ϵ-*Sparse Representations*: Generalized Sparse Approximation and the Equivalent Family of SVM Tasks*

Zoltán Szabó† András Lőrincz‡

Abstract

Relation between a family of generalized Support Vector Machine (SVM) problems and the novel ϵ-sparse representation is provided. In defining ϵ-sparse representations, we use a natural generalization of the classical ϵ-insensitive cost function for vectors. The insensitive parameter of the SVM problem is transformed into component-wise insensitivity and thus overall sparsification is replaced by component-wise sparsification. The connection between these two problems is built through the generalized Moore-Penrose inverse of the Gram matrix associated to the kernel.

*`http://www.inf.u-szeged.hu/actacybernetica/edb/vol17n3/pdf/Lorincz_2006_ActaCybernetica.pdf`, Acta Cybernetica 17(3):605-614, 2006.

†Department of Information Systems, Eötvös Loránd University, H-1117 Budapest, Hungary; e-mail: szzoli@cs.elte.hu

‡Corresponding author; Department of Information Systems, Eötvös Loránd University, H-1117 Budapest, Hungary; e-mail: andras.lorincz@elte.hu

1 Introduction

Girosi [3] has shown the equivalence of the classic Support Vector Machine (SVM) regression and the sparse approximation scheme [6], similar to the Basis Pursuit De-Noising algorithm [2] under the assumption of noiseless observation. The novelty of the approach is that the approximation is introduced directly in the Reproducing Kernel Hilbert Space (RKHS) and thus it avoids the empirical estimation of the estimation error. Equivalence is understood in the sense that the two optimization problems give rise to the same Quadratic Programming (QP) task.

Equivalence can be shown similarly to [3], but under the condition of noisy observation for linear and quadratic ϵ-insensitive SVM approximation costs [5]. The noise process was included into an extended RKHS. In both cases, however, the ϵ of the approximation cost is transformed onto the scalar multiplier of the parameter vector, which determines the linear combination in the approximation. We ask (i) if it is possible to embed the insensitivity parameter into a constraint on the searched representation, i.e, directly into the cost function, and (ii) if there is an extension of the SVM problems characterized by pair (C, ϵ) (where C is the multiplier of the ϵ-insensitive cost term of the cost function [12, 3]) to more general problems favoring sparse coding.

The paper is constructed as follows: Section 2 is about the notations and definitions used throughout this work. In Section 3 we sketch earlier correspondences between sparse coding and SVM. Section 4 defines the two generalized problem classes, ϵ-sparse problem class and the corresponding SVM problem class. These classes will be transformed onto each other in this section. Conclusions are drawn in Section 5.

2 Notations and Basic Concepts

For the sake of clarity, our notations and the basic concepts are provided.

2.1 Letter Types, Number Sets

Numbers (b), vectors[1] ($\mathbf{b}$), and matrices ($\mathbf{B}$) are distinguished from each other by letter types. Natural number sets are represented by $\mathbb{N}$, that is,

[1] Vector means column vector.

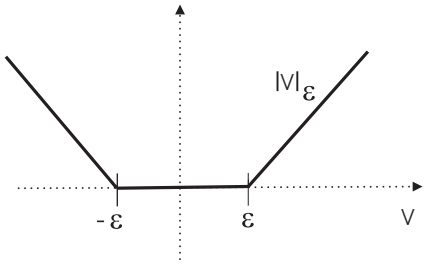

Figure 1: Vapnik's $|v|_\epsilon$ ϵ-insensitive cost function. One may think of this cost function that it represents a resolution not better than ϵ and errors smaller than ϵ are not detected and give rise to no cost. Errors larger than ϵ are, however, detected and – for mathematical tractability – make linear contributions to the cost function.

$\mathbb{N} := \{0, 1, 2, \dots\}$, whereas $\mathbb{R}$ stands for real numbers. Subsets restricted for positive values are indicated by + sign, e.g., $\mathbb{N}^+$ and $\mathbb{R}^+$.

