AVAILABLE AT
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT•

Neural Networks 17 (2004) 127–141

Neural Networks

www.elsevier.com/locate/neunet

Experimentally optimal ν in support vector regression for different noise models and parameter settings

Athanassia Chalimourda[a,*], Bernhard Schölkopf[b,1], Alex J. Smola[c,2]

[a]Ruhr-Universität Bochum, Institut für Neuroinformatik, D-44780 Bochum, Germany
[b]Max-Planck-Institut for Biological Cybernetics, Spemannstraße 38, D-72076 Tübingen, Germany
[c]Australian National University Canberra, ACT 0200, Australia

Received 23 August 2002; accepted 23 June 2003

Abstract

In Support Vector (SV) regression, a parameter ν controls the number of Support Vectors and the number of points that come to lie outside of the so-called ε-insensitive tube. For various noise models and SV parameter settings, we experimentally determine the values of ν that lead to the lowest generalization error. We find good agreement with the values that had previously been predicted by a theoretical argument based on the asymptotic efficiency of a simplified model of SV regression. As a side effect of the experiments, valuable information about the generalization behavior of the remaining SVM parameters and their dependencies is gained. The experimental findings are valid even for complex 'real-world' data sets. Based on our results on the role of the ν-SVM parameters, we discuss various model selection methods.

Keywords: Support Vector machines; ν-Support Vector machines; Support Vector regression; Support Vector machine parameters; Optimal ν; Gaussian kernel; Model selection; Risk minimization

1. Introduction

Support Vector (SV) machines comprise a new class of learning algorithms, motivated by results of the statistical learning theory (Vapnik, 1995). SV regression estimation seeks to estimate functions

$$f(\mathbf{x}) = (\mathbf{w}\cdot\mathbf{x}) + b, \qquad \mathbf{w}, \mathbf{x} \in \mathbb{R}^N, \qquad b \in \mathbb{R}, \tag{1}$$

based on data

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l) \in \mathbb{R}^N \times \mathbb{R}, \tag{2}$$

by minimizing the regularized risk functional

$$\|\mathbf{w}\|^2/2 + C\cdot R^{\varepsilon}_{\text{emp}}, \tag{3}$$

where C is a constant determining the trade-off between minimizing the training error, or *empirical risk*

$$R^{\varepsilon}_{\text{emp}} := \frac{1}{l}\sum_{i=1}^{l} |y_i - f(\mathbf{x}_i)|_{\varepsilon}$$

and the model complexity term $\|\mathbf{w}\|^2$. Here, we use the so-called ε-insensitive loss function $|y - f(\mathbf{x})|_{\varepsilon} = \max\{0, |y - f(\mathbf{x})| - \varepsilon\}$. It does not penalize errors below some $\varepsilon > 0$ chosen a priori. As a consequence, the data points that lie *inside* a tube with radius ε do not contribute directly to the solution. The latter will thus in many cases inherit the property of sparseness from its ancestor, the SV pattern recognition algorithm. Although the parameter ε does control the sparseness of the solution, it does this only in a rather indirect way. Lacking a priori information about the accuracy of the y-values, it can be difficult to come up with a reasonable value of ε a priori. Instead, one would rather specify the degree of sparseness and let the algorithm compute ε from the data. This is the idea of the ν-SVM, a modification of the original ε-SVM, introduced by Schölkopf, Smola, Williamson, and Bartlett (2000), which we will briefly review in Section 2. It turns out that to get the highest generalization accuracy, the sparsity parameter $\nu \in (0, 1]$ has to be chosen in accordance with the noise that is in the y-values. In Section 3, we describe the reasoning that leads to theoretical predictions of the optimal ν values. In Section 4, we experimentally test these predictions, and observe rather good agreement. Our experiments reveal a lot of interesting properties on the generalization behavior of ν and the other

[*] Corresponding author. Tel.: +49-6104-941784.
E-mail addresses: athanassia.chalimourda@neuroinformatik.ruhr-uni-bochum.de (A. Chalimourda); bernhard.schoelkopf@tuebingen.mpg.de (B. Schölkopf); alex.smola@anu.edu.au (A.J. Smola).
[1] Tel.: +49-7071-601-551; fax: +49-7071-601-552.
[2] Tel.: +61-2-6125-8652; fax: +61-2-6125-8651.

0893-6080/$ - see front matter © 2003 Published by Elsevier Ltd.
doi:10.1016/S0893-6080(03)00209-0

parameters C and σ_{kernel}. C weighs the data influence in the Support Vector Machine, see Eq. (4), and is thus responsible for the regularization in it. σ_{kernel} gives the width of the Gaussian kernel, k, that builds the regression estimate, see Eq. (11). ν, the sparsity parameter, seems to be largely insensitive to the choice of the other two parameters. In order to examine this assumption we extend in Section 5 the experiments of the previous section. While in Section 4 we computed risk versus ν varying only one parameter at a time, in Section 5 we compute the risk while varying all parameters at the same time. Valuable information on the combined regularization effects of C and σ_{kernel} is gained as a further side effect of this section's experiments. In Section 6 we repeat the above experiments for a complex, multidimensional data set, the Boston Housing Problem. The results confirm our previous findings.

2. ε-SVM regression and ν-SVM regression

The main insight of the statistical learning theory is that in order to obtain a small risk, one needs to control both training error and model complexity, i.e. explain the data with a simple model. The minimization of Eq. (3) is equivalent to the following constrained optimization problem (Vapnik, 1995):

$$\text{minimize } \tau(\mathbf{w}, \xi^{(*)}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{4}$$

subject to the following constraints

$$((\mathbf{w}\cdot\mathbf{x}_i) + b) - y_i \le \varepsilon + \xi_i \tag{5}$$

$$y_i - ((\mathbf{w}\cdot\mathbf{x}_i) + b) \le \varepsilon + \xi_i^* \tag{6}$$

$$\xi_i^{(*)} \ge 0, \qquad \varepsilon \ge 0. \tag{7}$$

As mentioned in Section 1, at each point $\mathbf{x}_i$ we allow an error of magnitude ε. Errors above ε are captured by the slack variables $\xi^{(*)}$ (see constraints (5) and (6)). They are penalized in the objective function via the regularization parameter C chosen a priori (Vapnik, 1995)

In the ν-SVM the size of ε is not defined a priori but is itself a variable. Its value is traded off against model complexity and slack variables via a constant $\nu \in (0, 1]$:

$$\text{minimize } \tau(\mathbf{w}, \xi^{(*)}, \varepsilon) = \frac{1}{2}\|\mathbf{w}\|^2 + C\cdot(\nu\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \tag{8}$$

subject to the constraints (5)–(7). Using Lagrange multipliers techniques, one can show (Vapnik, 1995) that the minimization of Eq. (4) under the constraints (5)–(7) results in a convex optimization problem with a global minimum. The same is true for the optimization problem (8) under the constraints (5)–(7). At the optimum, the regression estimate can be shown to take the form

$$f(\mathbf{x}) = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)(\mathbf{x}_i\cdot\mathbf{x}) + b. \tag{9}$$

In most cases, only a subset of the coefficients $(\alpha_i^* - \alpha_i)$ will be nonzero. The corresponding examples $\mathbf{x}_i$ are termed *support vectors* (SVs). The coefficients and the SVs, as well as the offset b, are computed by the ν-SVM algorithm. In order to move from linear (as in Eq. (9)) to nonlinear functions the following generalization can be done (Vapnik, 1995): we map the input vectors $\mathbf{x}_i$, into a high-dimensional feature space Z through some nonlinear mapping, $\Phi : \mathbf{x}_i \to \mathbf{z}_i$, chosen a priori. We then solve the optimization problem (8) in the feature space Z. In that case, the inner product of the input vectors $(\mathbf{x}_i\cdot\mathbf{x})$ in Eq. (9) is replaced by the inner product of their icons in feature space Z, $(\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}))$. The calculation of the inner product in a high-dimensional space is computationally very expensive. Nevertheless, under general conditions (see Vapnik, 1995 and references therein) these expensive calculations can be reduced significantly by using a suitable function k such that

$$(\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x})) = k(\mathbf{x}_i\cdot\mathbf{x}), \tag{10}$$

leading to nonlinear regression functions of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)k(\mathbf{x}_i, \mathbf{x}) + b. \tag{11}$$

The nonlinear function k is called a *kernel* (Vapnik, 1995). In our work we use a Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/(2\sigma_{\text{kernel}}^2))$.

Proposition 1 (Schölkopf et al., 2000) illuminates the significance of the new parameter ν. We will use the term *errors* to refer to training points lying outside of the tube, and the term *fraction* of errors/SVs to denote the relative numbers of errors/SVs, i.e. divided by l. The *modulus of absolute continuity* of a function f is defined as the function $\epsilon(\delta) = \sup\sum_i |f(b_i) - f(a_i)|$, where the supremum is taken over all disjoint intervals (a_i, b_i) with $a_i < b_i$ satisfying $\sum_i (b_i - a_i) < \delta$. Loosely speaking, the condition on the conditional density of y given $\mathbf{x}$ asks that it is absolutely continuous 'on average' (see Schölkopf et al., 2000).

Proposition 1. *(Schölkopf et al., 2000) Suppose the ν-SVM is applied to some data set and the resulting ε is nonzero. The following statements hold:*

(i) *ν is a upper bound on the fraction of errors.*
(ii) *ν is a lower bound on the fraction of SVs.*
(iii) *Suppose that the data (2) are generated iid from a distribution $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$ with $p(y|\mathbf{x})$ continuous and the expectation of the modulus of absolute continuity of its density satisfies $\lim_{\delta\to 0}\mathbf{E}\epsilon(\delta) = 0$. Then, asymptotically, ν equals both the fraction of SVs and the fraction of errors with probability 1.*

This shows that ν can be used to control the fraction of support vectors (and hence the sparsity of the solution's expansion) and the fraction of outliers (i.e. the amount of confidence that we have in the data).

