Asymptotically Optimal Choice of ε–Loss for Support Vector Machines

A.J. Smola[†,‡], N. Murata[*], B. Schölkopf[†,‡], and K.–R. Müller[†]
† GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
∗ RIKEN Brain Science Institute, Saitama 351-0198, Japan
‡ Australian National University, FEIT, Canberra ACT 0200, Australia
{smola, bs, klaus}@first.gmd.de, mura@brain.riken.go.jp

Abstract

Under the assumption of asymptotically unbiased estimators we show that there exists a nontrivial choice of the insensitivity parameter in Vapnik's ε–insensitive loss function which scales linearly with the input noise of the training data. This finding is backed by experimental results.

1 Introduction

Support Vector (SV) machines have shown to be an effective tool for regression and function estimation [9, 3]. However, it still lacks good theoretical results for controlling the model selection parameters. In particular there are two quantities which determine the behaviour of an SV machine once the kernel and the training data are fixed — the regularization constant and the insensitivity ε of the corresponding cost function. So far the approach mainly has been to use crossvalidation or bootstrap techniques for model selection (e.g. [3]).

The purpose of this paper is to give some theoretical analysis of the problem how to choose an optimal parameter ε for Vapnik's cost function. We will completely ignore the choice of different regularization schemes or kernels. Detailed information on this topic can be obtained in [2, 6]. For the sake of being self–contained let us briefly review some basic notions of SV regression.

2 Support Vector Regression

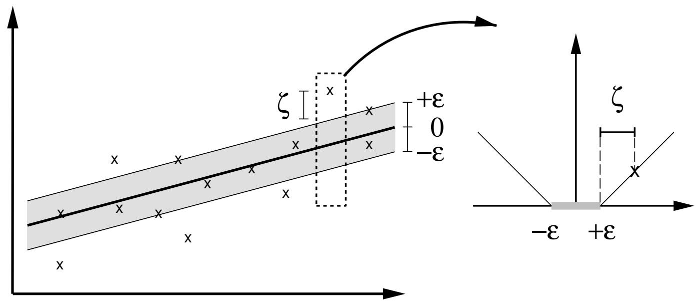
Given some training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\} \subset \mathbb{R}^n \times \mathbb{R}$ one tries to estimate a linear function

$$f : \mathbb{R}^n \to \mathbb{R} \text{ with } f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b \text{ and } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R} \tag{1}$$

that minimizes the following regularized risk functional

$$R_{\text{reg}}[f] = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{\ell} |f(\mathbf{x}_i) - y_i|_\varepsilon. \tag{2}$$

Here $|x|_\varepsilon := \max(0, |x| - \varepsilon)$ is Vapnik's ε–insensitive loss function [8].



The desired accuracy ε is specified a priori. It is then attempted to fit the flattest tube with radius ε to the data. The trade-off between model complexity and points lying outside of the tube (with positive slack variables ζ) is determined by minimizing (2). One can show [8] that this leads to a quadratic optimization problem which can be solved efficiently. It also can be shown that it is quite straightforward to estimate nonlinear functions by replacing the dot products in input space by kernels which represent dot products in feature space [1].

Vapnik's ε–insensitive cost function has algorithmical advantages such as providing sparse decompositions and rendering the computations more amenable [9]. In most cases, unfortunately, $|x|_\varepsilon$ is not the cost function with respect to which we would like to train our estimator. The problem that now arises is to find a value of ε such that, say, the variance of $y_i - f(\mathbf{x}_i)$ is minimized. Intuitively one might think that setting $\varepsilon = 0$ is the best choice that can be made, as $|x|_0 = |x|$ which leads to the standard Laplacian loss function. Experiments show, however, that this is not the case and hint that there exists a linear dependency between the level of the additive noise in the target values y_i and the size of ε. We will show that under some assumptions this finding is correct, indeed.

3 Asymptotic Efficiency

What we will show is that there is a nontrivial choice of ε for which the statistical efficiency of estimating a location parameter using the ε–insensitive loss function is maximized and that the optimal ε is proportional to the variance of the random variable (here additive noise) under consideration.