2.2 Vectors and Matrices

Relations concerning vectors (e.g.,: $\geq$) are to be meant for each coordinate separately. The i^{th} component of vector $\mathbf{v}$ is denoted by v_i, the ij^{th} component of matrix $\mathbf{V}$ by $V_{i,j}$. ϵ-insensitive cost of vectors is defined as

$$\|\mathbf{v}\|_{\mathbf{r}} := \sum_i |v_i|_{r_i},$$

where $|v|_r := \{0, \text{ if } |v| \leq r; |v| - r, \text{ otherwise}\}$ is the usual 'ϵ-insensitive' cost function[2], which is shown in Fig. 1.

Operations $\mathbf{v}^T$, $\mathbf{v} \circ \mathbf{z}$, and $\mathbf{V} \otimes \mathbf{Z}$ represent transposition, multiplication by elements and the Kronecker product, respectively. Symbol $\mathbf{1}$ has special meaning, it represents a vector having only 1s, i.e., $\mathbf{1} := [1, \dots, 1]^T$.

The *Moore-Penrose generalized inverse* of matrix $\mathbf{G} \in \mathbb{R}^{n \times m}$ is a unique matrix $\mathbf{G}^- \in \mathbb{R}^{m \times n}$, which has the following features:

$$\mathbf{G}\mathbf{G}^-, \mathbf{G}^-\mathbf{G} \quad : \quad \text{symmetric matrices} \tag{1}$$

$$\mathbf{G}\mathbf{G}^-\mathbf{G} \quad = \quad \mathbf{G} \tag{2}$$

$$\mathbf{G}^-\mathbf{G}\mathbf{G}^- \quad = \quad \mathbf{G}^-. \tag{3}$$

[2] Notice that $\mathbf{r} \equiv \mathbf{0}$ gives rise to the L_1 norm for vectors.

2.3 RKHS, Feature Mapping, Gram Matrix

Here, we review some basic properties of Reproducing Kernel Hilbert Spaces (RKHS), necessary for our considerations. For further details, the interested reader is referred to the literature [12, 11, 1, 4].

An RKHS is denoted by $\mathcal{H}$. We shall select functions from this space to approximate sample points $\{\mathbf{x}_i, y_i\}_{i=1..l}$, where $\mathbf{x}_i \in \mathcal{X}$ form the *input space* and $y_i \in \mathbb{R}$ (see, e.g, [7]). In space $\mathcal{H}$, the scalar product is computed by means of *kernel* k. Kernel k is also used to define the basic functions of the RKHS: $\phi(\mathbf{x}) := k(\cdot, \mathbf{x}) : \mathcal{X} \to \mathcal{H}$. Such functions are called *feature mappings* and function $\phi(\mathbf{x})$ is interpreted as the *representation* of $\mathbf{x}$ in space $\mathcal{H}$. Now, the scalar product of feature mappings is defined as

$$\langle \phi(\mathbf{s}), \phi(\mathbf{t}) \rangle_{\mathcal{H}} = \langle k(\cdot, \mathbf{s}), k(\cdot, \mathbf{t}) \rangle_{\mathcal{H}} = k(\mathbf{s}, \mathbf{t}) \quad (\mathbf{s}, \mathbf{t} \in \mathcal{X}). \tag{4}$$

It can be shown that the kernel satisfies the following *reproducing property*

$$\langle f(\cdot), k(\cdot, \mathbf{t}) \rangle_{\mathcal{H}} = f(\mathbf{t}) \quad (\mathbf{t} \in \mathcal{X}, \forall f \in \mathcal{H}). \tag{5}$$

This means that $k(\cdot, \mathbf{t})$ can be seen as the *evaluation functional* at position $\mathbf{t}$ of space $\mathcal{H}$. The *Gram matrix of* k defined by $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ l-tuples assumes the following form

$$\mathbf{G} := [G_{i,j}]_{i,j=1\ldots l} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1\ldots l}. \tag{6}$$