3. Asymptotically optimal choice of ν

In this section we will briefly review results of Smola, Murata, Schölkopf, and Müller (1998) and Schölkopf et al. (2000) concerning an asymptotically optimal choice of ν for a given class of noise models. For the following proposition two major assumptions are made: first, one assumes that SV regression produces an estimate $\hat{f}$ which converges to the underlying functional dependency f. Second, we replace the SVM for regression with a much simpler one-parametrical model. Instead of estimating the function f, we estimate the mean a, of an unknown distribution based on the data sample X.

Proposition 2. *Assume that the data were generated iid from a distribution $p(x,y) = p(x)p(y - f(x))$ with $p(y - f(x))$ continuous. p belongs to the family of noise models $\mathfrak{P} := \{p|p = 1/\sigma\mathfrak{p}(y/\sigma), \sigma > 0\}$. The family $\mathfrak{P}$ is generated from $\mathfrak{p}$, a density with unit variance. Then, the asymptotically optimal ν, described in Smola et al. (1998) is,*

$$\nu = 1 - \int_{-\varepsilon}^{\varepsilon} \mathfrak{p}(t)\mathrm{d}t \tag{12}$$

where

$$\varepsilon := \underset{\tau}{\operatorname{argmin}} \frac{1 - \int_{-\tau}^{\tau} \mathfrak{p}(t)\mathrm{d}t}{(\mathfrak{p}(-\tau) + \mathfrak{p}(\tau))^2} \tag{13}$$

To see Eq. (12), note that under the assumptions stated above the probability of a deviation larger than ε, $\Pr\{|y - \hat{f}(x)| > \varepsilon\}$, converges to

$$\begin{aligned}\Pr\{|y - f(x)| > \varepsilon\} &= \int_{\chi \times R \setminus [-\varepsilon,\varepsilon]} p(x)p(\zeta)\mathrm{d}x\,\mathrm{d}\zeta \\ &= 1 - \int_{-\varepsilon}^{\varepsilon} p(\zeta)\mathrm{d}\zeta \end{aligned} \tag{14}$$

Asymptotically, this is the fraction of examples that will become SVs, that is ν according to Proposition 1(iii). It corresponds to a tube of size ε. Consequently, given a noise model $p(\zeta)$, one can compute the optimal ε using Eq. (13) and then the corresponding optimal value of ν using Eq. (12).

The asymptotically optimal value of ε in Eq. (13) was estimated by Smola et al. (1998) by considering the estimation of the parameter a in a one-parametrical model instead of a regression SVM.

Example 1. For arbitrary polynomial noise models $p(\zeta)$, where

$$p(\zeta) \propto \exp(-b|\zeta|^P) \qquad \text{with } b, P > 0, \tag{15}$$

one obtains the optimal values of ν given in Table 1.

Table 1
Optimal ν for various degrees P of polynomial additive noise

Polynomial degree P	1	2	3	4	5	6	7	8
Optimal ν	1	0.54	0.29	0.19	0.14	0.11	0.09	0.07

For $P = 1$ and $P = 2$ we have Laplacian and Gaussian noise, respectively. Observe that the more lighter tailed the distribution becomes, i.e. for larger P, the smaller the optimal ν, i.e. the tube width increases. This is reasonable, since for very long tails of the distribution (data with many outliers) it appears natural to use an early cutoff of the influence of the data, thus reducing the width of the tube. The extreme case of Laplacian noise ($\nu = 1$) leads to a tube with width 0.

4. Experiments with different parameter settings and noise models

In the previous chapter we have reviewed theoretical results on the optimal values for ε and ν for the ε-SVM and ν-SVM, respectively. These theoretically optimal values were derived under very limiting assumptions, which are not satisfied in practical Support Vector machines such as replacing the SVM with a one-parametrical model and considering an asymptotical number of examples (Smola et al., 1998; Murata, Yoshizawa, & Amari, 1994). This makes the need evident to verify the above results in practice.

To this end, we numerically compute the risk (generalization error), here the root mean squared error (RMSE, or L_2 distance) as a function of ν. We concentrate on the optimal choice of ν, because it enables us to examine the whole parameter regime $\nu \in (0, 1]$. In particular we plot the risk versus ν while adding noise by different polynomial noise models to the data. Our aim is to see if the minima of the risk versus ν curves agree with the theoretically optimal ν value for each noise model. For the special case of Gaussian noise ($P = 2$), we also examine the dependence of the optimal ν values on the noise level and the remaining ν-SVM parameters: the regularization parameter C and the standard deviation σ_{kernel} of the Gaussian kernel k.

As a training set, we use 100 examples (x_i, y_i), generated by the sinc function with

$$y_i = \operatorname{sinc}(x) := \sin(\pi x_i)/(\pi x_i) + \zeta_i. \tag{16}$$

Here the x_i are drawn uniformly from the interval $[-3,3]$ and ζ_i is the additive noise, distributed according to a general polynomial distribution.[3] The test set consists of 500 equally spaced data points of the noiseless sinc function. The error bars represent 95% confidence intervals

[3] Adding noise would not change the location of minima but rather make the estimation of the latter less reliable. In the present setting we effectively compute the L_2 distance to the Bayes-optimal regressor.

for the mean risk. They were computed over 1000 trials, assuming Gaussian distribution on the risk.

For the experiments we used LOQO, an interior point algorithm developed by Vanderbei (1994). We used the duality property to recover b and ε directly from the dual variables of the optimizer.

4.1. Additive Gaussian noise

In the first experiment we added Gaussian noise to the data. Our aim was first to see whether the experimentally optimal values of ν agree with the theoretically predicted value of 0.54 (cf. Table 1) and, whether the noise level has any influence on the optimal ν. Therefore, we compute the risk (RMSE) versus ν varying only the noise level and keeping the other parameters, C and σ_{kernel} fixed. The results are shown in Fig. 1a. Observe that for all noise levels the curves are very flat and most ν values, except the smaller ones, result in a low risk. This holds in particular for the low noise case of $\sigma_{\text{noise}} = 0.1$ (corresponding to signal-to-noise ratio, SNR = 13.5) which would be most sensitive to misadjustment of ε. There exists a whole 'optimal area' for $\nu \in [0.3, 0.8]$, i.e. the theoretical value of 0.54 would be a good choice, independent of the noise level.

In Fig. 1b we examine whether the theoretically optimal value for Gaussian noise is still valid when we vary the regularization (complexity) parameter C of the function class. Again we obtain similar results for $C = 100$ and 1000. The curves are rather flat with a large optimal area for $\nu \in [0.3, 0.8]$. Thus, the theoretical $\nu_{\text{opt}}^{\text{theory}} = 0.54$ can be used for both $C = 100$ and $C = 1000$. Smaller values of C, however, lead to high risk and to atypical behavior in ν (Fig. 1b for

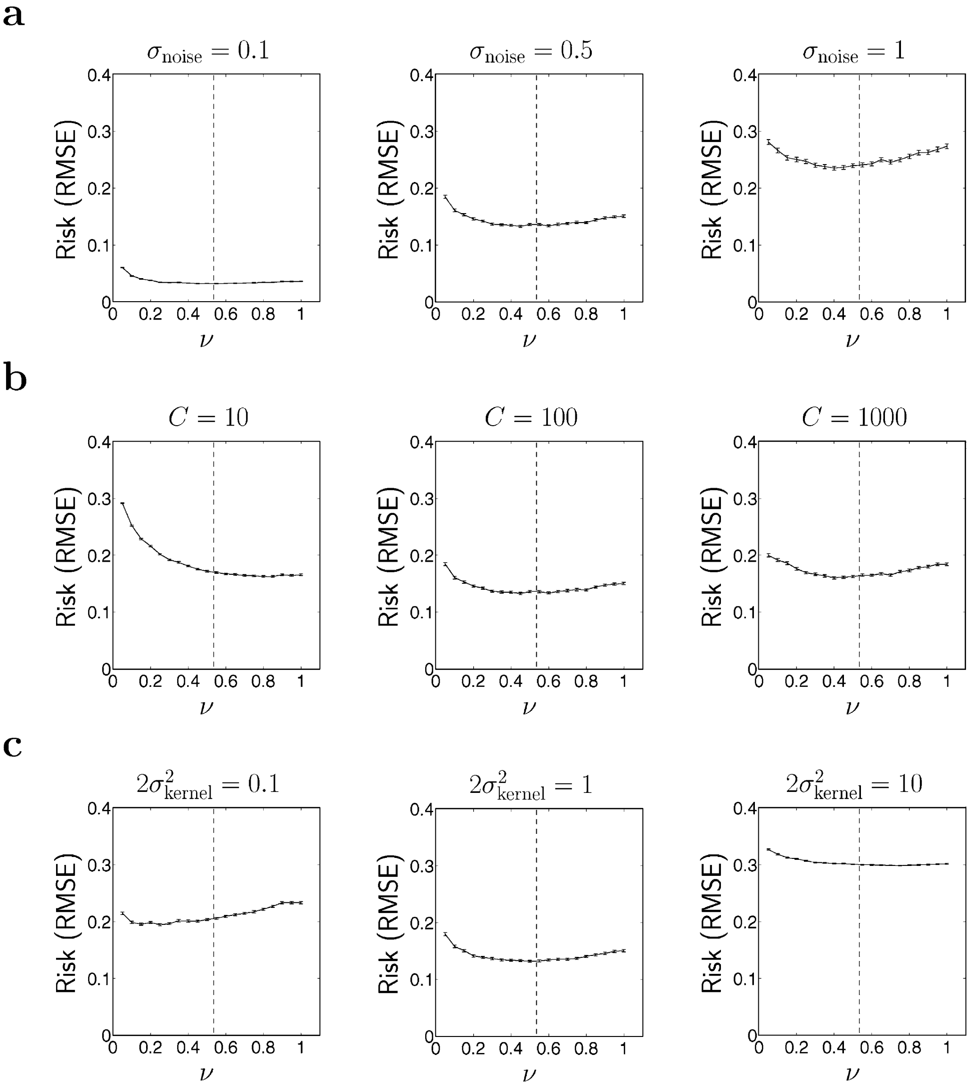

Fig. 1. Risk (RMSE) versus ν for the sinc data with added Gaussian noise, (a) with varying noise levels: from left to right, $\sigma_{\text{noise}} \in \{0.1, 0.5, 1\}$ corresponding to SNR $\in \{13.5, 0.54, 0.135\}$. For all three experiments $C = 100$, $2\sigma^2_{\text{kernel}} = 1$, constant. (b) With varying C: from left to right $C \in \{10, 100, 1000.\}$ $2\sigma^2_{\text{kernel}} = 1$, constant, additive Gaussian noise with $\sigma_{\text{noise}} = 0.5$ (SNR = 0.54). (c) With varying width of the Gaussian kernel: from left to right, $2\sigma^2_{\text{kernel}} \in \{0.1, 1, 10\}$. $C = 100$, and $\sigma_{\text{noise}} = 0.5$ (SNR = 0.54) constant. The dashed vertical line in each picture marks the position of $\nu_{\text{opt}}^{\text{theory}} = 0.54$ for Gaussian noise. Notice that for all three experiments the error bars are very small indicating that the variability between trials is very low.