For this purpose we need to introduce some statistical notations. Denote $\hat{\alpha}(\mathbf{X})$ an estimator of the parameters α based on the sample $\mathbf{X}$ and let $\mathbf{X}$ be drawn from some probability density function $p(\mathbf{X}, \alpha)$ (also parametrized by α). Finally denote $\langle \xi \rangle_\alpha$ the expectation of the random variable ξ with respect to $p(\mathbf{X}, \alpha)$. Now we can define an unbiased estimator $\hat{\alpha}(\mathbf{X})$ by requiring $\langle \hat{\alpha}(\mathbf{X}) \rangle_{\bar{\alpha}} = \bar{\alpha}$. Moreover in this case we can introduce the Fisher information matrix I with

$$I_{ij} := \left\langle \partial_{\alpha_i} \ln p(X, \bar{\alpha}) \cdot \partial_{\alpha_j} \ln p(X, \bar{\alpha}) \right\rangle_{\bar{\alpha}} \tag{3}$$

and the covariance matrix B of the estimator $\hat{\alpha}$ by

$$B_{ij} := \langle (\hat{\alpha}_i - \langle \hat{\alpha}_i \rangle_{\bar{\alpha}})(\hat{\alpha}_j - \langle \hat{\alpha}_j \rangle_{\bar{\alpha}}) \rangle_{\bar{\alpha}} . \tag{4}$$

Then the Cramér–Rao inequality [5] states that $\det IB \geq 1$ for all possible estimators $\hat{\alpha}$. This allows us to define the statistical efficiency e of an estimator as $e := 1/(\det IB)$. Comparing the quality of unbiased estimators in this context can be reduced to comparing their statistical efficiencies.

For a special class of estimators where $\hat{\alpha}$ is defined by

$$\hat{\alpha}(\mathbf{X}) := \operatorname*{argmin}_{\alpha} d(\mathbf{X}, \alpha) \tag{5}$$

and d is a two times differentiable function in α one can show [4, Lemma 3] that asymptotically $B = Q^{-1} G Q^{-1}$ with

$$G_{ij} := \mathrm{Cov}_{\bar{\alpha}} \left\langle \partial_{\alpha_i} d\left(\mathbf{X}, \bar{\alpha}\right), \partial_{\alpha_j} d\left(\mathbf{X}, \bar{\alpha}\right) \right\rangle \text{ and } Q_{ij} := \left\langle \partial^2_{\alpha_i \alpha_j} d(\mathbf{X}, \bar{\alpha}) \right\rangle_{\bar{\alpha}} \tag{6}$$

and therefore $e = (\det Q)^2 / (\det IG)$. In our case we are only dealing with a one–parametrical model, namely estimating a location parameter (the mean of a distribution). Denote $\Phi_\varepsilon(\xi)$ the noise model given by the ε–insensitive loss function, i.e.

$$\Phi_\varepsilon(\xi) = c_0 \exp(-|\xi|_\varepsilon) = \frac{1}{2(1+\varepsilon)} \begin{cases} 1 & \text{if } |\xi| \leq \varepsilon \\ \exp(\varepsilon - |\xi|) & \text{otherwise} \end{cases} \tag{7}$$

and $\phi(\xi)$ the actual noise model of the data. Without loss of generality we will assume the location parameter to be 0 or formally $\langle \mathbf{X}_i \rangle_\phi = 0$.[1] Then the maximum likelihood estimator α is given by setting $d(\mathbf{X}, \alpha) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \ln \Phi_\varepsilon(\mathbf{X}_i - \alpha)$ and therefore

$$I = \left\langle (\partial_\alpha \ln \phi(\alpha))^2 \right\rangle_\phi \tag{8}$$

$$G = \left\langle (\partial_\alpha \ln \Phi_\varepsilon(\alpha))^2 \right\rangle_\phi = 2 \int_\varepsilon^\infty \phi(\alpha) d\alpha \tag{9}$$

$$Q = \left\langle \partial^2_\alpha \ln \Phi_\varepsilon(\alpha) \right\rangle_\phi = 2\phi(\varepsilon) \tag{10}$$

Now let us consider different choices of ϕ and the corresponding values of e.

Gaussian Noise

Set $\phi(\xi) = 1/\sqrt{2\pi\sigma^2} \exp\left(-\xi^2 / \left(2\sigma^2\right)\right)$. Then $I = \sigma^{-2}, G = 1 - \mathrm{erf}\left(\frac{\varepsilon}{\sqrt{2}\sigma}\right)$ and therefore

$$\frac{1}{e} = \frac{\det GI}{\det Q^2} = 2\pi \exp(\varepsilon^2/\sigma^2) \left(1 - \mathrm{erf}\left(\frac{\varepsilon}{\sqrt{2}\sigma}\right)\right). \tag{11}$$

The maximum of e is obtained for $\varepsilon/\sigma = 0.6166$ and therefore we have a linear dependency between ε and the noise level. As a sanity check we will compute the optimal ε for Laplacian noise.