2.4 SVM

Function approximation based on sparse data is often hard and is typically ill-posed [4]: existence, uniqueness and stability conditions may not be met in these cases. Regularization theory [10] can be of help under these conditions. To solve such problems, Vapnik, in his pioneering works, formulated the Support Vector Machine (SVM) problem family [12, 11]. In the SVM problem, the approximating functions are searched in the form

$$f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b, \tag{7}$$

subject to ϵ-insensitive cost function

$$V(u, z) = |u - z|_{\epsilon}, \tag{8}$$

and with regularizer [10] of the form $\|\mathbf{w}\|_{\mathcal{H}}^2$ with norm $\|\cdot\|_{\mathcal{H}}$ defined by kernel k of RKHS $\mathcal{H} = \mathcal{H}(k)$. Then the SVM task is as follows:

$$\min_{\mathbf{w},b} H[\mathbf{w}, b] := C \sum_{i=1}^{l} |y_i - f_{\mathbf{w},b}(\mathbf{x}_i)|_\epsilon + \frac{1}{2} \|\mathbf{w}\|_{\mathcal{H}}^2 \quad (C > 0). \tag{9}$$

Optimization of Eq. (9) can be executed, e.g., by solving a Quadratic Programming (QP) task formulated in the dual space

$$\min_{\mathbf{d}^*,\mathbf{d}} \left[\frac{1}{2} (\mathbf{d}^* - \mathbf{d})^T \mathbf{G} (\mathbf{d}^* - \mathbf{d}) - (\mathbf{d}^* - \mathbf{d})^T \mathbf{y} + (\mathbf{d}^* + \mathbf{d})^T \epsilon \mathbf{1} \right] \tag{10}$$

$$\text{provided that} \left\{ \begin{array}{rcl} C\mathbf{1} \geq \mathbf{d}^*, \mathbf{d} & \geq & \mathbf{0} \\ (\mathbf{d}^* - \mathbf{d})^T \mathbf{1} & = & 0 \end{array} \right\}.$$

For the derivation, see, e.g., [9]. Here, matrix $\mathbf{G}$ is the Gram matrix introduced before.

3 Previous Results

3.1 Noiseless Case

Starting from the work [2] Girosi has formulated a modified sparse approximation task in RKHS [3]:

$$\min_{\mathbf{a}} \left[\frac{1}{2} \left\| f(\cdot) - \sum_{i=1}^{l} a_i k(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}}^2 + \epsilon \|\mathbf{a}\|_1 \right]. \tag{11}$$

The first term is about quadratic approximation but instead of $\mathbb{R}$ it is formulated through the norm $\|\cdot\|_{\mathcal{H}}^2$ on Hilbert space $\mathcal{H}$. The second term is the sparse constraint, or sparsifying cost term. Girosi has shown that Eq. (11) is equivalent to the SVM task of Eq. (9) provided that

1. objective f is in $\mathcal{H}$ and that $\langle f, 1 \rangle_{\mathcal{H}} = 0$,[3]
2. data are noise-free, that is $f(\mathbf{x}_i) = y_i$ $(i = 1, \ldots, l)$,

[3] This restriction gives rise to constraint $\sum_i a_i = 0$.

3. $C \to \infty$.

Equivalence is to be understood in the sense that by breaking the searched vector $\mathbf{a}$ into positive and negative parts, such as

$$\mathbf{a} = \mathbf{a}^+ - \mathbf{a}^-, \text{ where } \mathbf{a}^+, \mathbf{a}^- \geq \mathbf{0}, \text{ and } \mathbf{a}^+ \circ \mathbf{a}^- = \mathbf{0} \tag{12}$$

then the task for pair $(\mathbf{a}^+, \mathbf{a}^-)$ is identical to the optimal solution $(\mathbf{d}^*, \mathbf{d})$ for Eq. (10) in the dual QP space.

3.2 The Noisy Case

The solution was extended to the noise case [5]: the connection was formulated for the regression problem and for linear and quadratic ϵ-insensitive SVM approximation. The equivalence is based on a larger RKHS space, which encapsulates the noise process, too. For detailed description and for other similar equivalences, the interested reader is referred to the original work [5].

In the cited cases [3, 5], the insensitive parameter (ϵ) was transformed into the multiplier of the sparsifying cost term (compare, e.g., Eq. (9) and Eq. (11)). Our question is if the constant multiplier of the ϵ-insensitivity loss can be transformed directly into the different components of the loss function by generalizing uniform sparsification to a component-wise sparsification problem.