$C = 10$). This may be due to the fact that the assumptions of Proposition 2 are no longer fulfilled. The risk rises for small ν, where the optimum is for values $\nu > 0.8$, and thus, is far from the theoretical value of 0.54. Nevertheless, the risk curve is rather flat for $\nu > 0.5$ (Fig. 1b for $C = 10$). That means that since the risk does not change much with ν for $\nu > 0.5$, one could use $\nu_{\text{opt}}^{\text{theory}} = 0.54$ without much effect on the risk.

Finally, in Fig. 1c we vary the standard deviation σ_{kernel} of the Gaussian kernel, keeping C and the SNR of the noise constant. Again, for small values of $2\sigma_{\text{kernel}}^2 = 0.1$, the error curve has not the light convex behavior as one might expect. Still, through the flatness of the risk versus ν curve, one could use $\nu_{\text{opt}}^{\text{theory}} = 0.54$ even if the strict risk minimum is for a much smaller ν.

Deviations from the $\nu_{\text{opt}}^{\text{theory}}$ are sometimes conceivable, see Fig. 1b for $C = 10$ and Fig. 1c for $2\sigma_{\text{kernel}}^2 = 0.1$. This may be due to the fact that moving towards smaller values of C and $2\sigma_{\text{kernel}}^2 = 0.1$, the assumptions of Proposition 2 are no longer fulfilled. This is not a practical problem, however, since the risk curves are generally flat. One could thus use the $\nu_{\text{opt}}^{\text{theory}} = 0.54$ in absence of further information, assuming Gaussian additive noise in the data.

4.2. Additive polynomial noise

In the next set of experiments we examined whether the theoretically optimal values of ν agree with the experimental findings in the general case of polynomial models with different degrees, P. In all cases we kept σ_{kernel}, the SNR of the noise, and C constant. The results are shown in Fig. 2. Each panel shows risk versus ν for a different degree P, $P \in \{1, 2, 3, 4, 6, 8\}$. For smaller P values, $P \in \{1, 2, 3\}$, we notice that the risk curves are very flat around the theoretically predicted minima of $\{1, 0.54, 0.29\}$. Thus, we may use the theoretical minimum without much effect on the risk. For larger P, $P \in \{4, 6, 8\}$, the risk curves rise more rapidly with ν. Their minima are at the beginning of the curves, near the theoretically optimal values of $\{0.19, 0.11, 0.07\}$. We conclude that in all cases we may use $\nu_{\text{opt}}^{\text{theory}}$ without much effect on the risk, as in the Gaussian additive noise case (Section 4.1). Repeating the experiments of Fig. 2 for a higher noise level ($\sigma_{\text{noise}} = 1$, SNR $= 0.135$) showed that not only the positions of the minima were maintained but also the major characteristics of the curves. Clearly, a higher noise level causes a higher risk level.

5. Experimentally optimal ν varying all ν-SVM parameters

In Section 4 we showed that the theoretically optimal ν values agree with the experimentally optimal values using data from the sinc function. The experimentally optimal ν values are largely insensitive towards the noise level added on the data, the regularization parameter C and the kernel width σ_{kernel}, provided basic model selection assumptions are satisfied. We cannot postulate, of course, that the choice of ν is independent of the choice of all remaining SVM parameters, since each time that we varied one parameter we kept the others constant. Nevertheless, it seems that we can use the theoretically optimal ν value also in practice and

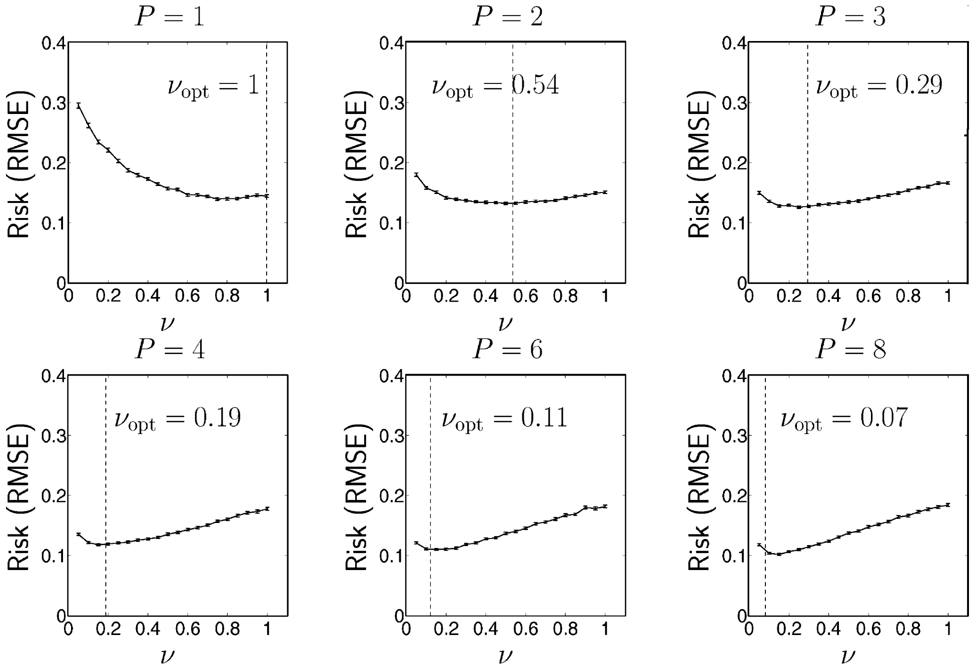

Fig. 2. Risk (RMSE) versus ν curves for the sinc data with added noise from different polynomial noise models. From left to right, from top to bottom: $P \in \{1, 2, 3, 4, 6, 8\}$. In all cases $\sigma_{\text{noise}} = 0.5$ (SNR $= 0.54$), $C = 100$ and $2\sigma_{\text{kernel}}^2 = 1$. The dashed vertical lines mark the position of the theoretical ν_{opt} for each polynomial noise model. Notice that for all three experiments the error bars are very small indicating that the variability between the 1000 trials is very low.

focus on the choice of the remaining ν-SVM parameters, C and σ_{kernel}. In order to be sure that we can do so, in the following experiments we determine the optimal ν while, this time, varying all the SVM parameters simultaneously.

The experimental setup is similar to Section 4: as training set, we used 100 examples (x_i, y_i) generated by the sinc function with $y_i = \sin(\pi x_i)/(\pi x_i) + \zeta_i$. x_i were drawn uniformly from the interval $[-3,3]$ and ζ_i is Gaussian additive noise. We concentrate on the most common noise distribution as the experiments are more complicated than in Section 4. The test set consists of 500 equally spaced data points of the noiseless sinc function. The results were averaged over 300 trials.

In order to find the optimal ν while varying all parameters at the same time, we extend the risk versus ν plots of our previous work by one more parameter, the kernel variance, σ^2_{kernel}.We preferred σ^2_{kernel} over C because the latter is the more ‘insensitive’ parameter. That means, that small changes of σ_{kernel} result in drastic changes of the risk, while C has to be changed over orders of magnitude to achieve a similar change of the risk. In order to take into account all degrees of freedom of our system, we compute the risk (here the mean squared error, MSE) versus ν and $2\sigma^2_{\text{kernel}}$ plots for different C values and noise levels of the Gaussian noise added to the data.

As in Section 4, we varied ν over its entire range, $(0,1]$, $2\sigma^2_{\text{kernel}} \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16.\}$ We used $C \in \{10, 100, 1000, 10\,000\}$, because of the insensitivity of the risk (MSE) with respect to C. Finally, we added Gaussian noise to the data with signal-to-noise ratio, $\text{SNR} \in \{20, 3, 0.5, 0.2\}$.