[1] One always could redefine the problem for a nonzero location parameter by shifting all variables by the corresponding amount.

Laplacian Noise

Set $\phi(\xi) = 1/(2\sigma)\exp(-|\xi|/\sigma)$. Then $I = \sigma^{-2}, G = \exp(-\varepsilon/\sigma)$ and therefore $\frac{1}{e} = \exp(\varepsilon/\sigma)$. Here the maximum of e is achieved for $\varepsilon/\sigma = 0$, i.e. the case where Φ_ε degenerates to the L^1 loss, exactly matching the Laplacian noise. In this case the estimator is asymptotically efficient ($e = 1$). Finally let us analyze the case of arbitrary polynomial noise models.

Polynomial Noise

Many cases, e.g. the one of Gaussian noise, can be dealt with in a more general approach, namely $\phi(\xi) = \frac{p}{2\beta\Gamma(1/p)}\exp(-(\xi/\beta)^p)$ where $\Gamma(x)$ is the gamma function, $\text{var}\phi = \beta^2\frac{\Gamma(3/p)}{\Gamma(1/p)}$. There we have $I = \frac{p^2}{\beta^2}\frac{\Gamma(2-1/p)}{\Gamma(1/p)}$ and $G = \frac{p}{\Gamma(1/p)}F_p(\varepsilon/\beta)$ where $F_p(x) := \int_x^\infty \exp(-x^p)dx$. In most cases $F_p(x)$ cannot be computed in closed form. Putting everything together we get

$$\frac{1}{e} = 4F_p(\varepsilon/\beta)\exp(2(\varepsilon/\beta)^p)\Gamma(2-1/p). \tag{12}$$

One can observe that e only depends on ε/β which in turn may be written in terms of ε/σ and p. Hence the optimal ϵ scales linearly with σ provided that the optimal quotient $\tau_p = \varepsilon/\sigma \neq 0$. This, however, can be shown for $p > 1$ by computing the derivative of $1/e$ w.r.t. ε/β for $\varepsilon/\beta = 0$. One gets $\frac{d(1/e)}{d(\varepsilon/\beta)}|_0 = -4\Gamma(2-1/p) < 0$ which proves that the minimum of $1/e$ is not obtained for $\tau_0 = 0$. Hence also in the case of general polynomial loss we have a linear dependency between the noise level and the optimal choice of ε.

Arbitrary Symmetric Noise

One can derive general conditions for a linar scaling behaviour of ϵ. Assume $\phi(\xi)$ to be a symmetric density with unit variance. Hence $1/\sigma\phi(\xi/\sigma)$ has standard deviation σ. Now rewrite G and Q in terms of $\tau := \epsilon/\sigma$ as

$$G = 2\int_\epsilon^\infty \frac{1}{\sigma}\phi(\xi/\sigma)d\xi = 2\int_\tau^\infty \phi(\xi)d\xi, \tag{13}$$

$I_\sigma = I_1\sigma^{-2}$ and $Q = \frac{2}{\sigma}\phi(\tau)$. This leads to

$$e = \frac{\det Q^2}{\det GI} = \frac{4\sigma^{-2}\phi(\tau)}{\sigma^{-2}I_1 2\int_\tau^\infty \phi(\xi)d\xi} = \frac{2}{I_1}\frac{\phi^2(\tau)}{\int_\tau^\infty \phi(\xi)d\xi}. \tag{14}$$

What remains is to check that e does not have a maximum for $\tau = 0$. Computing the derivative of e w.r.t. τ for $\tau = 0$ analogously to the polynomial case yields the sufficient condition $\partial_\xi\phi(0) + \phi^2(0) > 0$. For instance any density with $\partial_\xi\phi(0) = 0$ satisfies this property.

It is not directly possible, to carry over our conclusions to the SV case due to two assumptions that may not be satisfied. Neither are we normally dealing with the asymptotic case nor is a SV machine only dealing with a single parameter at a time. Instead we have finite sample size and estimate a function. Experiments illustrate, however, that the conclusions we obtained from this simplified situation are still valid in the more complex SV case.

