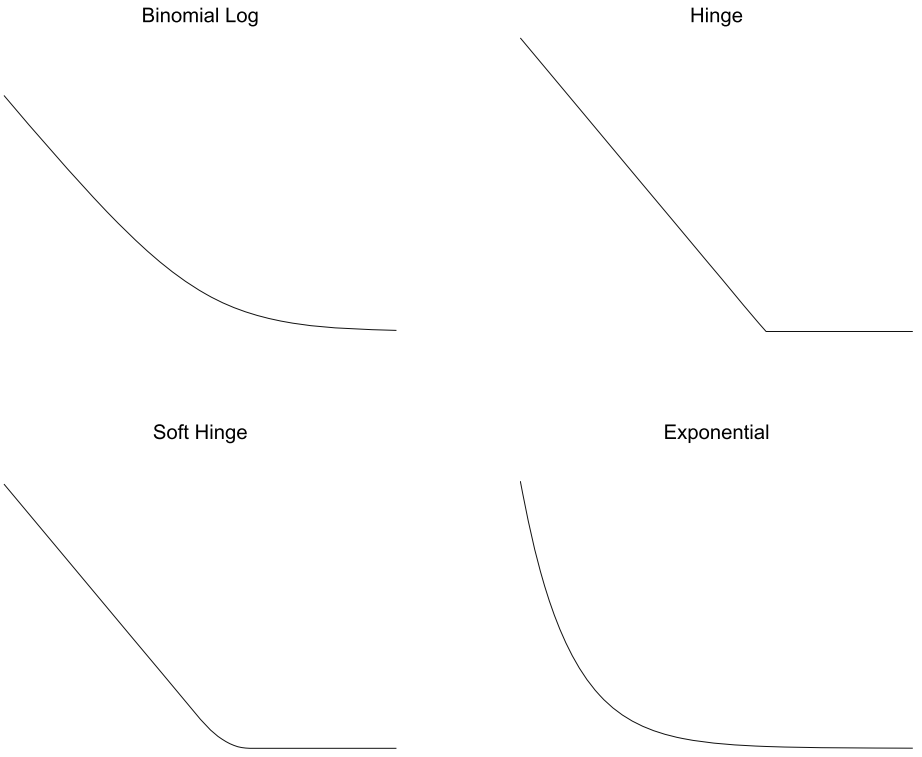

Fig. 2 Some common classification loss function

This function, as the previous one, is a regression loss that is everywhere twice differentiable except for a finite number of points. Let us write (13) in the two cases $|y_i - \hat{f}(x_i)| > d$ and $|y_i - \hat{f}(x_i)| < d$.

When $|y_i - \hat{f}(x_i)| > d$, it results that

$$a_i = dC \operatorname{sgn}(\eta_i - a_i K(x_i, x_i)).$$

If

$$\eta_i - a_i K(x_i, x_i) = y_i - \hat{f}(x_i) > 0,$$

then

$$a_i = dC, \qquad \eta_i = y_i - \hat{f}(x_i) + dC K(x_i, x_i) > d(1 + C K(x_i, x_i)).$$

Otherwise, if

$$\eta_i - a_i K(x_i, x_i) = y_i - \hat{f}(x_i) < 0,$$

it results that

$$a_i = -dC, \qquad \eta_i < -d(1 + C K(x_i, x_i)).$$

When $|y_i - \hat{f}(x_i)| < d$ we have:

$$a_i = C(\eta_i - a_i K(x_i, x_i)),$$

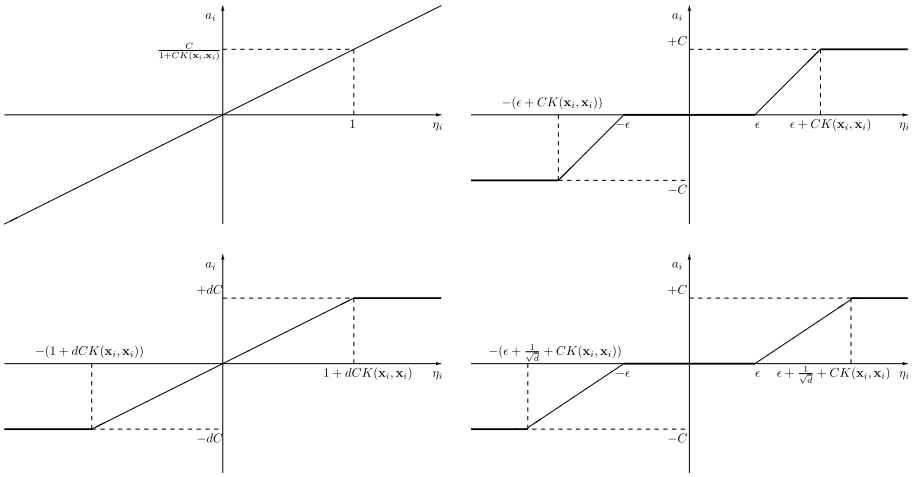

Fig. 3 Dependence of a_i on η_i for the regression methods analyzed in Examples 1–4. The top left panel is relative to the quadratic loss function (Example 1), the bottom left to the Huber's loss (Example 2), the top right to the ϵ-insensitive loss (Example 3), and the bottom right panel to the soft ϵ-insensitive loss (Example 4)

that is

$$a_i = \frac{C}{1 + CK(x_i, x_i)}\eta_i.$$

This last equation holds for

$$\eta_i \in (-d(1 + CK(x_i, x_i)), d(1 + CK(x_i, x_i))).$$

Finally, if $|y_i - \hat{f}(x_i)| = d$, observe that the left and right derivatives do coincide:

$$\widetilde{D}^-(\gamma_k^0) = \widetilde{D}^+(\gamma_k^0) = \widetilde{V}'(\gamma_k^0),$$

where $\gamma_1^0 = -d$, $\gamma_2^0 = d$. Then, the intervals $\widetilde{I}_k$ $(k = 1, 2)$, defined by (14), collapse to the points

$$\widetilde{I}_1 = \big\{-d(1 + CK(x_i, x_i))\big\}, \qquad \widetilde{I}_2 = \big\{d(1 + CK(x_i, x_i))\big\}.$$

In conclusion,

$$\begin{aligned} a_i &= \widetilde{S}_i(\eta_i(\mathbf{a}, \mathbf{b})) \\ &= \begin{cases} -dC, & \eta_i < -d(1 + CK(x_i, x_i)), \\ \frac{C}{1+CK(x_i,x_i)}\eta_i, & \eta_i \in [-d(1 + CK(x_i, x_i)), d(1 + CK(x_i, x_i))], \\ dC, & \eta_i > d(1 + CK(x_i, x_i)) \end{cases} \end{aligned} \tag{24}$$

as illustrated in Fig. 3, bottom left.

Example 3 (ϵ-insensitive loss) Let us consider the Vapnik's ϵ-insensitive loss function

$$\widetilde{V}(y_i - \hat{f}(x_i)) = \begin{cases} 0, & |y_i - \hat{f}(x_i)| \le \epsilon, \\ |y_i - \hat{f}(x_i)| - \epsilon, & |y_i - \hat{f}(x_i)| > \epsilon. \end{cases}$$

This loss specializes to the Laplace loss function when $\epsilon = 0$. It has two corner points for $\epsilon > 0$ ($\gamma_1^0 = -\epsilon$, $\gamma_2^0 = +\epsilon$), while just one for $\epsilon = 0$ ($\gamma_1^0 = 0$). As shown in (Dinuzzo et al. 2007),

$$a_i = \widetilde{S}_i(\eta_i(\mathbf{a}, \mathbf{b})) = \begin{cases} -C, & \eta_i \leq -(\epsilon + CK(x_i, x_i)), \\ \frac{\eta_i + \epsilon}{K(x_i, x_i)}, & -(\epsilon + CK(x_i, x_i)) \leq \eta_i \leq -\epsilon, \\ 0, & -\epsilon < \eta_i < \epsilon, \\ \frac{\eta_i - \epsilon}{K(x_i, x_i)}, & \epsilon < \eta_i < (\epsilon + CK(x_i, x_i)), \\ C, & \eta_i \geq (\epsilon + CK(x_i, x_i)) \end{cases} \tag{25}$$

which is illustrated in Fig. 3 (top right panel). Observe that for $\epsilon = 0$ the central plateau disappears and we have a simplified expression:

$$a_i = \begin{cases} -C, & \eta_i < -CK(x_i, x_i), \\ \frac{\eta_i}{K(x_i, x_i)}, & -CK(x_i, x_i) \leq \eta_i \leq +CK(x_i, x_i), \\ C, & \eta_i > CK(x_i, x_i). \end{cases}$$

Compare this expression with (24) obtained in Example 2 relative to Huber's loss function and let $d = 1$. The similarity is apparent and suggests the following: to obtain a linear trait in the function $\widetilde{S}_i$ we can either insert a quadratic trait in the loss function or a corner point. This fact can be exploited, for example, in order to approximate non-smooth loss functions using quadratic traits near the corner points. This possibility is explored in the subsequent example.