For notational simplicity, instead of approximating in semi-parametric form (e.g., $f + b$, where $f \in \mathcal{H}$), we shall deal with the so called non-parametric scheme [8] ($f \in \mathcal{H}$). This approach is well grounded by the representer theorem [8].

4 Generalized Problems

In this section we shall introduce the generalizations of the previous SVM and sparse tasks and we shall show that they are equivalent. Given this equivalence, the two problem family will be referred jointly as *ϵ-sparse representations*.

4.1 The $(\mathbf{c}, \mathbf{e})$-SVM Task

Below, we introduce an SVM task family, which can be connected to regularization theory and satisfies the conditions of the representer theorem [8]. The usual SVM task – Eq. (9) – is modified as follows:

1. We shall approximate in the form $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{\mathcal{H}}$. The representer theorem warrants that it is satisfactory to approximate in this special form from $\mathcal{H}$.

2. We shall use approximation errors that may differ for each sample point.

3. We shall use weights that may differ for each sample point.

Introducing vector $\mathbf{e}$ for the ϵ-insensitive costs and $\mathbf{c}$ for the weights, respectively, the generalized problem has the following form:

$$\min_{\mathbf{w}} \left[\sum_{i=1}^{l} c_i \left| y_i - f_{\mathbf{w}}(\mathbf{x}_i) \right|_{e_i} + \frac{1}{2} \left\| \mathbf{w} \right\|_{\mathcal{H}}^2 \right] \quad (\mathbf{c} > \mathbf{0}, \mathbf{e} \geq \mathbf{0}). \tag{13}$$

This task shall be called the $(\mathbf{c}, \mathbf{e})$-*SVM task*. The original task of Eq. (9) corresponds to the particular choice of $((C, \epsilon) \otimes \mathbf{1})$ and $b = 0$. Alike to the original SVM task, the new $(\mathbf{c}, \mathbf{e})$-SVM task also has its quadratic equivalent in the dual space, which is as follows

$$\min_{\mathbf{d}^*, \mathbf{d}} \left[\frac{1}{2} \left(\mathbf{d}^* - \mathbf{d} \right)^T \mathbf{G} \left(\mathbf{d}^* - \mathbf{d} \right) - \left(\mathbf{d}^* - \mathbf{d} \right)^T \mathbf{y} + \left(\mathbf{d}^* + \mathbf{d} \right)^T \mathbf{e} \right], \tag{14}$$
$$\text{provided that } \{ \ \mathbf{c} \geq \mathbf{d}^*, \mathbf{d} \ \geq \ \mathbf{0} \ \},$$

where $\mathbf{G}$ denotes the Gram matrix of kernel k that belongs to points $\mathbf{x}_i$.

4.2 The $(\mathbf{p}, \mathbf{s})$-Sparse Task

Let us consider the optimization problem

$$\min_{\mathbf{a}} F[\mathbf{a}] := \frac{1}{2} \left\| f(\cdot) - \sum_{i=1}^{l} a_i k(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}}^2 + \sum_{i=1}^{l} p_i \left| a_i \right|_{s_i} \quad (\mathbf{p} > 0, \mathbf{s} \geq 0) \tag{15}$$

on sample points $\{\mathbf{x}_i, y_i\}_{i=1..l}$ that intends to approximate objective function $f \in \mathcal{H}(k)$. This problem shall be referred to as $\mathbf{p}$-*weighted and* $\mathbf{s}$-*sparse task*, or $(\mathbf{p}, \mathbf{s})$-*sparse task*, for short. The particular choice of $((\epsilon, 0) \otimes \mathbf{1})$ recovers the sparse representation form of Eq. (11).

