Morozov, Ivanov and Tikhonov regularization based LS-SVMs

Kristiaan Pelckmans, Johan A.K. Suykens, and Bart De Moor
{kristiaan.pelckmans,johan.suykens}@esat.kuleuven.ac.be
http://www.esat.kuleuven.ac.be/sista/lssvmlab

KULeuven - ESAT - SCD,
Kasteelpark Arenberg 10, B - 3001 Leuven (Heverlee), Belgium
Tel. +32 - 16 - 32 11 45, Fax +32 - 16 - 32 19 70.

Abstract. This paper contrasts three related regularization schemes for kernel machines using a least squares criterion, namely Tikhonov and Ivanov regularization and Morozov's discrepancy principle. We derive the conditions for optimality in a least squares support vector machine context (LS-SVMs) where they differ in the role of the regularization parameter. In particular, the Ivanov and Morozov scheme express the trade-off between data-fitting and smoothness in the trust region of the parameters and the noise level respectively which both can be transformed uniquely to an appropriate regularization constant for a standard LS-SVM. This insight is employed to tune automatically the regularization constant in an LS-SVM framework based on the estimated noise level, which can be obtained by using a nonparametric technique as e.g. the differogram estimator.

1 Introduction

Regularization has a rich history which dates back to the theory of inverse ill-posed and ill-conditioned problems [9, 13, 15] inspiring many advances in machine learning [6, 16], support vector machines and kernel based modeling techniques [10, 7, 8]. Determination of the regularization parameter in the Tikhonov scheme is considered to be an important problem [16, 7, 2, 18]. Recently [3], this problem was approached from an additive regularization point of view: a more general parameterization of the trade-off was proposed generalizing different regularization schemes. Combination of this convex scheme and validation or cross-validation measures can be solved efficiently for the regularization trade-off as well as the training solutions.

This paper considers three classical regularization schemes [13, 15, 17] in a kernel machine framework based on LS-SVMs [8] which express the trade-off between smoothness and fitting in respectively the noise level and a trust region. It turns out that both result into linear sets of equations as in standard LS-SVMs where an additional step performs the bijection of those constants into an appropriate regularization constant of standard LS-SVMs. The practical relevance of this result is mainly seen in the exact derivation of the translation of prior knowledge as the noise level or the trust region. The importance of the noise level for (nonlinear) modeling and hyper-parameter tuning was already stressed in [18, 4]. The Bayesian framework (for neural networks,

Gaussian processes as well as SVMs and LS-SVMs, see e.g. [19, 8]) allows for a natural integration of prior knowledge in the derivations of a modeling technique, though it often leads to non-convex problems and computationaly heavy sampling procedures. Nonparameteric techniques for the estimation of the noise level were discussed in e.g. [4, 5] and can be employed in the discussed Morozov scheme.

This paper is organized as follows: Section 2 compares the primal-dual derivations of LS-SVM regressors based on Tikhonov regularization, Morozov's discrepancy principle and an Ivanov regularization scheme. Section 3 describes an experimental setup comparing the accuracy of the schemes in relation to classical model-selection schemes.

2 Tikhonov, Morozov and Ivanov based LS-SVMs

Let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be the training data where $x_1, \ldots, x_N$ are deterministic points (fixed design) and $y_i = f(x_i) + e_i$ with $f : \mathbb{R}^d \to \mathbb{R}$ an unknown real-valued smooth function and $e_1, \ldots, e_N$ uncorrelated random errors with $E[e_i] = 0$, $E\left[e_i^2\right] = \sigma_e^2 < \infty$. The model for regression is given as $f(x) = w^T \varphi(x) + b$ where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes a potentially infinite ($n_h = \infty$) dimensional feature map. In the following, the Tikhonov scheme [9], Morozov's discrepancy principle [15] and Ivanov Regularization scheme are elaborated simultaneously to stress the correspondences and the differences. The cost functions are given respectively as

– Tikhonov [8]

$$\min_{w,b,e_i} \mathcal{J}_T(w,e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \text{ s.t. } w^T \varphi(x_i) + b + e_i = y_i, \quad \forall i = 1, ..., N. \tag{1}$$

– Morozov's discrepancy principle [15], where the minimal 2-norm of w realizing a fixed noise level σ^2 is to be found:

$$\min_{w,b,e_i} \mathcal{J}_M(w) = \frac{1}{2} w^T w \text{ s.t. } \begin{cases} w^T \varphi(x_i) + b + e_i = y_i, \forall i = 1, \ldots N \\ N\sigma^2 = \sum_{i=1}^N e_i^2. \end{cases} \tag{2}$$

– Ivanov [13] regularization amounts at solving for the best fit with a 2-norm on w smaller than π^2. The following modification is considered in this paper:

$$\min_{w,b,e_i} \mathcal{J}_I(e) = \frac{1}{2} e^T e \text{ s.t. } \begin{cases} w^T \varphi(x_i) + b + e_i = y_i, \forall i = 1, \ldots N \\ \pi^2 = w^T w. \end{cases} \tag{3}$$

The use of the equality (instead of the inequality) can be motivated in a kernel machine context as these problems are often ill-conditioned and result in solutions on the boundary of the trust region $w^T w \leq \pi^2$.