Example 4 (Soft ϵ-insensitive loss) Suppose we want to approximate the ϵ-insensitive function analyzed in the previous example by means of an everywhere smooth function. The final observation of the previous example suggests the following family:

$$V(y_i, \hat{f}(x_i)) = \begin{cases} 0, & |y_i - \hat{f}(x_i)| \leq \epsilon, \\ d^2 \frac{(|y_i - \hat{f}(x_i)| - \epsilon)^2}{2}, & \epsilon \leq |y_i - \hat{f}(x_i)| \leq \epsilon + \frac{1}{d}, \\ |y_i - \hat{f}(x_i)| - \frac{1}{d} - \epsilon + \frac{1}{2}, & |y_i - \hat{f}(x_i)| \geq \epsilon + \frac{1}{d}. \end{cases} \tag{26}$$

In the last expression, the corner points of the ϵ-insensitive loss function have been replaced by quadratic traits whose length is regulated by a positive parameter d.

Proceeding as in the previous examples, the following expression is easily obtained:

$$a_i = \widetilde{S}_i(\eta_i(\mathbf{a}, \mathbf{b})) = \begin{cases} -C, & \eta_i \leq -(\epsilon + \frac{1}{d} + CK(x_i, x_i)), \\ \frac{\eta_i + \epsilon}{\frac{1}{d} + K(x_i, x_i)}, & -(\epsilon + \frac{1}{d} + CK(x_i, x_i)) < \eta_i < -\epsilon, \\ 0, & -\epsilon \leq \eta_i \leq \epsilon, \\ \frac{\eta_i - \epsilon}{\frac{1}{d} + K(x_i, x_i)}, & \epsilon < \eta_i < (\epsilon + \frac{1}{d} + CK(x_i, x_i)), \\ C, & \eta_i \geq (\epsilon + \frac{1}{d} + CK(x_i, x_i)). \end{cases} \tag{27}$$

Apparently, (27) reduces exactly to (25) when $d \to +\infty$. Therefore, if d is large enough (say $d \gg K(x_i, x_i)$, $\forall i$), the solution associated with the smooth loss function (26) will be a close approximation to the solution associated with the (non-smooth) ϵ-insensitive loss function.

Of course, in order to approximate the ϵ-insensitive loss function, it is possible to use other families different from (26). Any good approximation should preserve the two fundamental property of Support Vector Regression: robustness and sparsity. With reference to the functions $\widetilde{S}_i(\eta_i)$, the robustness properties is associated with saturation as $\eta_i \to +\infty$, whereas sparsity is associated with the dead zone around $\eta_i = 0$, see Fig. 3. There is a rather flexible family of loss functions, called SILF (Soft Insensitive Loss Function) (Chu et al. 2001), that includes as particular cases all the functions presented in Examples 1–4. However, in the following, we will use the name *soft insensitive loss function* to indicate only (26).

Example 5 (Symmetric ϵ-insensitive logistic loss) In this and in the next example we analyze two additional regression loss functions that can be used to approximate the ϵ-insensitive loss and are not obtainable from the SILF family. They are infinitely differentiable and, as such, they are more regular than the soft insensitive loss function.

The *symmetric ϵ-insensitive logistic loss* (Dekel et al. 2005) is defined as

$$\widetilde{V}(y_i - \hat{f}(x_i)) = \log(1 + e^{y_i - \hat{f}(x_i) - \epsilon}) + \log(1 + e^{\hat{f}(x_i) - y_i - \epsilon}) - 2\log(1 + e^{-\epsilon}).$$

From (13) we have

$$a_i = C\left(\frac{1}{e^{a_i K(x_i,x_i)+\epsilon-\eta_i}+1} - \frac{1}{e^{-a_i K(x_i,x_i)+\epsilon+\eta_i}+1}\right).$$

This time we do not find a closed form expression for the coefficient α_i as a function of the pseudo-margin η_i. However, we can easily derive the inverse relationship:

$$\eta_i = a_i K(x_i, x_i) + \log\left(\frac{C - a_i}{\sqrt{a_i^2 + \frac{C^2 - a_i^2}{(\cosh\epsilon)^2}} + a_i}\right) - \log(\cosh\epsilon).$$

Again, robustness is preserved ($-C \le a_i \le C$) and a very smooth relationship between the coefficients and the pseudo-residuals holds. However, the sparsity property is lost. In the next example we propose a new loss function that guarantees robustness, smoothness and sparsity.

Example 6 (Ultraflat loss) Let

$$\rho(\tau) = e^{-\frac{1}{1-\tau^2}}\chi_{[-1,1]}(\tau),$$

where $\chi_{[-1,1]}(\tau)$ is the indicator function of the interval $[-1, 1]$, that is

$$\chi_{[-1,1]}(\tau) = \begin{cases} 1, & \tau \in [-1, 1], \\ 0, & \tau \notin [-1, 1]. \end{cases}$$

The function $\rho(\tau)$ is infinitely differentiable and has finite integral

$$A := \int_{-\infty}^{+\infty} \rho(\tau) d\tau.$$

Let us define the *symmetric ultra-flat* family of loss functions, see Fig. 1:

$$\widetilde{V}(y_i - \hat{f}(x_i)) = \frac{d}{A}\int_0^{|y_i-\hat{f}(x_i)|}\int_0^{v}\rho(ud-\epsilon d-1)dudv.$$

Observe that we can also write:

$$\widetilde{V}(\tau) = \begin{cases} 0, & |\tau| \le \epsilon, \\ \frac{1}{A}\int_{-1}^{-1+d(|\tau|-\epsilon)}(|\tau|-\epsilon-\frac{1+v}{d})e^{-\frac{1}{1-v^2}}dv, & \epsilon < |\tau| < \epsilon + \frac{2}{d}, \\ |\tau|-\epsilon-\frac{1}{d}, & |\tau| \ge \epsilon + \frac{2}{d}. \end{cases}$$

We have obtained a family of loss functions that are infinitely differentiable and preserve both robustness and sparsity. Note that, as d tends to $+\infty$, we approach the ϵ-insensitive loss function. The price we must pay for these properties is that (13) is not easily solved for a_i or η_i due to the rather complicated expression of the loss function. However, it is easily checked that $a_i = 0$ for $|\eta_i| \le \epsilon$, while $a_i = \pm C$ for $|\eta_i| \ge \epsilon + CK(x_i, x_i) + \frac{2}{d}$. Conversely, when

$$\epsilon \le |\eta_i| \le \epsilon + CK(x_i, x_i) + \frac{2}{d},$$

we have

$$a_i = \operatorname{sgn}(\eta_i)\frac{C}{A}\int_{-1}^{-1+d(|\eta_i - a_i K(x_i,x_i)|-\epsilon)} e^{-\frac{1}{1-v^2}}dv,$$

which cannot be easily solved for either a_i or η_i.