4 Experiments and Conclusions

We consider a toy example, namely $f(x) = 0.9\,\mathrm{sinc}\,(10x/\pi)$ on $[-1, 1]$. 100 datapoints x_i were independently drawn from a uniform distribution on $[-1, 1]$ and generated the sample via $y_i = f(x_i) + \xi_i$. Here ξ_i was a gaussian random variable with zero mean and variance σ^2. In the implementation we used a spline kernel of degree 1 with 100 gridpoints (see [9] for details). As the purpose of these experiments was to exhibit the dependency on ϵ we carried out "model–selection" for the regularization parameter in such a way to always choose that value that led to the smallest (L^1 or L^2) error on the test set. By doing so it was possible to exclude side effects of possible model selection criteria. For statistical reliability we averaged the results over multiple trials.

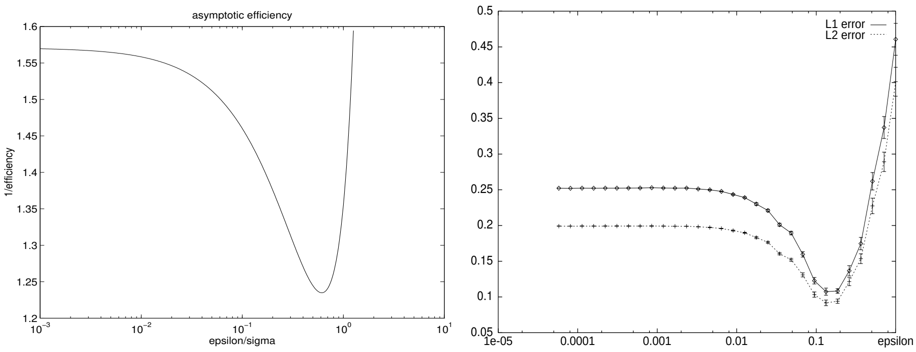


Figure 1: Left: Asymptotic efficiency for an ϵ–insensitive model and data with Gaussian noise. Right: L^1 and L^2 loss for different values of ϵ and fixed noise level $\sigma = 0.2$, averaged over 25 trials. Note the minimum for $\epsilon = 0.18$.

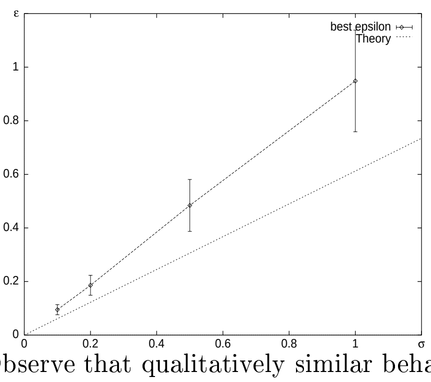


Figure 2: Experimentally optimal choice of ε for different levels of gaussian noise and the theoretically best value in the asymptic unbiased case. For each noise level σ, after determining the optimal regularization parameter, we chose the corresponding ϵ with minimal L^1 error. Although the obtained values do not match the exact value of the theoretical predictions, the scaling behaviour with the noise is clearly linear.

Observe that qualitatively similar behaviour in the errors and the efficiency of the estimator in figure 1. Also note the linear dependency in figure 2 of $\varepsilon_{\mathrm{opt}}$ on

σ. Even though it does not exhibit the same scaling factor one obtains from the asymptotic calculations still the scaling property remains unchanged.

Our finding is corroborated by results of Solla and Levin [7] where it was shown that for linear Boltzmann machines the best performance is achieved when the "internal noise" matches the external noise.

Acknowledgements: This work was supported in part by a grant of the DFG (# Ja 379/71) and the Australian Research Council. The authors thank Peter Bartlett and Robert Williamson for helpful discussions and comments.

References

[1] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

[2] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, MIT, 1993.

[3] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *Artificial Neural Networks — ICANN'97*, 1997.

[4] N. Murata, S. Yoshizawa, and S. Amari. Network information criterion—determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, 5:865–872, 1994.

[5] C. R. Rao. *Linear Statistical Inference and its Applications.* Wiley, New York, 1973.

[6] A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 1998. in press.

[7] S. A. Solla and E. Levin. Learning in linear neural networks: The validity of the annealed approximation. *Physical Review A*, 46(4):2124–2130, 1992.

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

[9] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In *NIPS 9*, San Mateo, CA, 1997.