The results of the experiments are shown in Fig. 3 as surface plots. Fig. 4 shows the corresponding contour plots. Each panel shows the risk (MSE) versus ν and $2\sigma^2_{\text{kernel}}$ for a different C and SNR of the Gaussian additive noise. $C \in \{10, 100, 1000, 10\,000\}$ from top to bottom and $\text{SNR} \in \{20, 3, 0.5, 0.2\}$ from left to right. In both Figs. 3 and 4, we first notice the simple form of the surfaces. For different noise levels (i.e. signal-to-noise ratios) as well as

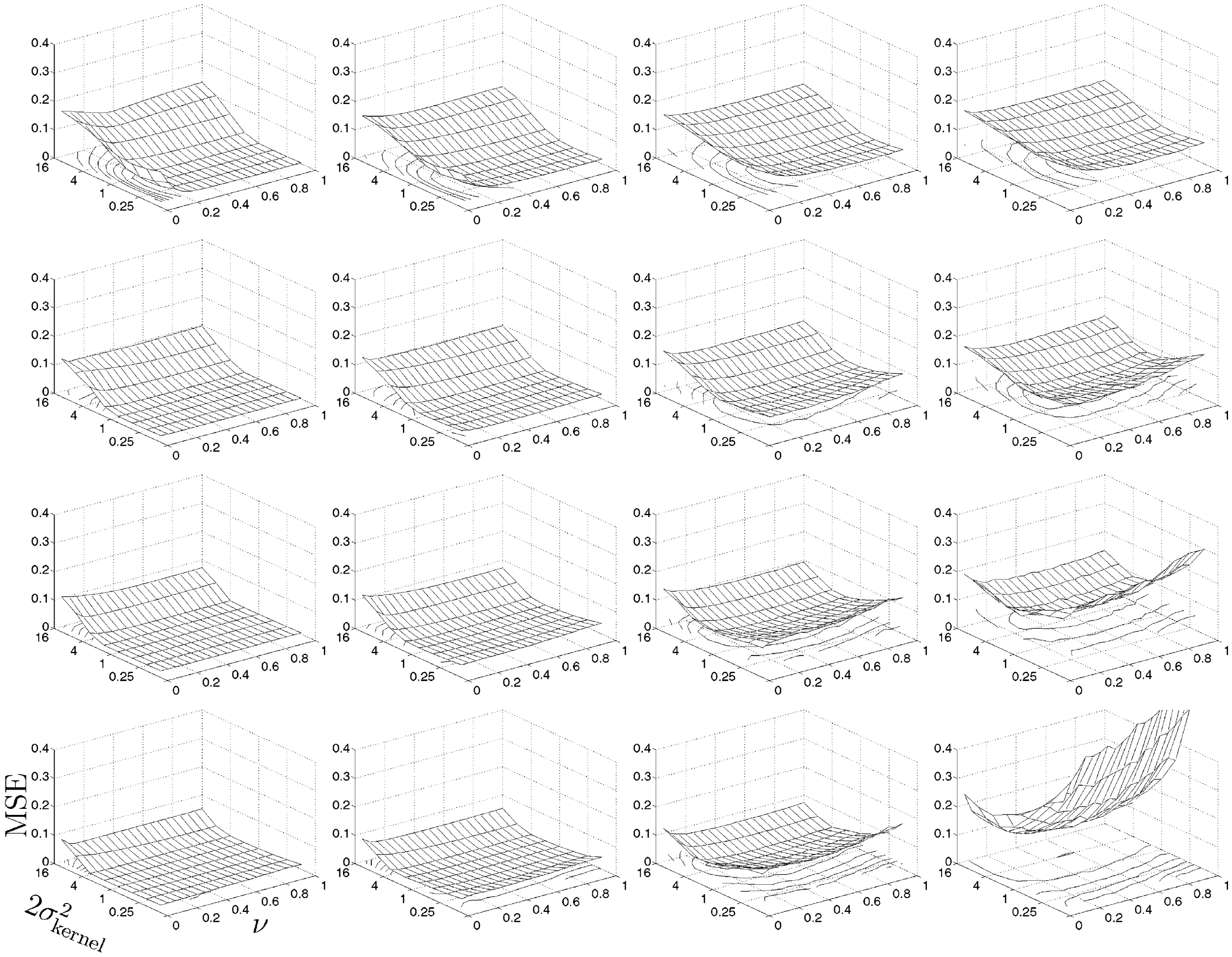

Fig. 3. Risk (mean squared error, MSE) versus ν and $2\sigma^2_{\text{kernel}}$ for the sinc function data. $\nu \in (0,1]$, (x-axis), $2\sigma^2_{\text{kernel}} \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$, ($y$-axis). Each panel shows the risk surfaces for a particular value of C and signal-to-noise-ratio (SNR). C changes from top to bottom, $C \in \{10, 100, 1000, 10\,000.\}$ SNR changes from left to right, $\text{SNR} \in \{20, 3, 0.5, 0.2\}$. The risk surfaces are convex and largely flat and smooth around their optimal area. For increasing C (panels from top to bottom) the optimal risk area moves to larger $2\sigma^2_{\text{kernel}}$. This is valid for all noise levels (panels from left to right).

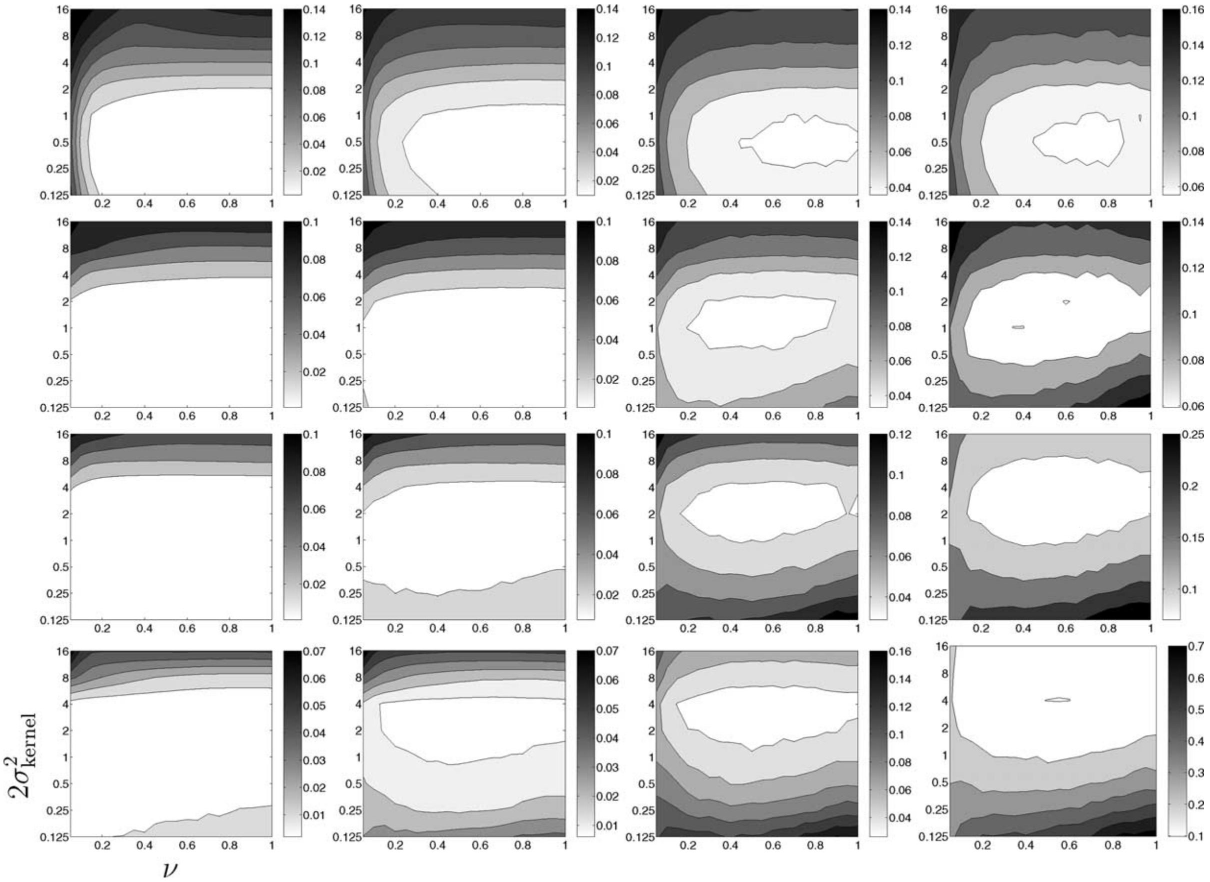

Fig. 4. Risk (Mean squared error, MSE) versus ν and $2\sigma^2_{\text{kernel}}$ for the sinc function data. Each panel shows the risk for a particular value of C and signal-to-noise ratio (SNR). $\nu \in (0,1]$, $2\sigma^2_{\text{kernel}} \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$. From top to bottom: $C \in \{10, 100, 1000, 10\,000\}$. From left to right: SNR $\in \{20, 3, 0.5, 0.2\}$. Light colors correspond to low values, dark colors to high values of MSE. Note that the scales are not equal across panels. The risk surfaces are convex and largely flat and smooth around their optimal area. For increasing C (panels from top to bottom) the optimal risk area moves to larger $2\sigma^2_{\text{kernel}}$. This is valid for all noise levels (panels from left to right).

for different C values all error surfaces are convex, having a global optimal area, without any local minima. For large signal-to-noise ratios (low noise), the optimal generalization area is large, allowing almost every value for ν. The optimal area becomes smaller for higher noise levels, nevertheless, it allows in any case $\nu \in [0.4, 0.8]$. As in Section 4, the theoretical ν value, $\nu_{\text{opt}}^{\text{theory}} = 0.54$ that was derived without taking into account the other SV-parameters can again be used in practice. We conclude from these results that we can use the theoretical value for almost all parameter settings (avoiding extreme situations, of course) even in the case that our data were completely masked by noise.

As in Section 4 we not only have results on the behavior of the experimental ν with respect to $\nu_{\text{opt}}^{\text{theory}}$, but we also reveal a number of useful properties of the generalization behavior of the ν-SVM parameters. The risk surfaces are flat and smooth with respect to ν and $2\sigma^2_{\text{kernel}}$for all different C and SNR values (Figs. 3 and 4). That means that neighboring ν and $2\sigma^2_{\text{kernel}}$ values result to neighboring risk values. This fact guarantees stability of the risk with respect to the ν-SVM parameters. It also enables us to use slightly smaller or larger ν values than the theoretically strict optimal ones, without much effect on the risk. Varying ν affects directly the number of support vectors that built the solution of the learning problem. This is useful when the number of support vectors is of importance. Smaller ν leads to wider tubes and less support vectors, i.e. larger data compression.

In Figs. 3 and 4 we also observe that varying C and the SNR of the additive noise has systematical effects on the risk surfaces. For increasing C values (Figs. 3 and 4 from top to bottom) the optimal risk area moves from smaller to larger $2\sigma^2_{\text{kernel}}$ values. In this way, the weak regularization by a large C value is out-weighted by the larger $2\sigma^2_{\text{kernel}}$ value. This effect is largely independent of the noise level in the data.