Example 7 (Binomial log-likehood) In this example and in the following ones we consider margin-dependent loss functions. Let us start with the binomial log-likehood that is differentiable everywhere:

$$\widetilde{V}(y_i\hat{f}(x_i)) = \log(1 + e^{-y_i\hat{f}(x_i)}).$$

Having no differentiability problems, we can simply write the system (21):

$$\alpha_i = C\frac{e^{-(\alpha_i K(x_i,x_i)-\eta_i)}}{1+e^{-(\alpha_i K(x_i,x_i)-\eta_i)}}.$$

As in Example 5, we do not find a closed form expression for α_i as a function of η_i. Nevertheless, one can obtain the inverse relation

$$\eta_i = \alpha_i K(x_i, x_i) + \log\left(\frac{\alpha_i}{C-\alpha_i}\right).$$

From this expression we see that α_i must lie in the interval $[0, C]$. Moreover, in accordance with Theorem 1, $\widetilde{S}_i$ is an increasing function of η_i. The coefficients α_i will never reach exactly the values 0 and C, but will always lie in the open interval $(0, C)$. In fact, boundary values for α_i imply infinite values for the pseudo-margins η_i, but this is impossible because the pseudo-margins are always limited:

$$|\eta_i| = \left|-y_i\sum_{j\neq i}^{\ell} a_j K(x_i, x_j)\right| \le \sum_{j\neq i}^{\ell}|a_j K(x_i, x_j)| \le \ell C \max_j |K(x_i, x_j)| < +\infty.$$

Example 8 (Hinge loss) The "hinge" loss function, associated with the Support Vector Classifier, is defined by

$$\widetilde{V}(y_i \hat{f}(x_i)) = \max\{0, 1 - y_i \hat{f}(x_i)\} = (1 - y_i \hat{f}(x_i))_+,$$

where $(\cdot)_+$ denotes the positive part.

The first step of the procedure for obtaining the function $\widetilde{S}_i$ described in Sect. 3 is to calculate the points γ_k^1 where $V(1, \cdot)$ is not twice differentiable. For the hinge loss there is only one such point: $\gamma_1^1 = 1$. Left and right derivatives are given by:

$$\widetilde{D}^-(1) = -1, \qquad \widetilde{D}^+(1) = 0.$$

Thus, we can determine the interval

$$\widetilde{I}_1 = [-1, -1 + CK(x_i, x_i)]$$

for the pseudo-margin η_i within which the affine relationship (22) holds:

$$\alpha_i = \frac{1 + \eta_i}{K(x_i, x_i)}.$$

From this relationship we see that α_i ranges from 0 to C. When $\alpha_i K(x_i, x_i) - \eta_i < 1$, we have from (21) that:

$$\alpha_i = C, \qquad \eta_i > -1 + CK(x_i, x_i).$$

At last, for

$$\alpha_i K(x_i, x_i) - \eta_i > 1,$$

we have

$$\alpha_i = 0, \qquad \eta_i < -1.$$

Putting pieces together, we obtain (see Fig. 4):

$$\alpha_i = \widetilde{S}_i(\eta_i(\mathbf{a}, \mathbf{b})) = \begin{cases} 0, & \eta_i < -1, \\ \frac{1+\eta_i}{K(x_i, x_i)}, & -1 \le \eta_i \le -1 + CK(x_i, x_i), \\ C, & \eta_i > -1 + CK(x_i, x_i). \end{cases}$$

In this case, since the values 0 and C can be actually reached by some coefficient, the classifier, as is well known, enjoys the typical sparseness property of Support Vector Machines. Interestingly, the expression for $\widetilde{S}_i$ is formally equivalent to an SMO (Sequential Minimal Optimization) update, see Platt (1998).

Example 9 (Soft Hinge loss) Also for the "hinge" loss function we can work out approximations having a quadratic trait that replaces the corner point. Consider the following class of functions:

$$\widetilde{V}(y_i \hat{f}(x_i)) = \begin{cases} \frac{3}{2} - \frac{1}{d} - y_i \hat{f}(x_i), & y_i \hat{f}(x_i) < 1 - \frac{1}{d}, \\ d^2 \frac{(y_i \hat{f}(x_i) - 1)^2}{2}, & 1 - \frac{1}{d} < y_i \hat{f}(x_i) < 1, \\ 0, & y_i \hat{f}(x_i) > 1. \end{cases}$$

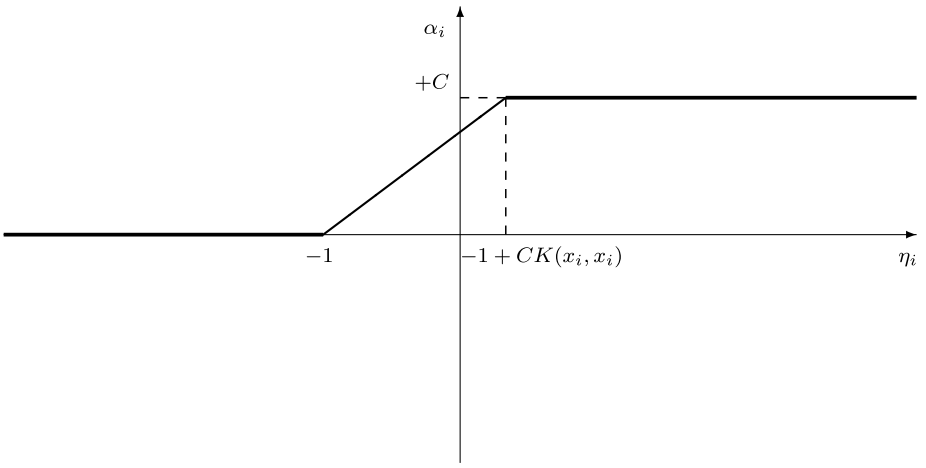

Fig. 4 Hinge loss (Example 6): dependence of α_i on η_i

Then, we obtain

$$\alpha_i = \widetilde{S}_i(\eta_i(\mathbf{a}, \mathbf{b})) = \begin{cases} 0, & \eta_i < -1, \\ \frac{1+\eta_i}{K(x_i,x_i)+\frac{1}{d}}, & -1 < \eta_i < -1 + CK(x_i, x_i) + \frac{1}{d}, \\ C, & \eta_i > -1 + CK(x_i, x_i) + \frac{1}{d}. \end{cases}$$

Example 10 (Adaboost loss) The "Adaboost" (exponential) loss function

$$\widetilde{V}(y_i \hat{f}(x_i)) = e^{-y_i \hat{f}(x_i)}$$

is everywhere differentiable. Then, it is sufficient to write the system (21):

$$\alpha_i = Ce^{-(\alpha_i K(x_i,x_i)-\eta_i)}.$$

As in the binomial log-likehood case, we cannot find a closed-form expression for α_i. However, we can solve for η_i:

$$\eta_i = \log\left(\frac{\alpha_i}{C}\right) + \alpha_i K(x_i, x_i).$$

From this expression we see that the coefficients α_i are always strictly positive. Moreover, they are always strictly increasing as a function of η_i (ranging from 0 to $+\infty$). When $\eta_i \to +\infty$, α_i tends to an asymptote with slope $\frac{1}{K(x_i,x_i)}$.

6 Degrees of freedom of kernel regression methods

In this section, attention is focused on regression methods and a probabilistic homoskedastic model is assumed for the generation of the dataset D. More precisely, it is assumed that $Y = \mathbb{R}$ and pairs (x_i, y_i) are i.i.d. (independently and identically distributed) samples drawn from

a probability measure P_{XY} on $X \times Y$. The conditional expectation of Y given X is defined as $E[Y|X=x] := \int_{\mathbb{R}} y dP_{Y|X=x}$. Hereafter, it is assumed that the conditional variance

$$\sigma^2 := \int_{\mathbb{R}} (y - E[Y|X=x])^2 dP_{Y|X=x}$$

is independent of x (homoskedasticity) and satisfies $0 < \sigma^2 < +\infty$. A convenient definition for the degrees of freedom of a nonlinear regression estimator $\hat{f}(x)$, see e.g. (Efron 2004), is

$$df := \sum_{i=1}^{\ell} \frac{\mathrm{cov}(\hat{f}(x_i), y_i)}{\sigma^2}.$$

Under mild assumptions, it can be shown that the formula

$$\widehat{df} = \sum_{i=1}^{\ell} \frac{\partial \hat{f}(x_i)}{\partial y_i}, \tag{28}$$

provides an unbiased estimate for df, see e.g. (Stein 1981).

