# A Convex Approach to Validation-Based Learning of the Regularization Constant

K. Pelckmans, J. A. K. Suykens, and B. De Moor

***Abstract*—This letter investigates a tight convex relaxation to the problem of tuning the regularization constant with respect to a validation based criterion. A number of algorithms is covered including ridge regression, regularization networks, smoothing splines, and least squares support vector machines (LS-SVMs) for regression. This convex approach allows the application of reliable and efficient tools, thereby improving computational cost and automatization of the learning method. It is shown that all solutions of the relaxation allow an interpretation in terms of a solution to a weighted LS-SVM.**

***Index Terms*—Convex optimization, model selection, regularization.**

## I. Introduction

The importance of setting the regularization constant has been emphasized for decades, for a full introduction into the topic we refer to [11]. Here, we confine ourselves to a summary of some key concepts: the regularization constant plays a crucial role in Tikhonov regularization [19], ridge regression [9], smoothing splines [22], regularization networks [6], support vector machines (SVMs) [20], and least squares support vector machines (LS-SVMs) [14], [17] among others. Different criteria were proposed to measure the appropriateness of a regularization constant for given data, including cross validation (CV), Moody's $C_p$, and minimum description length (MDL); see, e.g., [8] for references. A whole track of research is involved with finding good approximations to those criteria; see, e.g., generalized CV (GCV) [7], the span estimate [2], or methods exploiting matrix properties [1], while interest is arising in closed-form descriptions of the solution path [5]. A practical drawback of most model selection procedures (see, e.g., [2]) is that a global optimization needs to be performed, often resulting in (local) suboptimal results. This letter is related to the work on learning the kernel [10], but is more generic as it covers a whole range of learning methods and model selection criteria.

Motivation for this research is stimulated by practical and theoretical considerations: automatic tuning the model class towards the task at hand constitutes an important step towards fully standalone algorithms. Convexity not only avoids local optima and the lack of reproducibility, but allows for the application of efficient and reliable tools. The relaxation is furthermore useful as it delivers a good starting value if one insists on solving the original problem. A theoretical motivation is found in the fact that many complexity measures do not increase by considering, instead of a set of solutions, its minimal convex hull (e.g., in the case of Rademacher complexity [15] and others). Finally, this result on learning the regularization constant also provides a generic framework towards approaching more involved hyperparameter tuning problems. Specifically, the approach sets out a path to tune a learning machine efficiently for multiple hyperparameters as, e.g., in problems of input selection, while the extension to related model selection criteria such as CV follows along the same lines. Remark that the amount of regularization lies at the core of model selection in nonlinear models, e.g., in smoothing splines [22], this parameter also regulates the amount of smoothness [comparable to the role of the bandwidth in a radial basis function (RBF) kernel], while in techniques such as LASSO [18], the process of input selection (translated in terms of sparseness) is controlled again by the regularization tradeoff.

We proposed earlier (see, e.g., [11] and [12]) the formulation of the model selection problem as a constrained optimization problem, and eventually relaxing it into a convex problem which can be solved properly using standard tools. As a consequence, we are able to recover the optimal model with respect to a classical training criterion and, simultaneously, the corresponding optimal regularization constant with respect to a model selection criterion such as CV. This letter proposes a much tighter relaxation and gives an application to the elementary task of setting the regularization constant in LS-SVMs for regression [14], [17], and indicates the workability of the approach with some numerical results. Finally, it is shown that for a special form of weighted LS-SVMs the original problem of tuning all the regularization constants coincides exactly with the proposed relaxation.

This letter is organized as follows. Section II states the problem in general and the relaxation is introduced. Section III illustrates the implications in practice for setting the regularization constant in LS-SVMs for regression and shows the relationship with weighted LS-SVMs.