The risk levels do not change much with C. A significant change only takes place, as expected, when we increase the noise added to the data (from larger to smaller SNR values, Figs. 3 and 4 from left to right). Yet, it seems that the extent of the minimal risk areas is more affected and less their position. While we decrease the SNR for constant C, we notice that the optimal risk area becomes smaller with respect to the ν and $2\sigma^2_{\text{kernel}}$ values. Nevertheless, for a (constant) C value we can

Table 2
sinc Data: $2\sigma^2_{\text{kernel}}$ around which the risk is optimal, for $C \in \{10, 100, 1000, 10\,000\}$

C	10	100	1000	10 000
$2\sigma^2_{\text{kernel}}$	0.5	1	2	3

always find $2\sigma^2_{\text{kernel}}$ values which are optimal for all different noise levels. Table 2 gives the adequate $2\sigma^2_{\text{kernel}}$ around which the risk is optimal, for $C \in \{10, 100, 1000, 10\,000.\}$ These optimal $2\sigma^2_{\text{kernel}}$ values are independent of the noise level. That is, larger C results in larger $2\sigma^2_{\text{kernel}}$ as discussed above.

There is even an absolute minimum of the risk for these data sets for all parameter settings. It is reached for $C = 100$ and $2\sigma^2_{\text{kernel}} = 1$.

So far, our conclusions are based on the average case, estimated from 300 repetitions. For a more complete discussion, we also need to assess the variability of the risk over trials. From our previous argument, we expect that within the optimal area, the variability of the risk with respect to ν is small. That is, there is no distinguished value of ν for which the variability is exceptionally high (or low).

As a measure for the variability we chose the coefficient of variation (CV). It relates the standard deviation of a process to its mean and, thus, allows us to directly compare the relative dispersion of the risk for different C and signal-to-noise ratios. The CV is defined as

$$\text{CV} = \frac{\sigma_{\text{risk}}}{\mu_{\text{risk}}}, \tag{17}$$

where σ_{risk} and μ_{risk} are the standard deviation and the mean of the risk, respectively.

In Figs. 3 and 4 we saw that the risk, i.e. μ_{risk}, changes only little over the range of ν. If the standard deviation of the risk, i.e. σ_{risk}, exhibits a similar behavior, the CV will also show only a small dependency with respect to ν. However, if σ_{risk} exhibits a stronger dependency on ν, this will also show in the CV.

The results for the sinc data are shown in Fig. 5. Each panel shows the CV versus ν and $2\sigma^2_{\text{kernel}}$ for different C and SNR of the Gaussian additive noise. In analogy to

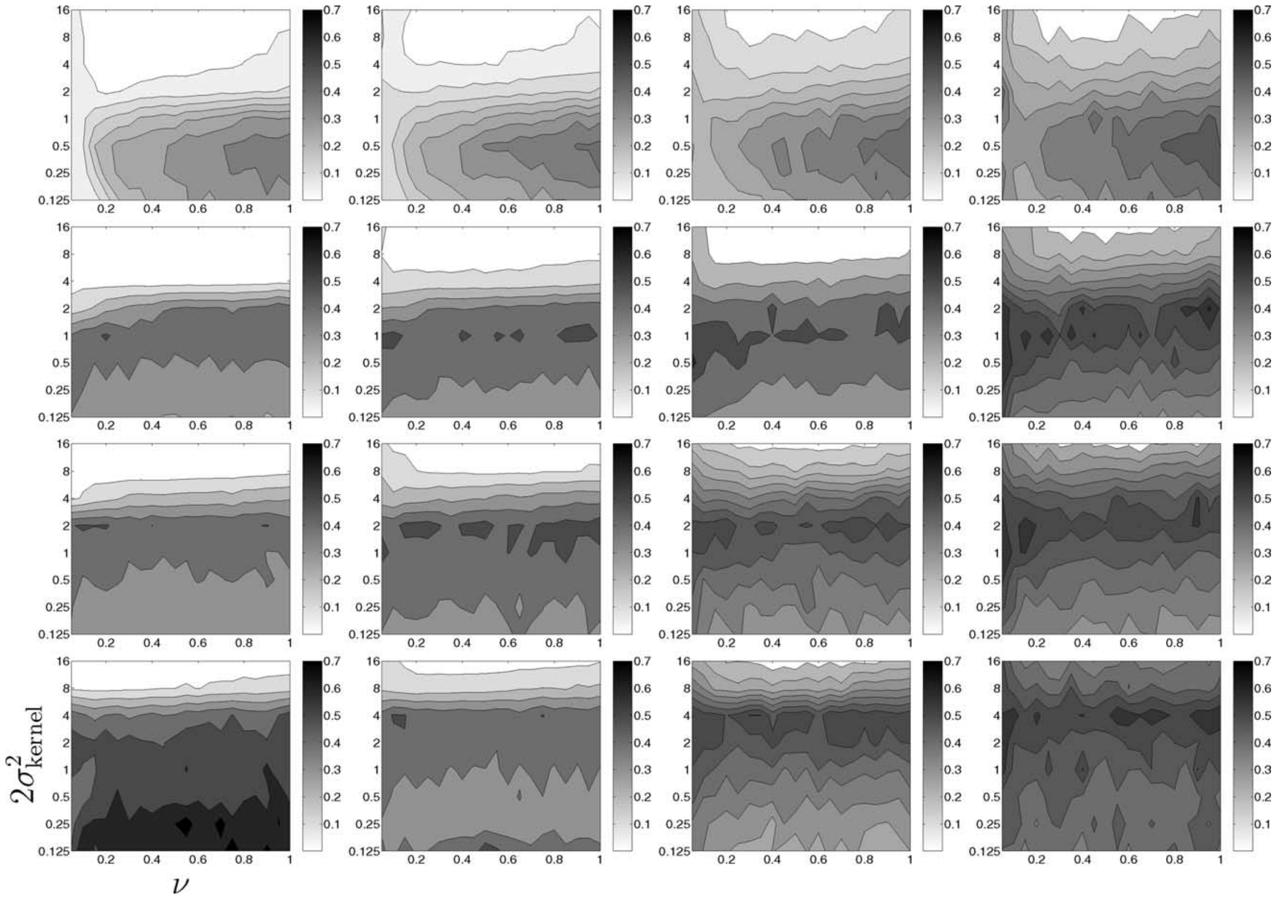

Fig. 5. The coefficient of variation (CV) versus ν and $2\sigma^2_{\text{kernel}}$ for the sinc function data. Each panel shows the risk for a particular value of C and signal-to-noise ratio (SNR) in analogy to Figs. 3 and 4. $\nu \in (0, 1]$, $2\sigma^2_{\text{kernel}} \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$. From top to bottom: $C \in \{10, 100, 1000, 10\,000\}$. From left to right: SNR $\in \{20, 3, 0.5, 0.2\}$. Light colors correspond to low values, dark colors to high values of CV. The grey levels of the coefficient of variation appear to form bands which run parallel to ν axis. The CV changes significantly with $2\sigma^2_{\text{kernel}}$ while showing only small dependency on ν.

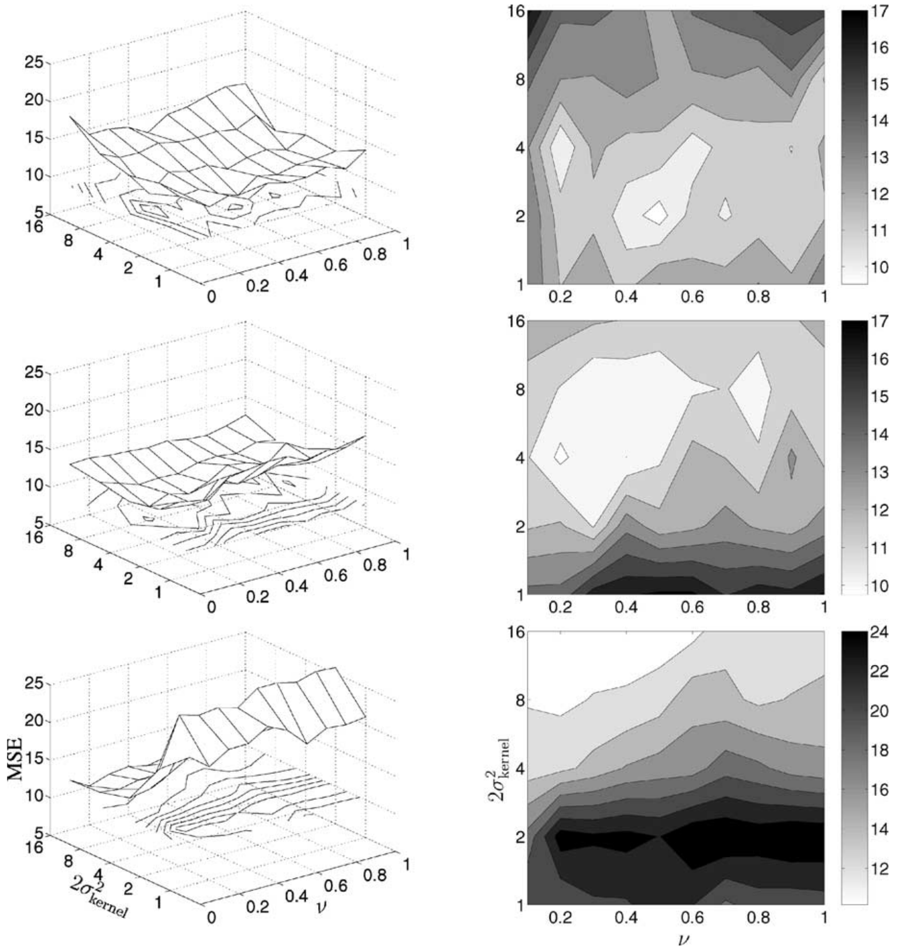

Fig. 6. Risk versus ν and $2\sigma^2_{\text{kernel}}$ for the Boston Housing Problem data as surface (left) and contour plots (right). $\nu \in (0, 1]$, $2\sigma^2_{\text{kernel}} \in \{1, 2, 4, 8, 16\}$. C changes from top to bottom: $C \in \{50 \cdot l, 10 \cdot 50 \cdot l, 100 \cdot 50 \cdot l\}$. $l = 406$ is the number of training points. The test set consists of 100 randomly chosen points. The results were averaged over 100 trials. The risk does not change much along the ν-axis. For increasing C (panels from top to bottom) the optimal risk area moves to larger $2\sigma^2_{\text{kernel}}$.