In this section, we exploit the analytic characterization of Sect. 2 in order to obtain $\widehat{df}$ for a generic regression kernel estimator. Let

$$\Lambda(\boldsymbol{\eta}) := \frac{\partial \widetilde{\mathbf{S}}(\eta)}{\partial \boldsymbol{\eta}},$$

and introduce the following index sets

$$I := \{1, \ldots, \ell\}, \qquad I_D := \{i \in I : \forall k = 1, \ldots, n_\gamma^i,\ \eta_i^* \notin \widetilde{I}_k\}, \qquad I_N := I \setminus I_D,$$

where $\boldsymbol{\eta}^* := (\eta_1^* \ \ldots \ \eta_\ell^*)^T$ solves (16). Let p denote the cardinality of I_N. Then, I_D has cardinality $\ell - p$ and we can partition vectors and matrices according to these index sets:

$$\mathsf{K} = \begin{pmatrix} \mathsf{K}_{DD} & \mathsf{K}_{DN} \\ \mathsf{K}_{ND} & \mathsf{K}_{NN} \end{pmatrix}, \qquad \mathsf{D} = \begin{pmatrix} \mathsf{D}_D & 0 \\ 0 & \mathsf{D}_N \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \Lambda_D & 0 \\ 0 & \Lambda_N \end{pmatrix}.$$

It is easily seen that

$$\Lambda_D = \mathrm{diag}\left(\frac{C\partial_2^2 V(y_1, \hat{f}(x_1))}{1 + CK(x_1,x_1)\partial_2^2 V(y_1, \hat{f}(x_1))}, \ldots, \frac{C\partial_2^2 V(y_p, \hat{f}(x_p))}{1 + CK(x_p,x_p)\partial_2^2 V(y_p, \hat{f}(x_p))} \right),$$

$$\Lambda_N = \mathsf{D}_N^{-1}.$$

Theorem 3 *Consider a generic kernel regression estimator without bias. Let the hypotheses of Theorem 1 be satisfied and assume that system (16) without bias (i.e. $\Psi = 0$) admits a unique solution $\boldsymbol{\eta}^*$ such that none of the coefficients η_i^* lies on the boundaries of the intervals $\widetilde{I}_k$ defined in (14). Then, the estimate (28) of the degrees of freedom is given by*

$$\widehat{df} = \mathrm{tr}\big[\mathsf{K}\Lambda^*(\mathsf{B}^*)^{-1}\big] = \mathrm{tr}\big[\mathsf{K}\Lambda^*(\mathsf{I} + (\mathsf{K} - \mathsf{D})\Lambda^*)^{-1}\big], \tag{29}$$

where

$$\Lambda^* := \operatorname{diag}\left(\frac{\partial \widetilde{S}_1}{\partial \eta_1}(\eta_1^*), \ldots, \frac{\partial \widetilde{S}_\ell}{\partial \eta_\ell}(\eta_\ell^*)\right), \qquad \mathsf{B}^* := \mathsf{I} + (\mathsf{K} - \mathsf{D})\Lambda^*.$$

Proof Note that, if η_i^* does not lie exactly on the boundaries of one of the intervals $\widetilde{I}_k$, the function $\widetilde{S}_i(\eta_i)$ is at least differentiable in a neighborhood of η_i^* (see Theorem 1). Therefore, both the matrices B^* and Λ^* are well defined.

Now, we show that if (16) admits a unique solution, then B^* is nonsingular. Assume by contradiction that B^* is singular. Then, there exists a direction $\mathbf{d} \neq \mathbf{0}$ such that

$$\mathsf{B}^*\mathbf{d} = (\mathsf{I} + (\mathsf{K} - \mathsf{D})\Lambda^*)\mathbf{d} = \mathbf{0}. \tag{30}$$

As shown below, for this to happen, a number of conditions must hold true:

1. The loss function $\widetilde{V}$ is non-differentiable.
2. The set I_N is nonempty.
3. The kernel $K(\cdot,\cdot)$ is not strictly positive.

With reference to the index sets I_N and I_D, introduce the following partitions:

$$\boldsymbol{\eta}^* = \begin{pmatrix} \boldsymbol{\eta}_D^* \\ \boldsymbol{\eta}_N^* \end{pmatrix}, \qquad \mathbf{d} = \begin{pmatrix} \mathbf{d}_D \\ \mathbf{d}_N \end{pmatrix}.$$

Note that B^* can be written as

$$\mathsf{B}^* = \begin{pmatrix} \mathsf{I}_D - \Lambda_D^*\mathsf{D}_D & 0 \\ 0 & 0 \end{pmatrix} + \mathsf{K}\Lambda^*.$$

In view of (30), we have:

$$\mathbf{d}^T\mathsf{B}^*\mathbf{d} = \mathbf{d}_D^T(\mathsf{I}_D - \Lambda_D^*\mathsf{D}_D)\mathbf{d}_D + \mathbf{d}^T\mathsf{K}\Lambda^*\mathbf{d} = 0.$$

Since $\mathsf{K}\Lambda^*$ is positive semi-definite while the matrix $(\mathsf{I}_D - \Lambda_D^*\mathsf{D}_D)$ is positive definite, it must be

$$\begin{cases} \mathbf{d}_D = 0, \\ \mathbf{d}^T\mathsf{K}\Lambda^*\mathbf{d} = 0. \end{cases} \tag{31}$$

From these equations we see that in order to find $\mathbf{d} \neq \mathbf{0}$ such that $\mathsf{B}^*\mathbf{d} = \mathbf{0}$ the set I_N must be nonempty. This means that the loss function $\widetilde{V}$ cannot be differentiable. Therefore, we assume that $\widetilde{V}$ is non-differentiable and that I_N is nonempty. From (31) we obtain

$$\mathbf{d}_N^T\mathsf{K}_{NN}\mathsf{D}_N^{-1}\mathbf{d}_N = 0.$$

Since D_N^{-1} is nonsingular, in order to satisfy this last equation with $\mathsf{D}_N \neq 0$, K_{NN} must be singular so that the kernel $K(\cdot,\cdot)$ cannot be strictly positive. Therefore, we suppose that the kernel is not strictly positive definite and that $\det(\mathsf{K}_{NN}) = 0$. Hence, any non-null direction $\mathbf{d}$ such that $\mathsf{B}^*\mathbf{d} = \mathbf{0}$ can be written as $\mathbf{d} = (\mathbf{0}^T\ \mathbf{d}_N^T)^T$, $\mathbf{d}_N \neq \mathbf{0}$.

Now, consider system (16), which, by assumption, admits a unique solution and can be rewritten as

$$\mathbf{F}(\boldsymbol{\eta}) = \mathbf{0}, \qquad \mathbf{F}(\boldsymbol{\eta}) := \eta - \mathbf{y} - (\mathsf{D} - \mathsf{K})\widetilde{\mathbf{S}}(\boldsymbol{\eta}).$$

Since each component of $\mathbf{F}$ is differentiable in a neighborhood of $\boldsymbol{\eta}^*$, we can choose $t > 0$ such that $\mathbf{F}$ is differentiable at $\boldsymbol{\eta}^* + t\mathbf{d}$. Locally, $\mathbf{F}$ is a linear operator along the direction $\mathbf{d}$ because the only components of $\boldsymbol{\eta}$ that vary belong to the $\widetilde{I}_k$ intervals (recall that $\mathbf{d}_D = \mathbf{0}$) within which all the functions $\widetilde{S}_i(y_i, \cdot)$ are affine, see (15). Thus, if t is small enough, $\mathbf{F}(\boldsymbol{\eta}^* + t\mathbf{d})$ coincides with its first order Taylor expansion around $\boldsymbol{\eta}^*$:

$$\mathbf{F}(\boldsymbol{\eta}^* + t\mathbf{d}) = \mathbf{F}(\boldsymbol{\eta}^*) + t\frac{\partial \mathbf{F}}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^*)\mathbf{d}.$$

Now, observe that the Jacobian $\frac{\partial \mathbf{F}}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^*)$ coincides with B^*:

$$\frac{\partial \mathbf{F}}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^*) = \mathsf{I} + (\mathsf{K} - \mathsf{D})\Lambda^* = \mathsf{B}^*.$$

Then, we have:

$$\mathbf{F}(\boldsymbol{\eta}^* + t\mathbf{d}) = \mathbf{F}(\boldsymbol{\eta}^*) + t\mathsf{B}^*\mathbf{d} = \mathbf{0}.$$

Therefore, $\boldsymbol{\eta}^* + t\mathbf{d}$ is a solution of $\mathbf{F}(\boldsymbol{\eta}) = \mathbf{0}$ different from $\boldsymbol{\eta}^*$, which contradicts the uniqueness assumption. Hence B^* is nonsingular.