## II. Ridge Solution Set

The problem and corresponding convex approach is stated in this section independently of a specific model representation. Let $b \in \mathbb{R}^N$ be a given vector and let $A = A^T \in \mathbb{R}^{N \times N}$ be a positive–semidefinite symmetric matrix. Tikhonov regularization schemes [19] of linear operators typically lead to the solution $\hat{u} \in \mathbb{R}^N$ of the following set of linear equations for a fixed $0 < \nu < \infty$:

$$(A + \nu I_N)u = b. \tag{1}$$

A wide class of algorithms for modeling leads to the solutions in this form, e.g., parametric methods as ridge regression [9] and RBF networks and nonparametric techniques as smoothing splines [22], regularization networks [6], and LS-SVMs [14], [17]. In the last case, one typically expresses (1) in terms of $\gamma = 1/\nu > 0$. Let the function $h_{A,b} : \mathbb{R}_0^+ \to \mathbb{R}^N$ be defined as $h_{A,b}(\nu) = (A + \nu I_N)^{-1}b$. This vector-valued function generates a set of solutions corresponding with all positive choices of $\nu \geq \nu_0$ for a fixed value of $\nu_0 > 0$.

*Definition 1 (Ridge Solution Set):* The ridge solution set $\mathcal{S}_0$ is defined as the set of all solutions $u$ to (1) corresponding with a value $\nu_0 < \nu < \infty$

$$\mathcal{S}_{\nu_0}(A,b) = \Big\{ h_{A,b}(\nu) = u \in \mathbb{R}^N | \exists 0 < \nu_0 \leq \nu < \infty \text{ s.t. } (A + \nu I_N)u = b \Big\}. \tag{2}$$

Manuscript received March 22, 2006; revised August 22, 2006; accepted November 12, 2006. This work was supported in part by the BOF PDM/05/161 under Grant V 4.090.05N, by the IPSI Fraunhofer FgS, Darmstadt, Germany, by the Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, by the Flemish Government (FWO) under Doctoral/Postdoctoral Grants and Projects G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, G.0302.07, by the research communities ICCoS, ANMMM, and MLDM, by the IWT Doctoral Grants GBOU (McKnow), by the Eureka-Flite2—Belgian Federal Science Policy Office under IUAP P5/22 and PODO-II,- EU: FP5-Quprodis, by ERNSI under contract research/agreements ISMC/IPCOS, Data4s, TML, Elia, LMS, and Mastercard.

The authors are with the Katholieke Universiteit Leuven (K. U. Leuven)—ESAT—SCD/SISTA, Leuven B-3001, Belgium (e-mail: kristiaan.pelckmans@esat.kuleuven.be).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2007.891187

1045-9227/$25.00 © 2007 IEEE

Authorized licensed use limited to: Uppsala Universitetsbibliotek. Downloaded on June 10, 2009 at 07:15 from IEEE Xplore. Restrictions apply.

The term *ridge solution set* is used instead of the term *regularization path* [5] in order to set the stage for more involved problems (tuning more hyperparameters). Let $A = U\Sigma U^T$ denote the SVD of the matrix $A$ with $U = [U_1, \ldots, U_N] \in \mathbb{R}^{N\times N}$ and $UU^T = U^TU = I_N$ and let $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_N) \in \mathbb{R}^{N\times N}$ containing all ordered positive singular values such that $\sigma_1 \geq \ldots \geq \sigma_N \geq 0$.

*Proposition 1 (Lipschitz Smoothness):* The function $h_{A,b} : \mathbb{R}_0^+ \to \mathbb{R}^N$ is Lipschitz smooth.

*Proof:* Let $\nu_0$ denote the minimal allowed regularization parameter. Let $R \in \mathbb{R}^+$ be defined as $R \geq \max_i |b_i|$. Then, for the two constants $\nu_0 \leq \nu_1 \leq \nu_2 < \infty$, the following holds:

$$\begin{aligned}
&\left\|(A+\nu_1 I_N)^{-1}b - (A+\nu_2 I_N)^{-1}b\right\|_2 \\
&\quad = \left\|U\left((\Sigma+\nu_1 I_N)^{-1} - (\Sigma+\nu_2 I_N)^{-1}\right)U^T b\right\|_2 \\
&\quad \leq \max_i \left|\frac{1}{\sigma_i+\nu_1} - \frac{1}{\sigma_i+\nu_2}\right| \|b\|_2 \\
&\quad \leq \|b\|_2 \left|\frac{1}{\nu_1+\sigma_N} - \frac{1}{\nu_2+\sigma_N}\right|. \qquad (3)
\end{aligned}$$