Fig. 3, $C \in \{10, 100, 1000, 10\,000\}$ from top to bottom and SNR $\in \{20, 3, 0.5, 0.2\}$ from left to right.

We indeed observe only small changes of the CV with respect to ν. Thus, σ_{risk} has a similarly small dependency on the choice of ν as the average risk itself.

The CV, however, does depend on $2\sigma^2_{\text{kernel}}$. In most panels we observe a pronounced horizontal structure: the grey-levels appear to form bands, which run parallel to the ν-axis.

These results are consistent with our view that ν is a well-behaved parameter which is easy to use in practice. The knowledge that we gained about ν and, as side effects, about the other SVM parameters are of great importance, as theoretical foundation about their effect on the risk is still incomplete.

6. The Boston Housing data: risk versus ν and $2\sigma^2_{\text{kernel}}$

In Section 5 we experimentally examined the optimal values of ν with respect to the remaining SVM parameters. As a side effect we derived valuable results on the generalization behavior of the other SVM parameters and their relations. In our experiments we used ‘toy’ data sets from the sinc function. The question arises if the above experimental results are still valid when we deal with complex ‘real-world’ data.

In this section we examine the behavior of ν and the other SVM parameters on a multidimensional data set, the *Boston Housing Problem*,[4] (Schölkopf et al., 2000; Stitson et al., 1999). The data set consists of 506 $(\mathbf{x}, y)$ data points. The $\mathbf{x}$ points are 13-dimensional vectors. Each coordinate stands for a quantity that influences the price of a house in a Bostonian suburb. The corresponding y value is the house price in thousand dollars.

As in the experiments with the sinc function data, we plot the risk (generalization error, here the MSE on the test set) versus ν and $2\sigma^2_{\text{kernel}}$. We randomly choose 406 data points for the training set, using the remaining 100 points for the test set. We average the results over 100 trials. We did not

[4] Clearly, the Boston Housing is ‘just another’ benchmark data set. With the term ‘real-world’ we would like to stress the fact that it is more complex and closer to reality than the sinc data.

add noise to the data in order not to affect their characteristics. We varied the parameter C by orders of magnitude as in Section 5, see Fig. 6 from top to bottom.

The generalization error surfaces are not so smooth as for the sinc data, but they show small fluctuations along both axes. This is natural as we now deal with a complex 'real-world' problem and not a well-controlled 'toy' data set as in the previous sections. Still, along the ν axis the risk does not change much, just like for the sinc data. As in the sinc data case, we have whole optimal ν areas rather than a sharp minimum. They are for $\nu \in [0.05, 0.5]$, for all three C values we used. We notice that the optimal area for the Boston Housing data is not in the middle of the ν interval, (0,1]. It is shifted towards lower ν values compared to the sinc data with added Gaussian noise (Sections 4 and 5). Since the position of the optimal ν depends on the noise model, this shift may be due to characteristics of the noise in the data. In the special case of the Boston Housing data maybe the distribution of the noise is shorter than normal tailed, because with house prices, everybody tries to get as close as possible to the average. Hence, large deviations from the average are rare. This is only an assumption though, as there are many markets that are known to behave in the opposite way. That is, the deviations from the average price are large giving long tails to the corresponding distribution.

The risk surfaces show similar dependencies between $2\sigma^2_{\text{kernel}}$ and C as for the sinc data. For small C values the error surface has its optimal area at small $2\sigma^2_{\text{kernel}}$ values, while for larger C values the optimal area moves towards larger $2\sigma^2_{\text{kernel}}$ values. Table 3 gives $2\sigma^2_{\text{kernel}}$ around which risk is optimal for $C \in \{50\cdot l, 10\cdot 50\cdot l, 100\cdot 50\cdot l\}$, see also Fig. 6. As in the case of the sinc data, when we too strongly weight the learning examples through a large C value a larger σ_{kernel} is needed for better regularization and thus better generalization results.

So far it appears that the risk changes rather smoothly with respect to ν. This could of course be the result of the sampling frequency along the ν axis. In order to verify that the fluctuations are indeed small, we recomputed the risk surface for one representative value ($C = 10\cdot 50\cdot l$, where $l = 406$ is the number of training data) with double resolution of ν and 300 instead of 100 trials. The results were qualitatively unchanged and we thus conclude that the risk is stable with respect to the ν-SVM parameters.

Our results so far are based on the risk (MSE) averaged over 100 trials. We now examine their variability, as we did for the sinc data. Since we want to compare the risk variability for different C values, we again use the CV. It is defined as the ratio of the standard deviation to the mean of the risk, see Eq. (17). The average risk (Fig. 6) is not as smooth as for the sinc data (Fig. 4), as we now deal with a much more complex problem. Nevertheless, it still shows rather small changes across the ν axis, allowing a large optimal area for ν. As for the sinc data, we expect that the risk variability (expressed by the CV) does not change much with respect to ν. The results are shown in Fig. 8. Each panel shows the CV versus ν and $2\sigma^2_{\text{kernel}}$ for $C \in \{50\cdot l, 10\cdot 50\cdot l, 100\cdot 50\cdot l\}$ from top to bottom, $l = 406$ is the number of training data. The left column shows the CV versus ν and $2\sigma^2_{\text{kernel}}$ as contour plots, the right column the corresponding surface plots.

Table 3
Boston Housing data: $2\sigma^2_{\text{kernel}}$ around which the risk is optimal for $C \in \{50\cdot l, 10\cdot 50\cdot l, 100\cdot 50\cdot l\}$

C	$50\cdot l$	$10\cdot 50\cdot l$	$100\cdot 50\cdot l$
$2\sigma^2_{\text{kernel}}$	2	4	8

$l = 406$, is the number of training data.

For $C = 50\cdot l$, Fig. 7 top panels, mainly the surface plot shows a rather flat CV with no significant preference to a particular ν value. The range of the CV values is small and does not exhibit an area of small risk variability either with respect to ν or the $2\sigma^2_{\text{kernel}}$.

For the two larger C, $C = 10\cdot 50\cdot l$ and $C = 100\cdot 50\cdot l$, Fig. 8 middle and bottom panels, the range of CV values is still small compared with the CV for the sinc data, see Fig. 5. The surface plots (right middle and bottom panels) do not reveal a small variability (CV) area that would indicate a preference to a ν value. The CV surfaces give the impression of being rather flat in spite of their fluctuations. The corresponding contour plots, Fig. 8 left middle and bottom panels, show a slight horizontal structure for the CV. That is CV seems to change less along the ν axis than along the $2\sigma^2_{\text{kernel}}$ axis. This strengthens the conclusion, that, with respect to the variability of the risk no ν is preferred upon another. The fluctuations of the CV surfaces, and the small range of its values comparing to the sinc data may be due to the relatively small number of trials (100) and the complexity of the particular data set.

The above experiments show that the generalization behavior of the ν-SVM with respect to its parameters is 'well-behaved' even for a complex, real-world data set, as for the Boston Housing data. For this data set we have only assumptions about the intrinsic noise or the effects of rather small test sets. Still, we observe similar risk properties as for the 'toy' sinc data. The risk (generalization error) shows small fluctuations along the ν axis for all parameter settings. This enables us to use the $\nu^{\text{theory}}_{\text{opt}} = 0.54$ in practice. The risk surfaces are largely smooth and flat around the optimal areas giving us much freedom for the choice of the other parameters. As a side effect, the connection between C and σ_{kernel} as regularization (complexity) parameters is obvious even for this complex 'real-world' problem.

7. Discussion

In this paper, we investigated in how far theoretical results on the optimal choice of ν can be used in practical

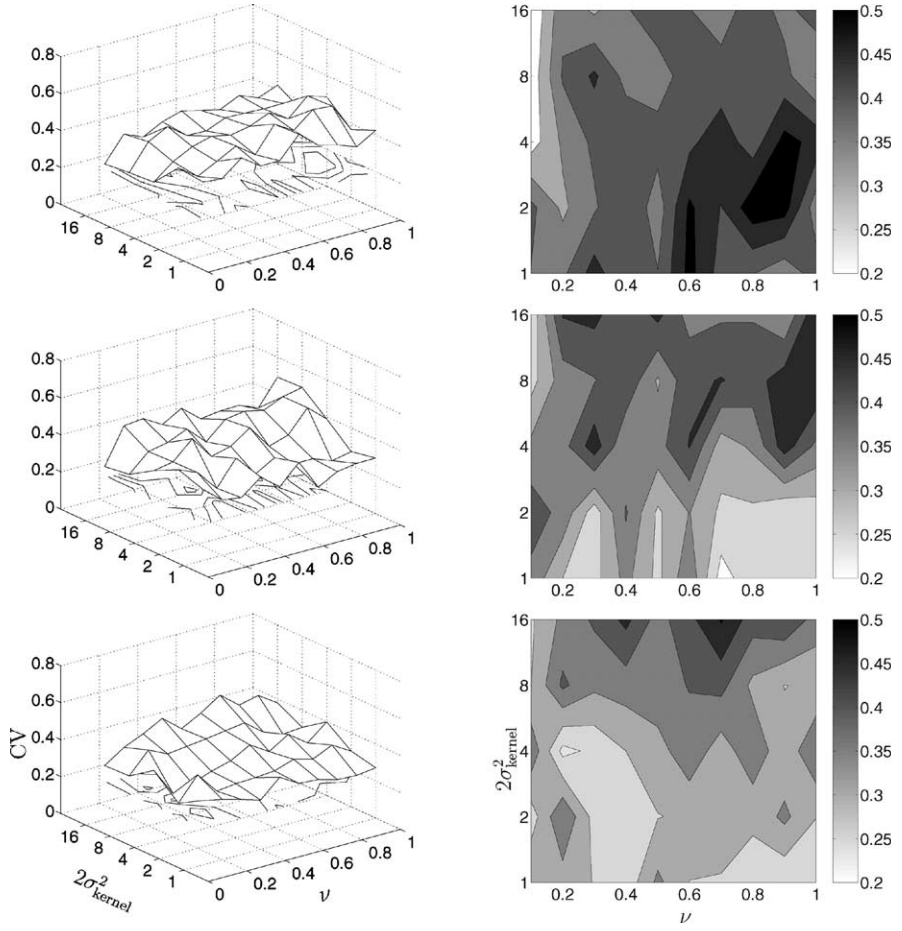

Fig. 7. The coefficient of variation (CV) versus ν and $2\sigma^2_{\text{kernel}}$ for the Boston Housing data as surface (left) and contour plots (right). $\nu \in (0,1]$, $2\sigma^2_{\text{kernel}} \in \{1,2,4,8,16\}$. C changes from top to bottom: $C \in \{50{\cdot}l, 10{\cdot}50{\cdot}l, 100{\cdot}50{\cdot}l\}$. $l = 406$ is the number of training points. The CV versus ν and $2\sigma^2_{\text{kernel}}$ is rather flat showing more variability along the $2\sigma^2_{\text{kernel}}$ than the ν axis.