Finally, we can write

$$\widehat{df} = \sum_{i=1}^{\ell} \frac{\partial \hat{f}(x_i)}{\partial y_i} = \mathrm{tr}\left[\frac{\partial}{\partial \mathbf{y}}(\mathsf{K}\widetilde{\mathbf{S}}(\boldsymbol{\eta}^*))\right] = \mathrm{tr}\left[\mathsf{K}\frac{\partial \widetilde{\mathbf{S}}(\boldsymbol{\eta}^*)}{\partial \mathbf{y}}\right] = \mathrm{tr}\left[\mathsf{K}\Lambda^*\frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}}\right].$$

From (16) we obtain

$$\frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}} = \mathsf{I} + (\mathsf{D} - \mathsf{K})\Lambda^*\frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}},$$

that is

$$\frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}} = (\mathsf{I} + (\mathsf{K} - \mathsf{D})\Lambda^*)^{-1} = (\mathsf{B}^*)^{-1},$$

which completes the proof. □

Next, we consider the case in which the model includes also a constant bias term. In the following, the symbol $\mathbf{1}$ denotes a vector of ones of proper dimension. In this case,

$$\eta_i = y_i + z_i = y_i - \sum_{j \neq i}^{\ell} a_j K(x_i, x_j) - b.$$

In the constant bias case, condition (3) reads $\sum_{i=1}^{\ell} a_i = 0$, that is

$$\mathbf{1}^T\widetilde{\mathbf{S}}(\boldsymbol{\eta}) = 0.$$

We have the following result.

Theorem 4 *Under the assumptions of Theorem 1, assume that system (16) with constant bias, given by*

$$\begin{cases} \boldsymbol{\eta} - \mathbf{y} - (\mathsf{D} - \mathsf{K})\widetilde{\mathbf{S}}(\eta) + b\mathbf{1} = \mathbf{0}, \\ \mathbf{1}^T\widetilde{\mathbf{S}}(\boldsymbol{\eta}) = 0, \end{cases} \tag{32}$$

admits a unique solution $(\boldsymbol{\eta}^*, b^*)$, *and that at least one of the following conditions is satisfied*:

1. *The loss function* $\widetilde{V}$ *is differentiable.*
2. *The kernel* $K(\cdot,\cdot)$ *is strictly positive.*

Suppose also that none of the coefficients η_i^* *lies on the boundaries of the intervals* $\widetilde{I}_k$ *defined in* (14). *Then, the unbiased estimate* (28) *of the degrees of freedom is given by*

$$\widehat{df} = \mathrm{tr}[\mathsf{K}\boldsymbol{\Lambda}^*(\mathsf{B}^*)^{-1}] + 1 - \frac{\mathbf{1}^T\boldsymbol{\Lambda}^*(\mathsf{B}^*)^{-1}\mathsf{K}\boldsymbol{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}},$$

where $\boldsymbol{\Lambda}^*$ *and* B^* *are as in Theorem* 3.

The proof of Theorem 4 is given in the appendix. Interestingly, introducing the bias term in the model does not entail, in general, a unitary increase of the degrees of freedom with respect to the no-bias case. Indeed, the increase is fractional and equal to

$$1 - \frac{\mathbf{1}^T\boldsymbol{\Lambda}^*(\mathsf{B}^*)^{-1}\mathsf{K}\boldsymbol{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}}.$$

In the following, Theorems 3 and 4 are used to derive the approximate degrees of freedom associated with some common loss functions.

Quadratic loss For the quadratic loss function analyzed in Example 1 of Sect. 5 we have

$$\boldsymbol{\Lambda}^* = C(\mathsf{I} + C\mathsf{D})^{-1}.$$

Then,

$$\begin{aligned}\widehat{df}^{\,Q} &= \mathrm{tr}\big[C\mathsf{K}(\mathsf{I} + C\mathsf{D})^{-1}(\mathsf{I} + (\mathsf{K} - \mathsf{D})C(\mathsf{I} + C\mathsf{D})^{-1})^{-1}\big] \\ &= \mathrm{tr}\big[C\mathsf{K}(\mathsf{I} + C\mathsf{D} + (\mathsf{K} - \mathsf{D})C(\mathsf{I} + C\mathsf{D})^{-1}(\mathsf{I} + C\mathsf{D}))^{-1}\big] = \mathrm{tr}\left[\mathsf{K}\left(\frac{\mathsf{I}}{C} + \mathsf{K}\right)^{-1}\right].\end{aligned}$$

Note that this is a very well known result (Wahba 1990). In fact, $\widehat{df}$ coincides with the trace of the so-called hat matrix and, due to linearity of the estimator, $df = \widehat{df}$. When also the bias term is introduced, we have

$$\widehat{df}_b^{\,Q} = \mathrm{tr}\left[\mathsf{K}\left(\frac{\mathsf{I}}{C} + \mathsf{K}\right)^{-1}\right] + 1 - \frac{\mathbf{1}^T(\frac{\mathsf{I}}{C} + \mathsf{K})^{-1}\mathsf{K}(\frac{\mathsf{I}}{C} + \mathsf{K})^{-1}\mathbf{1}}{\mathbf{1}^T(\frac{\mathsf{I}}{C} + \mathsf{K})^{-1}\mathbf{1}}.$$

Huber loss Consider the Huber loss function analyzed in Example 2 of Sect. 5. Let us introduce the following index sets:

$$I_Q = \{i \in I : |y_i - \hat{f}(x_i)| \le d\}, \qquad I_L = \{1, \ldots, \ell\} \setminus I_Q.$$

Note that I_Q is the index set associated with the data whose residuals fall in the quadratic trait of the Huber loss function, while I_L is the set relative to the data whose residuals fall in

the linear traits. Then, we have

$$\mathsf{K} = \begin{pmatrix} \mathsf{K}_{QQ} & \mathsf{K}_{QL} \\ \mathsf{K}_{LQ} & \mathsf{K}_{LL} \end{pmatrix}, \qquad \mathsf{D} = \begin{pmatrix} \mathsf{D}_Q & 0 \\ 0 & \mathsf{D}_L \end{pmatrix},$$

$$\Lambda^* = \begin{pmatrix} \Lambda_Q^* & 0 \\ 0 & 0 \end{pmatrix}, \qquad \Lambda_Q^* = \left(\frac{\mathsf{I}}{dC} + \mathsf{D}_Q \right)^{-1}.$$

It is easy to verify that

$$\mathsf{B}^* = \mathsf{I} + (\mathsf{K} - \mathsf{D})\Lambda^* = \begin{pmatrix} (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})(\frac{\mathsf{I}}{dC} + \mathsf{D}_Q)^{-1} & 0 \\ \mathsf{K}_{LQ}(\frac{\mathsf{I}}{dC} + \mathsf{D}_Q)^{-1} & \mathsf{I}_L \end{pmatrix},$$

$$(\mathsf{B}^*)^{-1} = \begin{pmatrix} (\frac{\mathsf{I}}{dC} + \mathsf{D}_Q)(\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} & 0 \\ -\mathsf{K}_{LQ}(\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} & \mathsf{I}_L \end{pmatrix}.$$

After some simplification, we obtain

$$\Lambda^*(\mathsf{B}^*)^{-1} = \begin{pmatrix} (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

In the no-bias case, we have

$$\widehat{df}^H = \mathrm{tr}[\mathsf{K}\Lambda^*(\mathsf{B}^*)^{-1}] = \mathrm{tr}\left[\mathsf{K}_{QQ} \left(\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ} \right)^{-1} \right],$$

while, if the bias is present,

$$\widehat{df}_b^H = \mathrm{tr}\left[\mathsf{K}_{QQ} \left(\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ} \right)^{-1} \right] + 1 - \frac{\mathbf{1}^T (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} \mathsf{K}_{QQ} (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} \mathbf{1}}{\mathbf{1}^T (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} \mathbf{1}}.$$

Soft ϵ-insensitive loss For the soft ϵ-insensitive loss function (Examples 4, 5) $\widehat{df}$ has expressions similar to that obtained for the Huber loss function:

$$\widehat{df}^S = \mathrm{tr}\left[\mathsf{K}_{QQ} \left(\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ} \right)^{-1} \right],$$

$$\widehat{df}_b^S = \mathrm{tr}\left[\mathsf{K}_{QQ} \left(\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ} \right)^{-1} \right] + 1 - \frac{\mathbf{1}^T (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} \mathsf{K}_{QQ} (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} \mathbf{1}}{\mathbf{1}^T (\frac{\mathsf{I}}{dC} + \mathsf{K}_{QQ})^{-1} \mathbf{1}}.$$

Again, the index set I_Q is associated with the data whose residuals fall in the quadratic traits of the loss function.