This quantity can be further bounded by application of the mean value theorem for the function $g(\nu') = 1/(\sigma_N + \nu_0 + \nu')$ with derivative $g' : \mathbb{R}^+ \to \mathbb{R}^-$. There exists a value $(\nu_1 - \nu_0) \leq \xi \leq (\nu_2 - \nu_0)$ such that

$$\begin{aligned}
\|b\|_2 \left|g(\nu_1-\nu_0) - g(\nu_2-\nu_0)\right| &\leq \|b\|_2 \left|g'(\xi)\right| |\nu_1-\nu_2| \\
&\leq \frac{\|b\|_2|\nu_1-\nu_2|}{(\sigma_N+\nu_0)^2} \leq \frac{R\sqrt{N}}{(\sigma_N+\nu_0)^2}|\nu_1-\nu_2| \qquad (4)
\end{aligned}$$

which follows from the Cauchy–Schwartz inequality and the fact that $\|b\|_2 \leq R\sqrt{N}$. ∎

This result can then be used to bound the complexity of the hypothesis class of the solution set (2) (*how "many" solutions should be examined in order to tune $\nu \geq \nu_0$?*). The result is stated without the proof due to the space limitations. For details on $\epsilon$-entropy, Lipschitz smooth functions, and learnability, see, e.g., [3] and [4]. The $\epsilon$-covering number of $\mathcal{S}_{\nu_0}(A,b)$ becomes $\mathcal{N}_\epsilon(\mathcal{S}_{\nu_0}(A,b),\epsilon) \leq \lceil R\sqrt{N}/\epsilon(\nu_0+\sigma_N)\rceil$, and the $\epsilon$-entropy of $\mathcal{S}_{\nu_0}(A,b)$ becomes $\mathcal{O}(\log\lceil R\sqrt{N}/\epsilon(\nu_0+\sigma_N)\rceil)$. It follows that the regularization constant is learnable if this term is finite. Specifically, this result indicates the importance of choosing a nonzero value $\nu_0$ in case the matrix $A$ becomes singular (i.e., $\sigma_N = 0$). For a discussion of the relation of the minimal (sample) eigenvalue to $N$, we refer the reader to [21] and citations.

We now study a convex relaxation to the set $\mathcal{S}_{\nu_0}(A,b)$. The importance of this relaxation is found in the possibility to search efficiently through a convex set of hypotheses as illustrated in Section III. The main idea is to rewrite

$$\begin{aligned}
u = h_{A,b}(\nu) &= (A+\nu I_N)^{-1}b = U(\Sigma+\nu I_N)^{-1}U^T b \\
&= \sum_{i=1}^{N}\left(\frac{1}{\sigma_i+\nu}\right)U_i U_i^T b \qquad (5)
\end{aligned}$$

and then replacing the term $1/(\sigma_i+\nu)$ by a new variable $\lambda_i$ for all $i = 1, \ldots, N$. Lemma 1 provides necessary, linear constraints on the set $\{\lambda_i\}_{i=1}^N$ following from this reparameterization.

*Lemma 1 (Convex Relaxation to $\mathcal{S}_{\nu_0}(A,b)$):* Let $\sigma_i' = \sigma_i + \nu_0$ for all $i = 1, \ldots, N$. Let $\mathcal{R}_{\nu_0}(A,b)$ in $\mathbb{R}^N$ be a polytope parametrized by $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$ as shown in (6a)–(6c) at the bottom of the page. Then, the set $\mathcal{R}_{\nu_0}(A,b)$ is convex and forms a convex hull to $\mathcal{S}_{\nu_0}(A,b)$.