4.3 Correspondence Between the Tasks

The tasks defined by Eq. (13) and Eq. (15), respectively will be connected to each other by means of the following theorem:

Theorem 1. *Let $\mathcal{X}$ denote and arbitrary non-empty set, k be a kernel on $\mathcal{X}$, $\{\mathbf{x}_i, y_i\}_{i=1..l}$ a sample set of l elements, where $\mathbf{x}_i \in \mathcal{X}, y_i \in \mathbb{R}$. Assuming that the values of RKHS objective $f \in \mathcal{H} = \mathcal{H}(k)$ can be observed in points $\mathbf{x}_i$ ($f(\mathbf{x}_i) = y_i$), then under the approximation*

$$f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{\mathcal{H}}$$

the dual problems of the

$$\min_{\mathbf{w}} \left[\sum_{i=1}^{l} c_i \left| y_i - f_{\mathbf{w}}(\mathbf{x}_i) \right|_{e_i} + \frac{1}{2} \left\| \mathbf{w} \right\|_{\mathcal{H}}^2 \right] \quad (\mathbf{c} > \mathbf{0}, \mathbf{e} \geq 0)$$

$(\mathbf{c}, \mathbf{e})$-SVM task and that of

$$\min_{\mathbf{a}} \left[\frac{1}{2} \left\| f(\cdot) - \sum_{i=1}^{l} a_i k(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}}^2 + \sum_{i=1}^{l} p_i \left| a_i \right|_{s_i} \right] \quad (\mathbf{p} > 0, \mathbf{s} \geq 0)$$

the $(\mathbf{p}, \mathbf{s})$-sparse task can be transformed onto each other through the generalized inverse $\mathbf{G}^-$ of Gram matrix

$$\mathbf{G} := [G_{i,j}]_{i,j=1\ldots l} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1\ldots l},$$

or, shortly,

$$\mathit{Dual}\left[(\mathbf{c}, \mathbf{e})\text{–}\mathit{SVM}\right] \overset{\mathbf{G}^-}{\longleftrightarrow} \mathit{Dual}\left[(\mathbf{p}, \mathbf{s})\text{–}\mathit{sparse}\right]$$

under correspondence

$$(\mathbf{d}^*, \mathbf{d}, \mathbf{G}, \mathbf{y}) \leftrightarrow (\mathbf{d}^+, \mathbf{d}^-, \mathbf{G}^-\mathbf{G}\mathbf{G}^-, \mathbf{G}^-\mathbf{y}) = (\mathbf{d}^+, \mathbf{d}^-, \mathbf{G}^-, \mathbf{G}^-\mathbf{y}).$$

Proof. *We shall modify Eq. (15) under the assumption of $f(\mathbf{x}_i) = y_i \quad (i = 1, \ldots, l)$. Given that norm $\|\cdot\|_{\mathcal{H}}^2$ is induced by a scalar product on $\mathcal{H}$, and utilizing the bilinear property of scalar products, we have*

$$\begin{aligned} F[\mathbf{a}] \;\; = \;\; & \frac{1}{2} \left\| f \right\|_{\mathcal{H}}^2 - \sum_i a_i \left\langle f(\cdot), k(\cdot, \mathbf{x}_i) \right\rangle_{\mathcal{H}} + \\ & + \frac{1}{2} \sum_{i,j} a_i a_j \left\langle k(\cdot, \mathbf{x}_i), k(\cdot, \mathbf{x}_j) \right\rangle_{\mathcal{H}} + \sum_i p_i \left| a_i \right|_{s_i}. \end{aligned} \tag{16}$$

The reproducing property of the kernel can be applied to show

$$\begin{aligned} \langle f(\cdot), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} &= f(\mathbf{x}) = y_i, & (17) \\ \langle k(\cdot, \mathbf{x}_i), k(\cdot, \mathbf{x}_j) \rangle_{\mathcal{H}} &= k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{G}_{i,j}, & (18) \end{aligned}$$

where the Gram matrix notation was used. Now, neglecting the first term of $F[\mathbf{a}]$*, which is independent of* $\mathbf{a}$*, one has*

$$\frac{1}{2}\mathbf{a}^T\mathbf{G}\mathbf{a} - \mathbf{y}^T\mathbf{a} + \sum_i p_i \left|a_i\right|_{s_i} \rightarrow \min_{\mathbf{a}}. \quad (19)$$