Vapnik ϵ-insensitive loss The degrees of freedom of the kernel regression estimator that uses the ϵ-insensitive loss function have been already discussed in (Gunter and Zhu 2007) and (Dinuzzo et al. 2007). When the kernel is strictly positive, the derivation of the result using the general formula (29) is straightforward. Rather interestingly, it can also be obtained via the following approximation argument.

Recall that the ϵ-insensitive loss can be obtained by taking the limit for $d \to +\infty$ in the soft insensitive loss function. Then, when $d \to +\infty$, the set of data whose residuals fall in the quadratic traits of the soft insensitive loss function reduce to the set of marginal support vectors. This means that there must exists $\bar{d}$ such that, for $d > \bar{d}$, the set I_Q does not change anymore. This final set must coincide with the set of indices I_M associated with the marginal support vectors (let m denote the set cardinality of I_M). Then, we can obtain $\widehat{df}$ taking the limits

$$\widehat{df}^E = \lim_{d\to+\infty} \widehat{df}^S(d), \qquad \widehat{df}_b^E = \lim_{d\to+\infty} \widehat{df}_b^S(d). \tag{33}$$

If the matrix K_{MM} is nonsingular, we obtain

$$\widehat{df}^E = \lim_{d\to+\infty} \operatorname{tr}\left[\mathsf{K}_{MM}\left(\frac{\mathsf{I}}{dC} + \mathsf{K}_{MM}\right)^{-1}\right] = \operatorname{tr}(\mathsf{I}_M) = m,$$

$$\widehat{df}_b^E = \lim_{d\to+\infty} \widehat{df}_b^S(d) = \operatorname{tr}(\mathsf{I}_M) + 1 - \frac{\mathbf{1}^T\mathsf{K}_{MM}^{-1}\mathbf{1}}{\mathbf{1}^T\mathsf{K}_{MM}^{-1}\mathbf{1}} = m,$$

which coincides with the results in (Gunter and Zhu 2007) and (Dinuzzo et al. 2007). Interestingly, the degrees of freedom are always equal to m irrespective of the presence of the bias term.

When K_{MM} is not full rank, some complications arise because Theorem 3 or 4 cannot be used. One possibility is to take (33) as the definition of $\widehat{df}$. Using the spectral decomposition

$$\mathsf{K}_{MM} = \mathsf{U}\Sigma\mathsf{U}^T, \quad \Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_m),$$

we can easily handle the no-bias case:

$$\widehat{df}^E = \lim_{d\to+\infty} \operatorname{tr}\left[\mathsf{U}\Sigma\mathsf{U}^T\left(\frac{\mathsf{I}}{dC} + \mathsf{U}\Sigma\mathsf{U}^T\right)^{-1}\right] = \lim_{d\to+\infty} \operatorname{tr}\left[\Sigma\left(\frac{\mathsf{I}}{dC} + \Sigma\right)^{-1}\right]$$

$$= \lim_{d\to+\infty} \sum_{i=1}^{m} \frac{\lambda_i}{\lambda_i + \frac{1}{dC}} = \lim_{d\to+\infty} \sum_{\lambda_i\neq 0} \frac{\lambda_i}{\lambda_i + \frac{1}{dC}} = \operatorname{rank}(\mathsf{K}_{MM}).$$

The calculation of $\widehat{df}_b^E$ is slightly more involved:

$$\widehat{df}_b^E = \operatorname{rank}(\mathsf{K}_{MM}) + 1 - \lim_{d\to+\infty} \frac{\mathbf{1}^T\mathsf{U}(\frac{\mathsf{I}}{dC} + \Sigma)^{-1}\Sigma(\frac{\mathsf{I}}{dC} + \Sigma)^{-1}\mathsf{U}^T\mathbf{1}}{\mathbf{1}^T\mathsf{U}(\frac{\mathsf{I}}{dC} + \Sigma)^{-1}\mathsf{U}^T\mathbf{1}}.$$

Introducing the vector $\mathbf{v} = \mathsf{U}^T\mathbf{1}$ we have

$$\lim_{d\to+\infty} \frac{\mathbf{1}^T\mathsf{U}(\frac{\mathsf{I}}{dC} + \Sigma)^{-1}\Sigma(\frac{\mathsf{I}}{dC} + \Sigma)^{-1}\mathsf{U}^T\mathbf{1}}{\mathbf{1}^T\mathsf{U}(\frac{\mathsf{I}}{dC} + \Sigma)^{-1}\mathsf{U}^T\mathbf{1}} = \lim_{d\to+\infty} \frac{\mathbf{v}^T(\frac{\mathsf{I}}{dC} + \Sigma)^{-2}\Sigma\mathbf{v}}{\mathbf{v}^T(\frac{\mathsf{I}}{dC} + \Sigma)^{-1}\mathbf{v}}$$

$$= \lim_{d\to+\infty} \frac{\sum_{i=1}^m \frac{v_i^2\lambda_i}{(\lambda_i+\frac{1}{dC})^2}}{\sum_{i=1}^m \frac{v_i^2}{(\lambda_i+\frac{1}{dC})}} = \lim_{d\to+\infty} \frac{\sum_{\lambda_i\neq 0} \frac{v_i^2\lambda_i}{(\lambda_i+\frac{1}{dC})^2}}{\sum_{\lambda_i\neq 0} \frac{v_i^2}{(\lambda_i+\frac{1}{dC})} + dC\sum_{\lambda_i=0} v_i^2}. \tag{34}$$

Now, that there are two cases. If

$$\sum_{\lambda_i=0} v_i^2 = 0, \tag{35}$$

the limit (34) equals 1. Conversely, if

$$\sum_{\lambda_i=0} v_i^2 > 0,$$

then the limit approaches zero. It can be easily seen that the condition (35) means that the vector $\mathbf{1}$ lies in the range of K_{MM}. Thus, denoting by $I(\cdot)$ the indicator function, we have

$$\widehat{df}_b^E = \operatorname{rank}(\mathsf{K}_{MM}) + I(\mathbf{1} \notin \operatorname{range}(\mathsf{K}_{MM})).$$

7 Conclusions

An algebraic characterization of the optimum associated with regularized kernel methods has been derived. In the case of non-smooth losses, the new characterization avoids the use of inclusions and is formulated as a system of algebraic equations.

Two consequences of this main result have been investigated. First, it has been shown that the original non-smooth problem can be given an unconstrained smooth reformulation. Second, the algebraic characterization has been exploited to derive a general formula for the degrees of freedom of kernel regression methods in the context of SURE (Stein's Unbiased Risk Estimation).

The present paper substantially improves and generalizes the results reported in Dinuzzo et al. (2007). Herein, rather than focusing on SVR, generic convex loss functions are considered, thus extending the algebraic characterization to virtually all kernel-based classification and regression methods. Furthermore, the presence of a parametric bias term is taken into account and non-strictly positive kernels are allowed. Of particular interest is the new derivation of explicit formulas for the degrees of freedom of any kernel regression estimator. Another novelty is the result on the smooth unconstrained reformulation.