The Lagrangians can be written respectively as

$$\mathcal{L}_T(w,b,e_i;\alpha_i) = \frac{1}{2}w^Tw - \frac{\gamma}{2}\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(w^Tx_i + b + e_i - y_i).$$

$$\mathcal{L}_M(w,b,e_i;\alpha_i,\xi) = \frac{1}{2}w^Tw - \xi(\sum_{i=1}^N e_i^2 - N\sigma^2) - \sum_{i=1}^N \alpha_i(w^Tx_i + b + e_i - y_i).$$

$$\mathcal{L}_I(w,b,e_i;\alpha_i,\xi) = \frac{1}{2}e^Te - \xi(w^Tw - \pi^2) - \sum_{i=1}^N \alpha_i(w^Tx_i + b + e_i - y_i).$$

The conditions for optimality are

Condition	Tikhonov	Morozov	Ivanov
$\frac{\partial\mathcal{L}}{\partial w} = 0$	$w = \sum_{i=1}^N \alpha_i\varphi(x_i)$	$w = \sum_{i=1}^N \alpha_i\varphi(x_i)$	$2\xi w = \sum_{i=1}^N \alpha_i\varphi(x_i)$
$\frac{\partial\mathcal{L}}{\partial b} = 0$	$\sum_{i=1}^N \alpha_i = 0$	$\sum_{i=1}^N \alpha_i = 0$	$\sum_{i=1}^N \alpha_i = 0$
$\frac{\partial\mathcal{L}}{\partial e_i} = 0$	$\gamma e_i = \alpha_i$	$2\xi e_i = \alpha_i$	$e_i = \alpha_i$
$\frac{\partial\mathcal{L}}{\partial \alpha_i} = 0$	$w^T\varphi(x_i) + b + e_i = y_i,$	$w^T\varphi(x_i) + b + e_i = y_i,$	$w^T\varphi(x_i) + b + e_i = y_i$
$\frac{\partial\mathcal{L}}{\partial \xi} = 0$	$-$	$\sum_{i=1}^N e_i^2 = N\sigma^2$	$w^Tw = \pi^2$

(4)

for all $i = 1,\ldots,N$. The kernel-trick is applied as follows: $\varphi(x_k)^T\varphi(x_l) = K(x_k,x_l)$ for an appropriate kernel $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ in order to avoid explicit computations in the high dimensional feature space. Let $\Omega \in \mathbb{R}^{N\times N}$ be such that $\Omega_{kl} = K(x_k,x_l)$ for all $k,l = 1,\ldots,N$. The Tikhonov conditions result in the following set of linear equations as classical [8]:

$$\textbf{Tikhonov}: \qquad \left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{\gamma}I_N \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right]. \tag{5}$$

Re-organizing the sets of constraints of the Morozov and Ivanov scheme results in the following sets of linear equations where an extra nonlinear constraint relates the Lagrange multiplier ξ with the hyper-parameter σ^2 or π^2

$$\textbf{Morozov}: \qquad \left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{2\xi}I_N \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \quad \text{s.t.} \quad N\sigma^2 = \alpha^T\alpha, \tag{6}$$

and π^2

$$\textbf{Ivanov}: \qquad \left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \frac{1}{2\xi}\Omega + I_N \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \quad \text{s.t.} \quad \pi^2 = \alpha^T\Omega\alpha. \tag{7}$$