Then the $\mathbf{s}$*-insensitive terms can be rewritten by introducing slack variables [9] and the following form can be derived*

$$\min_{\mathbf{a},\mathbf{a}^+,\mathbf{a}^-} \left[\frac{1}{2}\mathbf{a}^T\mathbf{G}\mathbf{a} - \mathbf{y}^T\mathbf{a} + \mathbf{p}^T\left(\mathbf{s}^+ + \mathbf{s}^-\right)\right], \quad (20)$$

$$\textit{provided that} \left\{ \begin{array}{rcl} \mathbf{a} & \leq & \mathbf{s} + \mathbf{s}^+ \\ -\mathbf{a} & \leq & \mathbf{s} + \mathbf{s}^- \\ \mathbf{0} & \leq & \mathbf{s}^+, \mathbf{s}^- \end{array} \right\},$$

with its dual form given as

$$\begin{aligned} &\max_{\mathbf{d}^+,\mathbf{d}^-,\mathbf{q}^+,\mathbf{q}^- \geq 0} L(\mathbf{d}^+, \mathbf{d}^-, \mathbf{q}^+, \mathbf{q}^-) = & (21) \\ &= \frac{1}{2}\mathbf{a}^T\mathbf{G}\mathbf{a} - \mathbf{y}^T\mathbf{a} + \mathbf{p}^T\left(\mathbf{s}^+ + \mathbf{s}^-\right) - (\mathbf{q}^+)^T\mathbf{s}^+ - (\mathbf{q}^-)^T\mathbf{s}^- - \\ &\quad -(\mathbf{d}^+)^T(\mathbf{s} + \mathbf{s}^+ - \mathbf{a}) - (\mathbf{d}^-)^T(\mathbf{s} + \mathbf{s}^+ + \mathbf{a}). \end{aligned}$$

According to the condition on the saddle-point, the derivatives of Langrangian L *taken by the prime variables disappear at optimum, that is*

$$\begin{aligned} \mathbf{0} = \frac{dL}{d\mathbf{a}} &= \mathbf{a}^T\mathbf{G} - \mathbf{y}^T + (\mathbf{d}^+ - \mathbf{d}^-)^T, & (22) \\ \mathbf{0} = \frac{dL}{d\mathbf{s}^+} &= \mathbf{p}^T - (\mathbf{d}^+)^T - (\mathbf{q}^+)^T, & (23) \\ \mathbf{0} = \frac{dL}{d\mathbf{s}^-} &= \mathbf{p}^T - (\mathbf{d}^-)^T - (\mathbf{q}^-)^T. & (24) \end{aligned}$$

Reordering and transposing Eq. (22)*, we have*

$$\begin{aligned} \mathbf{a}^T\mathbf{G} &= \left(\mathbf{y} - (\mathbf{d}^+ - \mathbf{d}^-)\right)^T, & (25) \\ \mathbf{G}\mathbf{a} &= \left(\mathbf{y} - (\mathbf{d}^+ - \mathbf{d}^-)\right), & (26) \end{aligned}$$

where the symmetric property of Gram matrix $\mathbf{G}$ *was exploited. One can replace matrix* $\mathbf{G}$ *of the Lagrangian by* $\mathbf{G}\mathbf{G}^{-}\mathbf{G}$ *according to Eq.* (2)*. Also, considering that*

$$\mathbf{a}^T\mathbf{G}\mathbf{a} = \mathbf{a}^T(\mathbf{G}\mathbf{G}^{-}\mathbf{G})\mathbf{a} = (\mathbf{a}^T\mathbf{G})\mathbf{G}^{-}(\mathbf{G}\mathbf{a}) \tag{27}$$

one can insert the expressions for $\mathbf{a}^T\mathbf{G}$ *and* $\mathbf{G}\mathbf{a}$ *from Eqs.* (25) *and* (26)*, respectively. Equations* (23) *and* (24) *can also be applied for Lagrangian* L*. Variables* $\mathbf{q}^+, \mathbf{q}^-$ *disappear from Lagrangian* L*, but the non-negativity conditions Eqs.* (23) *and* (24) *give rise to constraints* $\mathbf{p} \geq \mathbf{d}^+$ *and* $\mathbf{p} \geq \mathbf{d}^-$ *for variables* $\mathbf{d}^+$ *and* $\mathbf{d}^-$*. We can also change the minimization of Lagrangian* L *to maximization by changing the sign.*