Among possible future developments, we may mention two main directions of research. The first is extending the algebraic characterization in order to include also parametric models in a general framework. The second one deals with investigating the potential computational benefits of the smooth reformulation of the learning problem. In particular, note that the fixed-point equation (4)

$$a_i = S_i(y_i, z_i(\mathbf{a}, \mathbf{b})),$$

can be used to obtain a generalized version of an SMO update for generic convex loss functions. SMO-based algorithms are regarded as state-of-art algorithms for SVM computation. The combination of (4) with a suitable working selection strategy would lead to generalized SMO algorithms to be applied to a much larger class of regularization methods.

Acknowledgements This research has been partially supported by the Italian Ministry of University and Research through the FIRB Project "Learning Theory and Application".

Appendix

Proof of Theorem 1 By definition, the coefficients z_i are such that

$$a_i = \frac{1}{\delta} \frac{\hat{f}(x_i; \mathbf{a}, \mathbf{b}) + z_i(\mathbf{a}, \mathbf{b})}{K(x_i, x_i)}. \tag{36}$$

Now, we consider two cases.

First, when $\hat{f}(x_i) \neq \gamma_k$, $k = 1, \ldots, n_\gamma^i$, $V(y_i, \cdot)$ is twice differentiable at $\hat{f}(x_i)$ and its sub-differential is single-valued so that (2) yields (7). Now, using the Implicit Function Theorem, it is shown that, locally, a_i is a monotone nondecreasing Lipschitz continuous function of z_i. In fact, by taking the derivative of (7) with respect to z_i,

$$\frac{\partial a_i}{\partial z_i} = \frac{C\partial_2^2 V(y_i, a_i K(x_i, x_i)\delta - z_i)}{1 + \delta C K(x_i, x_i)\partial_2^2 V(y_i, a_i K(x_i, x_i)\delta - z_i)}.$$

The denominator is always different from zero because, by convexity, $\partial_2^2 V \geq 0$ whenever it exists. Therefore, locally, a_i is a differentiable function of z_i:

$$a_i = \overline{S}(y_i, z_i).$$

The function $\overline{S}(y_i, \cdot)$ is monotone nondecreasing and has bounded derivative because

$$0 \leq \frac{\partial a_i}{\partial z_i} < \frac{1}{\delta} \frac{1}{K(x_i, x_i)}. \tag{37}$$

Now, consider the second case. When $\hat{f}(x_i; \mathbf{a}, \mathbf{b}) = \gamma_k$ for some k, $V(y_i, \cdot)$ is not twice differentiable at $\hat{f}(x_i; \mathbf{a}, \mathbf{b})$. Then, (36) yields (6), so that a_i is an affine function of z_i. Recalling the properties of the sub-differential of a convex function, from (2) we have

$$a_i \in \left[-CD^+(\gamma_k), -CD^-(\gamma_k)\right],$$

so that

$$\begin{aligned} z_i \in I_k &:= [z_k^L, z_k^R], \\ z_k^L &:= -\big(\gamma_k + \delta C K(x_i, x_i) D^+(\gamma_k)\big), \\ z_k^R &:= -\big(\gamma_k + \delta C K(x_i, x_i) D^-(\gamma_k)\big). \end{aligned}$$

Finally, (6) implies

$$\frac{\partial a_i}{\partial z_i} = \frac{1}{\delta} \frac{1}{K(x_i, x_i)} > 0,$$

so that functions $S_i(y_i, \cdot)$ are locally monotone nondecreasing also in the second case. Combining this last inequality with the bound (37) that holds in differentiability points, we conclude that the derivative of $S_i(y_i, \cdot)$ is bounded everywhere, possibly except for discontinuity points. Hence, in order to prove Lipschitz continuity it suffices to show that $S_i(y_i, \cdot)$ is continuous. Therefore, we conclude the proof by showing that the set of discontinuity points is actually empty. In this respect, the only points that must be analyzed are the boundaries of the intervals I_k. In fact, in the interior of I_k, $S_i(y_i, \cdot)$ is infinitely differentiable because it is affine, while, outside, it has the same regularity of $\partial_2 V(y_i, \cdot)$, by the Implicit Function Theorem. Hence, it suffices to prove continuity at the left boundary z_k^L of I_k. Consider (6) and take the limit from the right (see also Fig. 5):

$$\lim_{z_i \to (z_k^L)^+} S_i(y_i, z_i)\big|_{z_i \in I_k} = \lim_{z_i \to (z_k^L)^+} \frac{1}{\delta} \frac{\gamma_k + z_i}{K(x_i, x_i)} = -CD^+(\gamma_k).$$

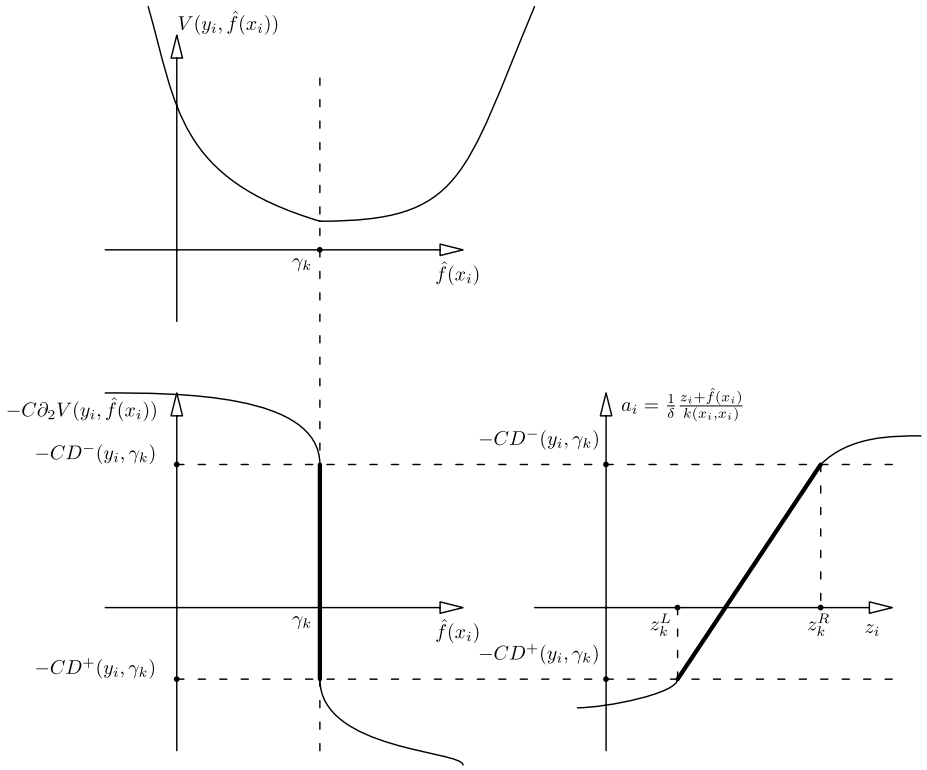

Fig. 5 The construction of Theorem 1

Now, observe that, if a_i tends to $-CD^+(\gamma_k)$ from below, then z_i tends to z_k^L from the left. Indeed, if $\hat{f}(x_i) = a_i K(x_i, x_i)\delta - z_i \neq \gamma_k$, it results that

$$a_i = -CD^+(a_i K(x_i, x_i)\delta - z_i).$$

Then,

$$\begin{aligned} -CD^+(\gamma_k) &= \lim_{a_i \to -CD^+(\gamma_k)} a_i = \lim_{a_i \to -CD^+(\gamma_k)} -CD^+(a_i K(x_i, x_i)\delta - z_i) \\ &= -CD^+(-CD^+(\gamma_k)K(x_i, x_i)\delta - \bar{z}), \end{aligned}$$

where

$$\bar{z} = \lim_{a_i \to -CD^+(\gamma_k)} z_i(a_i).$$

By comparing the first and the last term of the previous chain of equalities, it follows that

$$\gamma_k = -CD^+(\gamma_k)K(x_i, x_i)\delta - \bar{z},$$

so that $\bar{z} = z_k^L$. □

Proof of Theorem 4 Following the first part of the proof of Theorem 3, we obtain that Λ^* is well defined and B^* is nonsingular. Moreover, it turns out that $\Lambda^* \neq 0$. In fact, suppose by contradiction that $\Lambda^* = 0$. Note that, by definition of Λ^*, this can only happen when the set I_N (defined in Sect. 6) is empty. Moreover, since

$$\frac{\partial \widetilde{S}_i}{\partial \eta_i}(\eta_i^*) = 0, \quad \forall i = 1, \ldots, \ell,$$

we have that $\widetilde{\mathbf{S}}(\boldsymbol{\eta})$ is constant in a neighborhood of $\boldsymbol{\eta}^*$. Thus, taking any direction $\mathbf{d} \in \mathbb{R}^\ell$, there exists a positive number t such that

$$\widetilde{\mathbf{S}}(\boldsymbol{\eta}^* + t\mathbf{d}) = \widetilde{\mathbf{S}}(\eta^*).$$