*Proof:* The inequalities (6a) and (6b) are easily verified by studying the monotonically decreasing function $g(\nu)$. The necessity of $(\sigma_{i+1}'/\sigma_i')\lambda_{i+1} \leq \lambda_i$ is proven as follows with $\sigma_{i+1}' \leq \sigma_i'$:

$$\begin{aligned}
\sigma_i' = \sigma_{i+1}' + \left(\sigma_i' - \sigma_{i+1}'\right) \Leftrightarrow \lambda_i &= \frac{\lambda_{i+1}}{1+\lambda_{i+1}\left(\sigma_i' - \sigma_{i+1}'\right)} \\
&\geq \frac{\lambda_{i+1}}{1+\frac{1}{\sigma_{i+1}'}\left(\sigma_i' - \sigma_{i+1}'\right)} \\
&= \lambda_{i+1}\left(\frac{\sigma_{i+1}'}{\sigma_i'}\right) \qquad (7)
\end{aligned}$$

where the last inequality follows from $\lambda_{i+1} \leq (1/\sigma_{i+1}')$ which decreases the denominator. Denote furthermore that the set $\mathcal{R}_{\nu_0}(A,b)$ is characterized entirely by equalities and inequalities which are linear in the unknowns $\{\lambda_i\}_{i=1}^n$, and, hence, it forms a polytope. The statement that the set $\mathcal{R}_{\nu_0}(A,b)$ is a convex relaxation to the set $\mathcal{S}_{\nu_0}(A,b)$ follows now from the previous necessity result together with the property of convexity of a polytope. ∎

*Corollary 1 (Maximal Difference):* The maximal distance between a solution $\Lambda$ in $\mathcal{R}_{\nu_0}(A,b)$ and its closest counterpart in $\mathcal{S}_{\nu_0}(A,b)$ is bounded as follows:

$$\min_\nu \left\|U\Lambda U^T b - (A+\nu I_N)^{-1}b\right\|_2 \leq \frac{N\sqrt{R}}{\sigma_N+\nu_0}. \qquad (8)$$

*Proof:* The maximal difference between a solution $u$ in the set $\mathcal{R}_{\nu_0}(A,b)$ and its corresponding closest $u$ in $\mathcal{S}_{\nu_0}(A,b)$ can be written as

$$\begin{aligned}
&\min_\nu \left\|U\Lambda U^T b - (A+\nu I_N)^{-1}b\right\|_2 \\
&\qquad \leq \|b\|_2 \min_\nu \max_{i=1,\ldots,N}\left(\left|\lambda_i - \frac{1}{\sigma_i+\nu}\right|\right) \qquad (9)
\end{aligned}$$

which follows along the same lines as in (4). Then, using the property for any two values of $\lambda_i$ and $\lambda_k$, the minimum $\min_\nu \max\{|\lambda_i - 1/(\sigma_i-\nu)|, |\lambda_k - 1/(\sigma_k-\nu)|\})$ is bounded as follows: $\min_\nu \max_i(|\lambda_i - (1/(\sigma_i+\nu))|) < (1/(\sigma_N+\nu_0))$, as $\lambda_i \leq 1/(\sigma_i+\nu)$. Combining this equation with (9) gives the inequality $\forall u \in \mathcal{R}_{\nu_0}(A,b), \exists\hat{\nu}$ such that $\|u - h_{A,b}(\hat{\nu})\|_2 \leq N\sqrt{R}/(\sigma_N+\nu_0)$.

Fig. 1(a) displays a solution path $\mathcal{S}_{\nu_0}(A,b)$ and the corresponding relaxation $\mathcal{R}_{\nu_0}(A,b)$. This polytope is defined with $\mathcal{O}(N)$ (in)equality constraints. Note that this relaxation is not minimal (the set $\mathcal{S}_{\nu_0}(A,b)$ has smooth boundaries), such that the complexity is an order of magnitude higher than the original problem of tuning $\nu$ in (1).