Taken together, we have the QP task

$$\min_{\mathbf{p}\geq\mathbf{d}^+,\mathbf{d}^-\geq\mathbf{0}} \left[\frac{1}{2}(\mathbf{y} - (\mathbf{d}^+ - \mathbf{d}^-))^T\mathbf{G}^{-}(\mathbf{y} - (\mathbf{d}^+ - \mathbf{d}^-)) + (\mathbf{d}^+ + \mathbf{d}^-)^T\mathbf{s}\right]. \tag{28}$$

The terms of the quadratic expression can be expanded and reordered. Upon dropping terms not containing variables $\mathbf{d}^+$ *or* $\mathbf{d}^-$*, and making use of the symmetric property of* $\mathbf{G}^{-}$ *inherited from* $\mathbf{G}$*, one has*

$$\min_{\mathbf{p}\geq\mathbf{d}^+,\mathbf{d}^-\geq\mathbf{0}} \left[\frac{1}{2}(\mathbf{d}^+ - \mathbf{d}^-)^T\mathbf{G}^{-}(\mathbf{d}^+ - \mathbf{d}^-) - (\mathbf{d}^+ - \mathbf{d}^-)^T\mathbf{G}^{-}\mathbf{y} + (\mathbf{d}^+ + \mathbf{d}^-)^T\mathbf{s}\right]. \tag{29}$$

Now, we are in the position to compare this optimization task with Eq. (14) *by making use of the generalized inverse* $\mathbf{G}^{-}$ *of Gram matrix* $\mathbf{G}$*. The result is that*

$$\textit{Dual}[(\mathbf{c},\mathbf{e})\textit{–SVM}] \leftrightarrow \textit{Dual}[(\mathbf{p},\mathbf{s})\textit{–sparse}]\,.$$

In short, we proved that the two tasks transform onto each other through $\mathbf{G}^{-}$ *in the following way*

$$(\mathbf{d}^*, \mathbf{d}, \mathbf{G}, \mathbf{y}) \leftrightarrow (\mathbf{d}^+, \mathbf{d}^-, \mathbf{G}^{-}\mathbf{G}\mathbf{G}^{-}, \mathbf{G}^{-}\mathbf{y}) = (\mathbf{d}^+, \mathbf{d}^-, \mathbf{G}^{-}, \mathbf{G}^{-}\mathbf{y}), \tag{30}$$

where in the last step, property $\mathbf{G}^{-}\mathbf{G}\mathbf{G}^{-} = \mathbf{G}^{-}$ *of the generalized inverse (Eq.* (3)*) was exploited.* □

5 Conclusions

We have extended the concept of sparse representation in RKHSs to a larger class of tasks, where individual components can have individual sparsifying terms. We showed that alike to the original sparse formulation, the generalized ϵ-sparse approach also has an equivalent SVM task family. This novel formulation may gain applications in signal processing, clustering and categorization problems.

References

[1] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] S.S. B. Chen, D.L. Donoho, and M.A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal of Scientific Computing*, 20(1):33–61, 1999.

[3] F. Girosi. An Equivalence Between Sparse Approximation and Support Vector Machines. *Neural Computation*, 10(6):1455–1480, 1998.

[4] R. Herbrich. *Learning Kernel Classifiers*. The MIT Press, 2002.

[5] S.Y. Huang and Y.J. Lee. Equivalence Relations Between Support Vector Machines, Sparse Approximation, Bayesian Regularization and Gauss-Markov Prediction. Technical report, Inst. of Stat. Science, Academia Sinica, Computer Sci. and Inf. Engn., National Taiwan University, 2003.

[6] B. Olshausen and D.J. Field. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision Research*, 37:3311–3325, 1997.

[7] B. Schölkopf, C.J.C. Burges, and A.J. Smola. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, Ma., 1999.

[8] B. Schölkopf, R. Herbrich, and A.J. Smola. A Generalized Representer Theorem. In *Proc. of the 14th Ann. Conf. on Comp. Learning Theory*, volume 2111 of *Lecture Notes In Computer Science*, pages 416–426, London, UK, 2001. Springer-Verlag.

[9] A.J. Smola. *Learning with Kernels*. PhD thesis, Technische Universität Berlin, 1998.

[10] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston, Washington, DC, USA, 1977.

[11] V. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.

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