In particular, we can choose $\mathbf{d} = \mathbf{1}$ and find a corresponding t. Then, it is immediate to see that the pair $(\eta^* + t\mathbf{1}, b^* - t)$ is a solution of (32) different from $(\boldsymbol{\eta}^*, b^*)$. In fact,

$$\begin{cases} \boldsymbol{\eta}^* + t\mathbf{1} - \mathbf{y} - (\mathsf{D} - \mathsf{K})\widetilde{\mathbf{S}}(\boldsymbol{\eta}^* + t\mathbf{1}) + (b^* - t)\mathbf{1} = \mathbf{0}, \\ \mathbf{1}^T \widetilde{\mathbf{S}}(\boldsymbol{\eta}^* + t\mathbf{1}) = \mathbf{1}^T \widetilde{\mathbf{S}}(\boldsymbol{\eta}^*) = 0. \end{cases}$$

Since this contradicts the uniqueness hypothesis, it must be $\Lambda^* \neq 0$. Now, we can take the derivative with respect to $\mathbf{y}$ in system (32) obtaining:

$$\begin{cases} \mathsf{B}^* \frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}} + \mathbf{1} \frac{\partial b^*}{\partial \mathbf{y}} = \mathsf{I}, \\ \mathbf{1}^T \Lambda^* \frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}} = 0. \end{cases}$$

The first equation can be solved for $\frac{\partial \eta^*}{\partial \mathbf{y}}$.

$$\begin{cases} \frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}} = (\mathsf{B}^*)^{-1}(\mathsf{I} - \mathbf{1} \frac{\partial b^*}{\partial \mathbf{y}}), \\ \mathbf{1}^T \Lambda^* (\mathsf{B}^*)^{-1}(\mathsf{I} - \mathbf{1} \frac{\partial b^*}{\partial \mathbf{y}}) = 0. \end{cases}$$

Observe that $\mathbf{1}^T \Lambda^* (\mathsf{B}^*)^{-1} \mathbf{1} \neq 0$. In fact, if this were not the case, from the second equation we would obtain

$$\mathbf{1}^T \Lambda^* (\mathsf{B}^*)^{-1} = \mathbf{0}^T,$$

that, by non-singularity of B^*, implies $\Lambda^* \mathbf{1} = \mathbf{0}$. Since Λ^* is a diagonal matrix, this would imply $\Lambda^* = 0$, which, as seen before, is not possible. Now, we can solve for $\frac{\partial b^*}{\partial \mathbf{y}}$ in the second equation obtaining

$$\frac{\partial b^*}{\partial \mathbf{y}} = \frac{\mathbf{1}^T \Lambda^* (\mathsf{B}^*)^{-1}}{\mathbf{1}^T \Lambda^* (\mathsf{B}^*)^{-1} \mathbf{1}},$$

which gives also

$$\frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}} = (\mathsf{B}^*)^{-1} - \frac{(\mathsf{B}^*)^{-1} \mathbf{1} \mathbf{1}^T \Lambda^* (\mathsf{B}^*)^{-1}}{\mathbf{1}^T \Lambda^* (\mathsf{B}^*)^{-1} \mathbf{1}}.$$

Now, it is easy to obtain $\widehat{df}$:

$$\widehat{df} = \sum_{i=1}^{\ell} \frac{\partial \hat{f}(x_i)}{\partial y_i} = \operatorname{tr}\left[\frac{\partial}{\partial \mathbf{y}} (\mathsf{K}\widetilde{\mathbf{S}}(\boldsymbol{\eta}^*) + b\mathbf{1}) \right] = \operatorname{tr}\left[\mathsf{K}\Lambda^* \frac{\partial \boldsymbol{\eta}^*}{\partial \mathbf{y}} \right] + \operatorname{tr}\left[\mathbf{1} \frac{\partial b^*}{\partial \mathbf{y}} \right]$$

$$= \operatorname{tr}\left[\mathsf{K}\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1} - \frac{\mathsf{K}\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}\mathbf{1}^T\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}}{\mathbf{1}^T\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}}\right] + \operatorname{tr}\left[\mathbf{1}\frac{\mathbf{1}^T\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}}{\mathbf{1}^T\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}}\right]$$

$$= \operatorname{tr}\left[\mathsf{K}\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}\right] - \frac{\mathbf{1}^T\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}\mathsf{K}\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}}{\mathbf{1}^T\mathsf{\Lambda}^*(\mathsf{B}^*)^{-1}\mathbf{1}} + 1. \qquad \square$$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second international symposium on information theory*. Budapest: Académiai Kiadó.

Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Computation*, *19*(5), 1155–1178.

Chu, W., Keerthi, S. S., & Ong, C. J. (2001). A unified loss function in Bayesian framework for support vector regression. In *Proceedings of the 18th international conference on machine learning* (pp. 51–58). San Francisco: Morgan Kaufmann.

Cox, D., & O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics*, *18*, 1676–1695.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*, 377–403.

De Vito, E., Rosasco, L., Caponnetto, A., Piana, M., & Verri, A. (2004). Some properties of regularized kernel methods. *Journal of Machine Learning Research*, *5*, 1363–1390.

Dekel, O., Shalev-Shwartz, S., & Singer, Y. (2005). Smooth ϵ-insensitive regression by loss symmetrization. *Journal of Machine Learning Research*, *6*, 711–741.

Dinuzzo, F., Neve, M., De Nicolao, G., & Gianazza, U. P. (2007). On the representer theorem and equivalent degrees of freedom of SVR. *Journal of Machine Learning Research*, *8*, 2467–2495.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, *99*(14), 619–632.

Fukushima, M., & Qi, L. (1999). *Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods*. Dordrecht: Kluwer Academic.

Gunter, L., & Zhu, J. (2007). Efficient computation and model selection for the support vector regression. *Neural Computation*, *19*, 1633–1655.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. (2001). *The elements of statistical learning. Data mining, inference and prediction*. Canada: Springer.

Kimeldorf, G., & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, *33*(1), 82–95.

Lee, Y. J., & Mangasarian, O. L. (2001). SSVM: A smooth support vector machine for classification. *Computational Optimization and Applications*, *20*, 5–22.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675.

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

Osborne, M. R. (2001). *Simplicial algorithms for minimizing polyhedral functions*. Cambridge: Cambridge University Press.

Perez-Cruz, F., Bousono-Calzon, C., & Artes-Rodriguez, A. (2005). Convergence of the IRWLS procedure to the support vector machine solution. *Neural Computation*, *17*(1), 7–18.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—support vector learning*. Cambridge: MIT Press.

Poggio, T., & Girosi, F. (1992). A theory of networks for approximation and learning. In *Foundation of neural networks* (pp. 91–106).

Pontil, M., & Verri, A. (1998). Properties of support vector machines. *Neural Computation*, *10*, 955–974.

Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. *Neural Networks and Computational Learning Theory*, *81*, 416–426.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning*. Cambridge: MIT Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007). PEGASOS: primal estimated sub-gradient solver for svm. In *ICML '07: proceedings of the 24th international conference on machine learning* (pp. 807–814). New York: ACM.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, *9*, 1135–1151.
Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research*, *4*, 1071–1105.
Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.
Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.
Wahba, G. (1998). *Support vector machines, reproducing kernel Hilbert spaces and randomized GACV* (Tech. Rep. 984). Department of Statistics, University of Wisconsin.
Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, *93*, 120–131.
Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, *35*(5), 2173–2192.