*Corollary 2 ($\epsilon$-Entropy of $\mathcal{R}_{\nu_0}(A,b)$):* The $\epsilon$-entropy $\mathcal{N}_\epsilon(\mathcal{R}_{\nu_0}(A,b))$ becomes $\mathcal{O}(\log(\sqrt{N}/(\epsilon(\nu_0+\sigma_N))^N))$.

$$\mathcal{R}_{\nu_0}(A,b) = \begin{cases}
U_i^T u = \lambda_i U_i^T b & \forall i = 1,\ldots,N \qquad (6a)\\
0 < \lambda_i \leq \frac{1}{\sigma_i'} & \forall i = 1,\ldots,N \qquad (6b)\\
\left(\frac{\sigma_{i+1}'}{\sigma_i'}\right)\lambda_{i+1} \leq \lambda_i \leq \lambda_{i+1} & \forall\sigma_i' \geq \sigma_{i+1}' \quad \forall i = 1,\ldots,N-1. \qquad (6c)
\end{cases}$$

Authorized licensed use limited to: Uppsala Universitetsbibliotek. Downloaded on June 10, 2009 at 07:15 from IEEE Xplore. Restrictions apply.

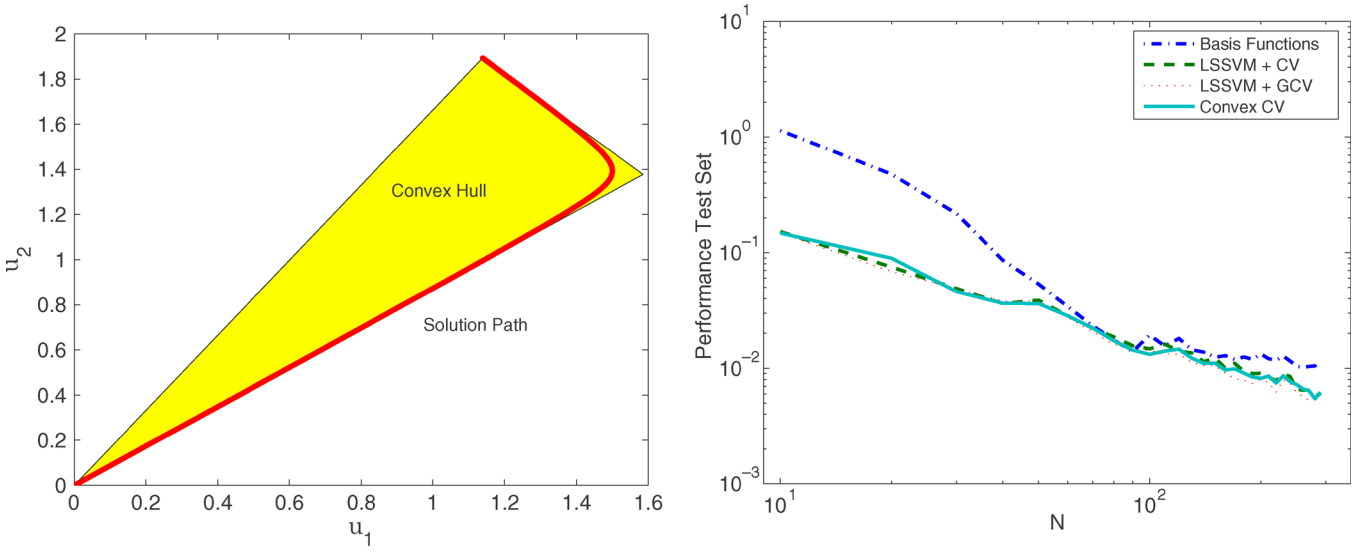


Fig. 1. (a) Display of the proposed convex relaxation to the solution path. The smooth curve represents the original solution path $\mathcal{S}_{\nu_0}(A,b)$ and the shaded polytope represents the relaxation $\mathcal{R}_{\nu_0}(A,b)$. (b) Comparison between LS-SVM tuned by CV and GCV with a gradient-descent method and the proposed approach based on (12). A basis-function approach with $n_b = \lceil n/2\log(n)x \rceil$ equispaced RBFs. An artificial data set is constructed for a univariate toy problem, where $f(x) = \mathrm{sinc}(x) + e_i$ and $\{e_i\} \sim \mathcal{N}(0,1)$.