2.1 Formulation in terms of the singular value decomposition

This subsection rephrases the optimization problem (2) in terms of the Singular Value Decomposition (SVD) of Ω [12]. For notational covenience, the bias term b is omitted from the following derivations. The SVD of Ω is given as

$$\Omega = U\Gamma U^T \text{ s.t. } U^T U = I_N, \tag{8}$$

where $U \in \mathbb{R}^{N\times N}$ is orthonormal and $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_N)$ with $\gamma_1 \geq \cdots \geq \gamma_N$. Using the orthonormality [12], the conditions (6) can be rewritten as

$$\begin{cases} \alpha & = U\left(\Gamma + \frac{1}{2\xi} I_N\right)^{-1} p & (a) \\ N\sigma^2 & = \frac{1}{4\xi^2} p^T \left(\Gamma + \frac{1}{2\xi} I_N\right)^{-2} p = \sum_{i=1}^{N} \left(\frac{p_i}{2\xi\gamma_i + 1}\right)^2, & (b) \end{cases} \tag{9}$$

where $p = U^T y \in \mathbb{R}^N$. Rewriting the Ivanov scheme (7) yields

$$\begin{cases} \alpha & = U\left(\frac{1}{2\xi}\Gamma + I_N\right)^{-1} p & (a) \\ \pi^2 & = \sum_{i=1}^{N} \frac{\gamma_i p_i^2}{(\frac{1}{2\xi}\gamma_i + 1)^2}. & (b) \end{cases} \tag{10}$$

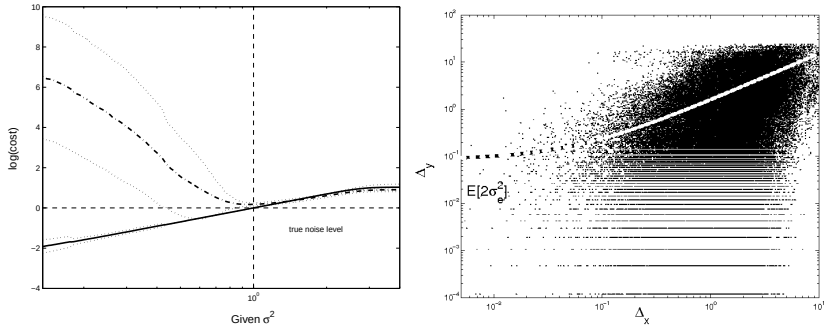
One refers to the equations (9.b) and (10.b) as secular equations [12, 17].

The previous derivation can be exploited in practical algorithms as follows. As the secular equation (9.b) is strictly monotone in the Lagrange multiplier ξ, the roles can be reversed (the inverse function exist for a nontrivial positive interval): once a regularization constant ξ is chosen, a unique corresponding noise level σ^2 is fixed. Instead of translating the prior knowledge σ^2 or π^2 using the secular equation (which needs an SVD), one can equivalently look for a ξ value resulting respectively in exactly the specified σ^2 or π^2. This can be done efficiently in a few steps by using e.g. the bisection algorithm [12]. The previous derivation of the monotone secular equations states that one obtains not only a model resulting in respectively the specified noise level or trust region, but one gets in fact the optimal result in the sense of (2) and (3).

Figure 1.a illustrates the training and validation performance for Morozov based LS-SVMs for a sequence of strictly positive noise levels. The figure indicates that overfitting on the training comes into play as soon as the noise level is underestimated. The error bars were obtained by a Monte-Carlo simulation as described in the next Section. Figure 1.b shows the technique for model-free noise variance estimation using the differogram [4, 5]. This method is based on a scatterplot of the differences Δ_x of any two input points and the corresponding output observations Δ_y. It can be shown that the curve $E[\Delta_y|\Delta_x = \delta]$ gives an estimate of the noise level at the value where it intersects the Y-axis ($\Delta_x = 0$).

3 Experiments

The experiments focus on the choice of the regularization scheme in kernel based models. For the design of a Monte-Carlo experiment, the choice of the kernel and kernel-parameter should not be of critical importance. To randomize the design of the underlying functions in the experiment with known kernel-parameter, the following class of



(a) (b)

Fig. 1. (a) *Training error (***solid line***) and validation error (***dashed-dotted line***) for the LS-SVM regressor with the Morozov scheme as a function of the noise level σ^2 (the dotted lines indicate error-bars by randomizing the experiment). The* **dashed lines** *denote the true noise level. One can see that for small noise levels, the estimated functions suffer from overfitting.* **(b)** *Differogram cloud of the Boston Housing Dataset displaying all differences between two inputs (Δ_x) and two corresponding outputs (Δ_y). The location of the curve passing the Y-axis given as $E[\Delta_y|\Delta_x = 0]$ results in an estimate of the noise variance.*

functions is considered

$$f(\cdot) = \sum_{i=1}^{N} \bar{\alpha}_i K(x_i, \cdot) \tag{11}$$

where x_i is equidistantly taken between 0 and 5 for all $i = 1, \ldots, N$ with $N = 100$ and $\bar{\alpha}_i$ is an i.i.d. uniformly randomly generated term. The kernel is fixed as $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2)$ for all $i, j = 1, \ldots, N$. Datapoints were generated as $y_i = f(x_i) + e_i$ for $i = 1, \ldots, N$ where e_i are N i.i.d. samples of a Gaussian distribution. Although no true underlying regularization parameter is likely to exist for (1), the *true* regularization parameter $\bar{\gamma}$ is estimated by optimizing w.r.t. a noiseless test set of size 10000.