situations. In this investigation we also obtained important information on the properties on the remaining parameters C, and σ_{kernel} with respect to the generalization properties of the SVM:

1. The theoretically optimal value of ν can be used in practice, since the risk curves are largely smooth and flat, showing a wide optimal area rather than a sharp minimum for ν.
2. The choice of C is not critical, since it only has a significant effect on the risk if changed over orders of magnitude. Generally speaking, it should not be chosen too small, since for too low C the SVM regression function cannot grow enough to reach the output values y.
3. $2\sigma^2_{\text{kernel}}$ is the most sensitive parameter to choose, since it has the strongest influence on the risk.
4. C and $2\sigma^2_{\text{kernel}}$ should not be chose independently. A small C should be accompanied by a small $2\sigma^2_{\text{kernel}}$ and vice versa.

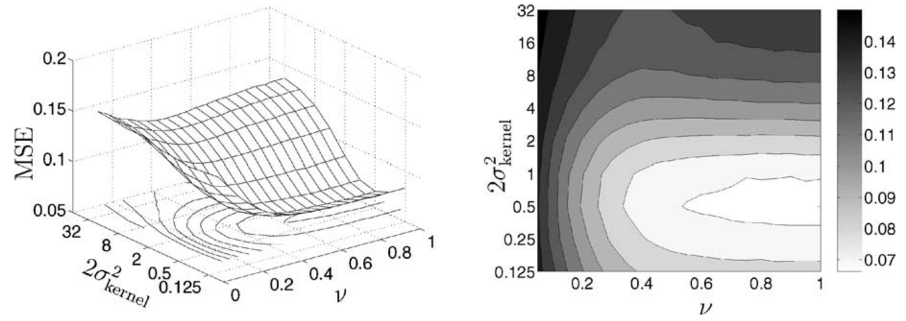

Fig. 8. The sinc function data: surface (left) and contour (right) plots of the risk versus ν and $2\sigma^2_{\text{kernel}}$ for $C = 1.5$ and SNR $= 3$ of the additive Gaussian noise. $C = 1.5$ is the optimal value for this data set according to Cherkassky and Ma (2002). The risk is higher than for $C = 10$ and $C = 100$ for the same SNR. See Fig. 3 second column, first and second panels.

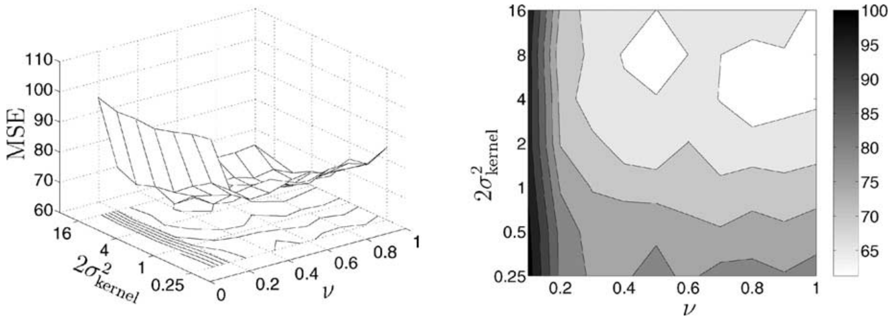

Fig. 9. The Boston Housing data: surface (left) and contour (right) plots of the risk versus ν and $2\sigma^2_{\text{kernel}}$ for $C = 50$, the optimal C value according to Cherkassky and Ma (2002). The risk is much higher than for $C = 50 \cdot l$ or $C = 500 \cdot l$ ($l = 406$ is the number of training data), see Fig. 6, first and second rows.

The optimal choice of SVM parameters is a problem which belongs to the larger context of model selection. While it is beyond the scope of the current paper to discuss the general model selection problem, it is instructive to discuss our findings in the context of recent results on model selection for support vector regression.

For model selection, general methods from statistical inference, like cross validation can be used. Alternatively, one can use results that are specific to statistical learning theory and, thus, are valid only for support vector machines, like the VC bounds on the actual risk. Finally, there are heuristic methods that are intuitively plausible and work well for many practical situations.

7.1. Empirical model selection

Cherkassky and Ma (2002) used heuristic methods for choosing the SVM parameters with good results. Their results are of particular interest, since they considered the same problem as we did, that is, model selection on SV regression with Gaussian kernels. The authors use a similar experimental setup for illustration (the sinc function data in the interval [−10,10] with additive Gaussian noise) and concentrate on the selection of the parameters C and ε of the ε-insensitive loss function (Section 1). They suggest that the third parameter, the kernel width, σ_{kernel}, can be easily chosen, based on the distribution of the $\mathbf{x}$ values of the training data.[5]

A first choice for C, also used in Mattera and Haykin (1999), is to set it equal to the range of response values, y, of the training data. Cherkassky and Ma (2002) propose the following prescription for C instead, to be insensitive towards outliers:

$$C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|), \tag{18}$$

where $\bar{y}$ is the mean and σ_y the standard deviation of the output values.

For ε, the following empirical dependency is proposed:

$$\varepsilon = \tau\sigma\sqrt{\frac{\ln l}{l}}, \tag{19}$$

where $\tau = 3$ is empirically chosen, σ is the standard deviation of the additive noise in the data points, and l is the number of data in the training set.

The fact that ε is proportional to the standard deviation of the noise (also theoretically stated in e.g. Smola et al., 1998; Vapnik, 1995) requires that we have an idea of the intrinsic noise in the data. This, however, is not always possible. In this sense, the parameter ν is more convenient to choose than ε, as it adjusts automatically to the intrinsic noise (Schölkopf et al., 2000). The choice of ν becomes easier, since our experimental study suggests that the risk is not very sensitive to ν. For instance, for Gaussian noise on the data, we can choose ν in the interval [0.3,0.6], depending on how many support vectors we would like to have in the solution.

In the following we will try the model selection procedure suggested by Cherkassky and Ma (2002) on our data sets. We consider 100 points of the sinc function with additive Gaussian noise and signal-to-noise ratio SNR = 3 (see Figs. 3 and 4, column 2 for different C values). SNR = 3 corresponds to $\sigma_{\text{noise}} \approx 0.2$. Using Eq. (19), we find that the optimal value is $\varepsilon = 0.13$ which corresponds to $\nu \approx 0.5$, a value very close to the theoretically optimal value of 0.54 and inside our proposed interval $\nu \in [0.3, 0.6]$. In this case, the method thus works well for choosing ε.The optimal C value according to Cherkassky and Ma (2002) is $C = 1.5$. As our sinc data are in the interval [−3,3], we assume $2\sigma^2_{\text{kernel}} = 2$ to be optimal in the sense of Cherkassky and Ma (2002). We compute the MSE on the test set for $C = 1.5$ and SNR = 3 for $\nu \in (0, 1]$ and $2\sigma^2_{\text{kernel}} \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$. The results show (Fig. 9) that $2\sigma^2_{\text{kernel}} = 0.5$ is optimal for $C = 1.5$. That means that $2\sigma^2_{\text{kernel}} = 2$ is somewhat too large for $C = 1.5$.

For the Boston Housing data, Cherkassky and Ma (2002), propose $C = 50$ as the optimal value. The risk as a function of ν and $2\sigma^2_{\text{kernel}}$ is shown in Fig. 9. Comparing the results to our experiments (see Figs. 6 and 9), we see that the risk for $C = 50$ is not optimal. We reach state of the art

[5] Note that Cherkassky and Ma (2002) are not more specific about the choice of σ_{kernel}. For our purposes, we tried to estimate its value by comparison with the values used by Cherkassky and Ma (2002).

performances (see Schölkopf et al., 2000; Stitson et al., 1999) starting with $C = y_{\max} \cdot l$ where $y_{\max}$ is the maximum output value and $l = 406$ is the number of the training data. Therefore, we conclude that $C = 50$ is too low for the particular data set. As we saw in our experiments (consistent with Cherkassky and Ma, 2002; Schölkopf et al., 2000), when the parameter C is above a certain value, it does not significantly affect the risk. Therefore, the danger is rather to choose a C too low than too high for the problem at hand. In that case, the Lagrange multipliers α, which are upper-bounded by C/l, cannot grow enough for the regression function to reach the output values y.

Our suggestion is to start with $C = y_{\max} \cdot l$. One can even try an order of magnitude lower. For 'easy' data sets, like the sinc data, this C value can work. For the sinc data, $C = 1.5$, proposed by Cherkassky and Ma (2002), works well, two orders of magnitude lower than the suggested $C = 100$. If we carefully compare Fig. 8 to Figs. 3 and 4 for SNR $= 3$, we see that the risk (MSE on the test set) for $C = 1.5$ is slightly higher than for $C = 10$ and $C = 100$.

This discussion shows that the choice of σ_{kernel} is a difficult problem. Moreover, we should not choose σ_{kernel} independently of the C value. There can be several optimal $(C, \sigma_{\text{kernel}})$ pairs, as we saw in our experiments (Figs. 4 and 6). Over at least two orders of magnitude in C the minimum risk value is about the same, with σ_{kernel} moving to larger values with larger C.