TABLE I
NUMERICAL RESULTS OF EXPERIMENTS. THE PERFORMANCES ARE EXPRESSED IN TERMS OF THE $R^2$ OF THE ERROR (AND THE PCC) ON AN INDEPENDENT TEST SET. THE LAST DATA SET ILLUSTRATES THAT THE SAME APPROACH YIELDS GOOD RESULTS IN A CONTEXT OF CLASSIFICATION. THE PERFORMANCE IS INDICATED IN TERMS OF THE INCREASING NUMBER OF SAMPLES $n$ INDICATING THAT THE PROPOSED APPROACH PERFORMS EQUALLY WELL AS THE LS-SVM MODEL TUNED WITH CV AND GRADIENT-DESCENT SEARCH

| | Basis Functions | LS-SVM + CV | LS-SVM + GCV | Convex CV |
|---|---|---|---|---|
| Sinc ($n = 50$) | 0.058 | 0.043 | 0.044 | 0.042 |
| Motorcycle | 0.207 | 0.139 | 0.148 | 0.131 |
| Abalone | 0.035 | 0.033 | 0.074 | 0.034 |
| Ionosphere | 92.26% | 94.74% | 93.25% | 94.74% |

## III. APPLICATION TO LS-SVMS FOR REGRESSION

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^D \times \mathbb{R}$ be a data set with $n$ independent identically distributed (i.i.d.) samples, and let $\varphi : \mathbb{R}^D \to \mathbb{R}^{D_\varphi}$ be a mapping to a possibly infinite dimensional ($D_\varphi \to \infty$) feature space. The LS-SVM model (without intercept term) is $f(x) = w^T\varphi(x)$ with $w \in \mathbb{R}^{D_\varphi}$ where parameters $w$ are estimated as $(\hat{w}, \hat{e}) = \arg\min_{w,e} \mathcal{J}_\nu(w,e) = (1/2\nu)\sum_{i=1}^n e_i^2 + (1/2)w^T w$ such that $w^T\varphi(x_i) + e_i = y_i$ for $i = 1,\ldots,n$ and $\gamma = 1/\nu$. Let the symmetric positive–semidefinite matrix $\Omega$ be defined as $\Omega_{ij} = K(x_i,x_j) = \varphi(x_i)^T\varphi(x_j)$ for $i,j = 1,\ldots,n$ and for an appropriate kernel function $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. The dual solution is given by the linear system [14], [17], which is of the same form as (1)

$$(\Omega + \nu I_n)\alpha = Y \tag{10}$$

where $\alpha \in \mathbb{R}^n$ is the vector of unknowns and $Y = (y_1,\ldots,y_n)^T \in \mathbb{R}^n$ is a given vector. The corresponding estimate can be evaluated as $\hat{f}(x_*) = \sum_{i=1}^n \hat{\alpha}_i K(x_i, x_*)$. Let $\mathcal{D}^v = \{(x_j^v, y_j^v)\}_{j=1}^{n_v} \subset \mathbb{R}^D \times \mathbb{R}$ be a validation data set, sampled i.i.d. from the same distribution underlying $\mathcal{D}$. The optimization problem of finding the optimal regularization parameter corresponding to a minimal validation performance is given as

$$(\hat{\alpha}, \hat{\nu}) = \arg\min_{\alpha,\nu} \sum_{j=1}^{n_v} \ell\left(\Omega_j^{v\,T}\alpha - y_j^v\right) \quad \text{s.t. } (\alpha,\nu) \in \mathcal{S}_{\nu_0}(\Omega, Y) \tag{11}$$