The experiment tests the accuracy of the regularization constant tuning for Morozov's discrepancy principle (see Table 1). It compares results obtained when using exact prior knowledge of the noise level, a model-free estimate of the noise level using the differogram and using data-driven model selection methods as L-fold Cross-Validation (CV), leave-one-out CV, Mallows C_p statistic [14] and Bayesian inference [8]. An important remark is that the method based on the differogram is orders of magnitudes faster than any data-driven method. This makes the method suited as a method for picking a good starting-value for a local search typically associated with a more powerful and computationally intensive generalization criterion. Experiments on the higher dimensional Boston housing data (with standardized inputs and outputs) even suggest that the proposed measure can be sufficiently good as a model selection criterion on its own. For this experiment, one third of the data was reserved for test purposes, while the

remaining data were used for the training and selection of the regularization parameter. This procedure was repeated 500 times in a Monte-Carlo experiment.

	Morozov	Differogram	10-fold CV	leaveoneout	Bayesian	C_p	"true"
	Toy example: 25 datapoints						
mean(MSE)	0.4238	0.4385	0.3111	0.3173	0.3404	1.0072	0.2468
std(MSE)	1.4217	1.9234	0.3646	1.5926	0.3614	1.0727	0.1413
	Toy example: 200 datapoints						
mean(MSE)	0.1602	0.2600	0.0789	0.0785	0.0817	0.0827	0.0759
std(MSE)	0.0942	0.5240	0.0355	0.0431	0.0289	0.0369	0.0289
	Boston Housing Dataset						
mean(MSE)	-	0.1503	0.1538	0.1518	0.1522	0.3563	0.1491
std(MSE)	-	0.0199	0.0166	0.0217	0.0152	0.1848	0.0184

Table 1. *Numerical results on testdata of the experiments as described in Section 3.*

4 Conclusions

This paper compared derivations based on regularization schemes as Morozov discrepancy principle, Ivanov and Tikhonov regularization schemes. It employs these to incorporate prior or model-free estimates of the noise variance for tuning the regularization constant in LS-SVMs.

Acknowledgements. This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven. It is supported by grants from several funding agencies and sources: Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP/40: Sustainibility effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS and BDM are an associate and full professor with K.U.Leuven Belgium, respectively.

References

1. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
2. O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.
3. K. Pelckmans, J.A.K. Suykens, and B. De Moor. Additive regularization: fusion of training and validation levels in kernel methods. *Internal Report 03-184, ESAT-SCD-SISTA, K.U.Leuven (Leuven, Belgium)*, 2003, submitted for publication.
4. K. Pelckmans, J. De Brabanter, J.A.K. Suykens, B. De Moor. Variogram based noise variance estimation and its use in Kernel Based Regression. *in Proc. of the IEEE Workshop on Neural Networks for Signal Processing*, 2003.
5. K. Pelckmans, J. De Brabanter, J.A.K. Suykens, B. De Moor. The differogram: nonparametric noise variance estimation and its use in model selection. *Internal Report 04-41, ESAT-SCD-SISTA, K.U.Leuven (Leuven, Belgium)*, 2004, submitted for publication.
6. T. Poggio and F. Girosi. Networks for approximation and learning. In *Proceedings of the IEEE*, volume 78, pages 1481–1497. Proceedings of the IEEE, septenber 1990.
7. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
8. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.
9. A.N. Tikhonov and V.Y. Arsenin. *Solution of Ill-Posed Problems*. Winston, Washington DC, 1977.
10. V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.
11. V. Cherkassky and F. Mulier. *Learning from Data*. Wiley, New York, 1998.
12. G.H. Golub and C.F. Van Loan. *Matrix Computations*. The John Hopkins University Press, 1989.
13. V.V. Ivanov. *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*. Nordhoff International, 1976.
14. C.L. Mallows. Some comments on Cp. *Technometrics*, 40, 661-675, 1973.
15. V.A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag, 1984.
16. G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, 59, SIAM, Philadelphia, 1990.
17. A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40, 636-666, 1988.
18. V. Cherkassky. Practical selection of svm parameters and noise estimation for svm regression. *Neurocomputing, Special Issue on SVM*, 17(1), 113-126, 2004.
19. D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3), 415-447, 1992.