The question is how to choose σ_{kernel} effectively for a given C. As we see from our experiments (Figs. 3 and 6), for constant C, the risk versus ν and $2\sigma^2_{\text{kernel}}$ surfaces are convex. For the real-world Boston Housing Problem they show more fluctuations, of course, than for the 'toy' sinc data. We can use the convexity of the surfaces and apply gradient descent methods on a validation set in order to find the optimal risk area. Robust gradient descent methods should not be affected by the starting point in the parameter space or by local fluctuations of the risk.

Up to this point, we discussed only heuristic methods for model selection, based largely on our experimental results. However, since theoretical understanding of our experimental results is still incomplete one should also employ more general methods for model selection.

7.2. Cross validation

Our results can be used for model selection, since it is clear for which parameter values the average optimal risk area is reached (see Figs. 4 and 6). However, the results were averaged over 300 and 100 trials, respectively. These involved considerable computational costs, considering that each trial requires the training of a SVM. Larger data sets may render the computation of the average risk over many trials impossible. In this situation, cross validation may offer a solution.

The general idea of cross validation is to divide the data into training, validation, and test sets. One then estimates the regression function with a given set of parameters on the training set, chooses the best parameter set through the performance on the validation set and test its goodness on the yet unseen test set. There are several variants of the cross validation method, differing in the way they divide the data set into training, validation and test set. Some of these variants can be computationally extremely expensive.

In our context, k-fold cross validation offers the best compromise between computational cost and reliable parameter estimates. It was applied with very good results by Duan, Keerthi, and Poo (2001) in the classification context. In k-fold cross validation the training data set is randomly split into k mutually exclusive subsets (folds) of approximately equal size. We build the regression function with a given set of parameters $\{\nu, C, \sigma_{\text{kernel}}\}$, using the $k-1$ subsets as training set. The performance of the parameter set is measured by the MSE on the last subset. The above procedure is repeated k times, so that each subset is used once for testing. Averaging the MSE over the k trials gives an estimate of the expected generalization error for training on sets of size $\frac{k-1}{k} \cdot l$, l is the number of training data. Finally, one chooses the parameter set which performed best.

Keeping the number of folds, k, moderate ($k = 5$ suggested by Duan et al., 2001) we can keep k-fold cross validation down to reasonable computational costs. One should be careful not to lose this advantage, because of an extensive search in the three dimensional parameter space. The insight we gained from our experiments can help us to find some shortcuts, especially regarding the choice of ν and C. We can then concentrate on the choice of σ_{kernel}. Keeping in mind that the risk surface is convex helps us to decide on the update steps for σ_{kernel} and when to stop the search.

7.3. Theoretical bounds on the risk

Another way of performing model selection, is to use theoretical results specific to statistical learning theory. For example, the expected MSE, $E(y - f(\mathbf{x}, \mathbf{w}))^2$ in the regression case is bounded by the empirical risk and a term depending on h, the VC dimension of the set of the approximating functions (Vapnik, 1998). The bound holds with probability $1 - \eta$ (the confidence level of the bound).

The main difficulty of applying this bound in practice, is to estimate the VC dimension of the set of regression functions. For special regression functions, for example polynomial kernels of degree k, the VC dimension is simply given by $h = k + 1$. For Gaussian kernels, however, we cannot use the bound, as the VC dimension in the space of regression functions is infinite.

In order to overcome this difficulty, Cristianini and Shawe-Taylor (2000) propose a bound on err(f), that is the probability that the linear function f with weight vector $\mathbf{w}$ (in the feature space) has output more than θ away from its true value. Again, it is a probabilistic bound that applies with probability $1 - \eta$. In this bound the VC dimension h is not directly involved. In practice, first one has to choose carefully a number of constants. After training the SVM for a particular parameter set $\{\nu, C, \sigma_{\text{kernel}}\}$ we can compute the bound. The idea is to plot it as function of the parameter of interest, e.g. σ_{kernel}, hoping that it will show a minimum for a particular value. While using this bound for model selection, one has to keep in mind the following: first, err(f) is *not* the expected MSE. Consequently, it may have a behavior with respect to the SVM parameters different from the MSE on the test set, which we widely used as an estimate of the actual risk in the heuristic, empirical and statistical model selection methods described so far. Second, the bound of Cristianini and Shawe-Taylor (2000) as well as bound of Vapnik (1998) are derived based on methods which are known to be somewhat loose[6] and may not have similar behavior with respect to the parameters as the estimated MSE. It is still very interesting to investigate the applicability of theoretical bounds on the risk as they offer a different point of view for performing model selection.

8. Conclusions

In this paper we showed that theoretically optimal values for the parameter ν can also be used in practice, although they have been derived under strong theoretical assumptions that are not satisfied in practical SVMs.

We started reviewing theoretical results on the optimal choice of ν in Section 3. In Section 4, we verified the theoretical results on toy data generated by the sinc function. For general polynomial noise added to the data, one can use the theoretically optimal ν without much effect on the risk. This is due to the fact that the risk versus ν curves are largely flat around their minima showing a whole optimal area for ν rather than a sharp minimum. For the special case of Gaussian noise our experiments show that the optimal ν is rather insensitive towards each of the other SVM parameters, C and σ_{kernel}. From our experiments, we obtained additional information on the behavior of ν concerning the risk (generalization error). The risk versus ν curves are flat and smooth, indicating that there is an optimal area rather than a sharp optimum for ν. That enables us to use smaller or larger ν values than the theoretical optimum, without much effect on the risk. Varying ν directly affects the number of support vectors that build the solution of the learning problem at hand. At this point we note that we mainly used the MSE on the test set as a reliable representation of the actual risk. Although it is quite different from the ε-insensitive loss (Section 1) that we used during the training of the SVM, the MSE is widely used for measuring the performance of learning methods.

In Section 5 we extended our experiments with data from the sinc function with additive Gaussian noise while varying all degrees of freedom at the same time. The risk surfaces showed for different C values and noise levels similar behavior to the risk versus ν curves in Section 4. The error curves are convex, flat, and there is a large optimal ν area around the theoretical optimal ν value of 0.54 for Gaussian noise. Therefore, we can use the theoretically optimal $\nu_{\text{opt}}^{\text{theory}}$ in practice for almost any parameter setting (avoiding extreme situations, of course), even in the case that our data are completely masked by noise.

Moreover, from the extended experiments of Section 5, we were able to obtain additional information on the behavior of the other two SVM parameters, $2\sigma_{\text{kernel}}^2$ and C as well as their dependencies. For increasing C, the optimal error area moves from smaller to larger $2\sigma_{\text{kernel}}^2$ values, thus out-weighting the weak regularization by a large C value. This effect seems largely independent of the noise level.

In Section 6, we examined whether our results still hold when we deal with real world data. For this purpose we used the widely used Boston Housing benchmark. As for the sinc function data, we notice that for larger C, good regularization is restored through a wider Gaussian kernel. The experimentally optimal ν values, however, are smaller than for the sinc data with additive Gaussian noise. This may be due to shorter than Gaussian tailed noise in the data. Still, the optimal ν areas are largely insensitive towards the C and $2\sigma_{\text{kernel}}^2$ values, even for this complex real world problem. We could, however, use the theoretically optimal $\nu_{\text{opt}}^{\text{theory}} = 0.54$ value since the risk does not vary much along the ν axis. Even for this data set, the risk shows only small deviations around the optimal areas but it still seems largely flat and smooth.

We reached our conclusions from looking at trial averages as well as the corresponding inter-trial variability. The number of trials as well as the amount of noise in the data determine the degree of variability. More important for our conclusions is that the dependence of the variability on ν is generally low and in most cases negligible.

In Section 7 we embedded our results on the behavior of the SV parameters in the more general discussion of model selection. We first tested heuristic methods proposed in the literature followed by a brief review of cross validation and theoretical bounds on the risk along with a discussion of their applicability. The comparison with heuristic methods confirmed our results. It became again obvious that the kernel width is the most sensitive parameter and should be chosen in accordance with C.

The theoretical values of ν are mostly useful in practical applications where we have some knowledge of the distribution of the noise. This restriction may not be as severe as it seems: our experiments show that the risk versus ν curves are rather smooth and flat around their minima. The fact that the curves are flat allows us to choose a ν which is slightly off the optimal value without sacrificing too much accuracy. This is also useful when the number of support vectors is of importance. Smaller ν leads to wider tubes and fewer support vectors, i.e. larger data compression. The flatness of the curves indicates that the ν-SVM is insensitive with respect to ν. Our experiments, therefore, support the view that ν is a well-behaved parameter which is easy to use in practice.

[6] N. Cristianini, private communication.

References

Cherkassky, V., & Ma, Y. (2002). *Selection of meta-parameters for support vector regression. In Proceedings of the International Conference on Artificial Neural Networks (ICANN), Madrid Spain*.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge University Press.

Duan, K., Keerthi, S. & Poo, A (2001). *Evaluation of simple performance measures for tuning SVM hyperparameters* (Tech. Rep. No. Control Division Technical Report CD-01-11). Department of Mechanical Engineering, National University of Singapore.

Mattera, D., & Haykin, S. (1999). Support vector machines for dynamic reconstruction of a chaotic system. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—Support vector learning* (pp. 211–241). MIT Press.

Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information criterion—Determining the number of hidden units for artificial neural networks. *IEEE Transactions on Neural Networks*, *5*, 865–872.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neutral Computation*, *12*(4), 1207–1245.

Smola, A. J., Murata, N., Schölkopf, B., & Müller, K.-R. (1998). Asymptotically optimal choice of ε-loss for support vector machines. In L. Niklasson, M. Bodèn, & T. Ziemke (Eds.), *Perspectives in neural computing* (pp. 105–110). *Proceedings of the Eighth International Conference on Artificial Neural Networks*, Springer Verlag.

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., & Weston, J. (1999). Support vector regression with ANOVA decomposition kernels. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—Support vector learning* (pp. 285–291). MIT Press.

Vanderbei, R. J (1994). *An interior point code for quadratic programming* (Tech. Rep. Nos. TR SOR-94-15, Statistics and Operations Research). Princeton University, NJ.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.