where $\Omega_j^v = (K(x_i,x_j^v),\ldots,K(x_n,x_j^v))^T \in \mathbb{R}^n$ and $\ell : \mathbb{R} \to \mathbb{R}^+$ is a convex loss function. The main idea is then to replace the nonconvex set $\mathcal{S}_{\nu_0}(\Omega, Y)$ by the convex relaxation $\mathcal{R}_{\nu_0}(\Omega, Y)$. Let $\Lambda = \{\lambda_1,\ldots,\lambda_n\}$ parameterize the convex set $\mathcal{R}_{\nu_0}(\Omega, Y)$, then

$$(\hat{\alpha}, \hat{\Lambda}) = \arg\min_{\alpha,\Lambda} \sum_{j=1}^{n_v} \ell\left(\Omega_j^{v\,T}\alpha - y_j^v\right) \quad \text{s.t. } (\alpha,\Lambda) \in \mathcal{R}_{\nu_0}(\Omega, Y) \tag{12}$$

which can be solved as a linear programming problem in the case of the robust loss function $\ell(z) = |z|$. The extension to $L$-fold CV follows along the same lines, making use of $\mathcal{O}(Ln)$ (in)equality constraints (see [12] and [13]). Fig. 1(b) compares the performance on a toy example of the proposed convex approach with respect to tuning by using a gradient descent. The following result states that for a closely related weighted [16] training criterion, $\mathcal{R}_{\nu_0}(\Omega, Y)$ coincides with the proposed convex solution path.

*Lemma 2 (Weighted LS-SVM Yielding a Convex Solution Path):* The convex set $\mathcal{R}_{\gamma_0}(\Omega, Y)$ spans exactly the solution set for the following modified weighted LS-SVM problem:

$$(\hat{w}, \hat{e}) = \arg\min_{w,e} \mathcal{J}_\Gamma(w,e) = (U^T e)^T\Gamma(U^T e) + \frac{1}{2}w^T w$$
$$\text{s.t. } w^T\varphi(x_i) + e_i = y_i \qquad \forall i = 1,\ldots,n \tag{13}$$

Authorized licensed use limited to: Uppsala Universitetsbibliotek. Downloaded on June 10, 2009 at 07:15 from IEEE Xplore. Restrictions apply.

where $e = (1, \ldots, 1)^T \in \mathbb{R}^n$ and $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_n)$ contain the weighting terms. These weighting terms $\{\gamma_i\}_{i=1}^n$ are chosen such that the constraints 1) $\gamma_i \leq \gamma_0$ for all $i = 1, \ldots, n$, 2) $\gamma_{i+1} \leq \gamma_i$ for all $i = 1, \ldots, n-1$, and 3) $((\sigma_i + \gamma_0^{-1})/(\sigma_{i+1} + \gamma_0^{-1}))(\sigma_{i+1} + \gamma_{i+1}^{-1}) \geq (\sigma_i + \gamma_i^{-1}) \geq (\sigma_{i+1} + \gamma_{i+1}^{-1})$ for all $i = 1, \ldots, n-1$ are satisfied.

*Proof:* Let $\Omega$ be decomposed as $\Omega = U \Sigma U^T$ with $U$ orthonormal and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$. A primal-dual derivation of the Lagrange as in [11] and [17] states that for weighting constants $\Gamma$ solutions follow from the dual system $(\Omega + U \Gamma^{-1} U^T)\alpha = U(\Sigma + \Gamma^{-1})U^T \alpha = Y$. Now, equating the definition of the terms $\{\lambda_i\}_{i=1}^N$ in Lemma 1 and the terms $\{1/(\sigma_i + \gamma_i^{-1})\}_{i=1}^N$ gives the relation $\lambda_i = 1/(\sigma_i + \gamma_i^{-1}) \Leftrightarrow \gamma_i^{-1} = (1/\lambda_i) - \sigma_i$. Translating the inequalities in (6) in terms of $\gamma_i^{-1}$ gives the result. ∎

This result proofs that the solutions in $\mathcal{R}_{\nu_0}(\Omega, Y)$ which are not contained in the original path $\mathcal{S}_{\nu_0}(\Omega, Y)$ are the optimal solution to a slightly related learning machine, and indicate that convexity of the model selection problem can be obtained through considering alternative parameterizations of the regularization scheme.

A proof of concept is given using various small-to-medium-sized data sets. Table I reports the result on an artificial regression data set ($y = \mathrm{sinc}(x) + e$ with $e \sim \mathcal{N}(0, 1)$), where $n = 50$ and $D = 1$, the motorcycle data set ($n = 133$ and $D = 1$; see, e.g., [22]), and the University of California at Irvine (UCI) abalone data set ($n = 4177$ and $D = 8$), respectively. The performance is expressed as the $R^2$ performance of an independent test set. Kernel parameters where optimized by CV using a gradient-descent procedure. The outlined method can also be used for classification problems in the context of LS-SVMs [17]. The performance of the UCI ionosphere classification data set ($n = 351$ and $D = 34$) is expressed as the percentage correctly classified (PCC) samples of an independent test set. Fig. 1(b) illustrates the behavior of the relaxation with respect to a classical gradient-descent method for the artificial data when $n$ is growing. The general conclusion of these studies is that a relaxation does not imply a detoriation of the generalization ability.

## IV. Conclusion

This letter established a methodology for casting the problem of tuning the regularization constant for a class of regularized problems using standard convex problem solvers, hereby, preventing the need of user interaction in tuning learning machines towards the application at hand. The analysis indicates the need for a proper choice of a minimal amount of regularization, even in the original nonconvex case. The prototypical case of tuning the regularization constant with respect to a validation criterion is studied in some detail. We have shown that for a weighted form of LS-SVM for regression, the original problem and its relaxation are identical.

This principled approach towards automatizing model tuning is particularly promising for further model selection procedures as input selection. The application of this technique to learning machines as SVMs remains a challenge, mainly because the corresponding QP is not completely determined in terms of an eigenvalue decomposition.

## Acknowledgment

The authors would like to thank M. Pontil, U. von Luxburg, and O. Chapelle for discussions from which the result presented in this letter emerged.

## References

[1] G. C. Cawley and N. L. C. Talbot, "Fast exact leave-one-out cross-validation of sparse least-squares support vector machines," *Neural Netw.*, vol. 17, no. 10, pp. 1467–1475, 2004.

[2] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 131–159, 2002.

[3] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[4] R. M. Dudley, "Universal donsker classes and metric entropy," *Ann. Probability*, vol. 15, no. 4, pp. 1306–1326, 1987.

[5] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[6] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Comput.*, vol. 7, pp. 219–269, 1995.

[7] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215–223, 1979.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Heidelberg, Germany: Springer-Verlag, 2001.

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–82, 1970.

[10] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, M. I. Jordan, and L. El Ghaoui, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.

[11] K. Pelckmans, "Primal-dual kernel machines," Ph.D. dissertation, Faculty Eng., K. U. Leuven, Leuven, Belgium, 2005.

[12] K. Pelckmans, J. A. K. Suykens, and B. De Moor, "Additive regularization trade-off: Fusion of training and validation levels in kernel methods," *Mach. Learn.*, vol. 62, no. 3, pp. 217–252, Mar. 2006.

[13] K. Pelckmans, J. A. K. Suykens, and B. De Moor, ESAT-SISTA, K. U. Leuven, "A convex approach to learning the ridge based on CV," Leuven, Belgium, Tech. Rep. 05-216, 2005.

[14] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, 1998, pp. 515–521.

[15] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[16] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: Robustness and sparse approximation," *Neurocomput.*, vol. 48, no. 1–4, pp. 85–105, 2002.

[17] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[18] R. J. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc.*, vol. 58, pp. 267–288, 1996.

[19] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*. Washington, DC: Winston, 1977.

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

[21] U. von Luxburg, O. Bousquet, and M. Belkin, "Limits of spectral clustering," in *Advances in Neural Information Processing Systems (NIPS)*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, vol. 17.

[22] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.

Authorized licensed use limited to: Uppsala Universitetsbibliotek. Downloaded on June 10, 2009 at 07:15 from IEEE Xplore. Restrictions apply.