# Linear Dependency Between $\epsilon$ and the Input Noise in $\epsilon$-Support Vector Regression

James T. Kwok, Ivor W. Tsang

***Abstract*— In using the $\epsilon$-support vector regression ($\epsilon$-SVR) algorithm, one has to decide a suitable value for the insensitivity parameter $\epsilon$. Smola *et al.* considered its "optimal" choice by studying the statistical efficiency in a location parameter estimation problem. While they successfully predicted a linear scaling between the optimal $\epsilon$ and the noise in the data, their theoretically optimal value does not have a close match with its experimentally observed counterpart in the case of Gaussian noise. In this paper, we attempt to better explain their experimental results by studying the regression problem itself. Our resultant predicted choice of $\epsilon$ is much closer to the experimentally observed optimal value, while again demonstrating a linear trend with the input noise.**

***Index Terms*— Support vector machines, Support vector regression.**

## I. INTRODUCTION

IN recent years, the use of support vector machines (SVMs) on various classification and regression problems have been increasingly popular. SVMs are motivated by results from statistical learning theory and, unlike other machine learning methods, their generalization performance does not depend on the dimensionality of the problem [3], [24]. In this paper, we focus on regression problems and consider the $\epsilon$-support vector regression ($\epsilon$-SVR) algorithm [4], [20] in particular. The $\epsilon$-SVR has produced the best result on a timeseries prediction benchmark [11], as well as showing promising results in a number of different applications [7], [12], [22].

One issue about $\epsilon$-SVR is on how to set the insensitivity parameter $\epsilon$. Data-resampling techniques such as cross-validation can be used [11], though they are usually very expensive in terms of computation and/or data. A more efficient approach is to use a variant of the SVR algorithm called $\nu$-support vector regression ($\nu$-SVR) [17]. By using another parameter $\nu$ to trade-off $\epsilon$ with model complexity and training accuracy, $\nu$-SVR allows the value of $\epsilon$ to be automatically determined. Moreover, it can be shown that $\nu$ represents an upper bound on the fraction of training errors and a lower bound on the fraction of support vectors. Thus, in situations where some prior knowledge on the value of $\nu$ is available, using $\nu$-SVR may be more convenient than $\epsilon$-SVR. Another approach is to consider the theoretically "optimal" choice of $\epsilon$. Smola *et al.* [19] tackled this by studying the simpler location parameter estimation problem, and derived the asymptotically optimal choice of $\epsilon$ by maximizing statistical efficiency. They also showed that this optimal value scales linearly with the noise in the data, which is confirmed in the experiment. However, in the case of Gaussian noise, their predicted value of this optimal $\epsilon$ does not have a close match with their experimentally observed value.

J.T. Kwok and I.W. Tsang are with the Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. Email: [jamesk,ivor]@cs.ust.hk.

In this paper, we attempt to better explain their experimental results. Instead of working on the location parameter estimation problem as in [19], our analysis will be based on the original $\epsilon$-SVR formulation. The rest of this paper is organized as follows. Brief introduction to the $\epsilon$-SVR is given in Section II. The analysis of the linear dependency between $\epsilon$ and the input noise level is given in Section III, while the last section gives some concluding remarks.

## II. $\epsilon$-SUPPORT VECTOR REGRESSION

In this section, we introduce some basic notations for $\epsilon$-SVR. Interested readers are referred to [3], [20], [24] for more complete reviews.

Let the training set $D$ be $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, with input $\mathbf{x}_i \in \Re^n$ and output $y_i \in \Re$. In $\epsilon$-SVR, $\mathbf{x}$ is first mapped to $\mathbf{z} = \psi(\mathbf{x})$ in a Hilbert space $\mathcal{F}$ (with inner product $\langle\cdot,\cdot\rangle$) via a nonlinear map $\psi : \Re^n \to \mathcal{F}$. This space $\mathcal{F}$ is often called the *feature space*, and its dimensionality is usually very high (sometimes infinite). Then, a linear function $f(\mathbf{x}) = \langle \mathbf{w}, \psi(\mathbf{x})\rangle + b$ is constructed in $\mathcal{F}$ such that it deviates least from the training data according to Vapnik's $\epsilon$-insensitive loss function

$$|y - f(\mathbf{x})|_\epsilon = \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \le \epsilon, \\ |y - f(\mathbf{x})| - \epsilon & \text{otherwise,} \end{cases}$$

while at the same time is as "flat" as possible (i.e., $\|\mathbf{w}\|$ is as small as possible). Mathematically, this means:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^N(\xi_i + \xi_i^*), \qquad (1)$$

subject to

$$\begin{cases} y_i - f(\mathbf{x}_i) \le \epsilon + \xi_i, \\ f(\mathbf{x}_i) - y_i \le \epsilon + \xi_i^*, \; i = 1, \ldots, N, \\ \xi_i, \xi_i^* \ge 0, \end{cases}$$

where $C$ is a user-defined constant. It is well-known that (1) can be transformed to the following quadratic programming (QP) problem:

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^N (\gamma_i - \gamma_i^*)(\gamma_j - \gamma_j^*)\langle\psi(\mathbf{x}_i), \psi(\mathbf{x}_j)\rangle \\ & -\epsilon\sum_{i=1}^N(\gamma_i + \gamma_i^*) + \sum_{i=1}^N y_i(\gamma_i - \gamma_i^*), \end{aligned} \qquad (2)$$

subject to

$$\sum_{i=1}^{N}(\gamma_i - \gamma_i^*) = 0, \quad \text{and} \quad \gamma_i, \gamma_i^* \in [0, C].$$

However, recall that the dimensionality of $\mathcal{F}$, and thus also of $\psi(\mathbf{x}_i)$ and $\psi(\mathbf{x}_j)$, is usually very high. Hence, in order for this approach to be practical, a key characteristic of $\epsilon$-SVR, and kernel methods in general, is that one can obtain $\langle\psi(\mathbf{x}_i), \psi(\mathbf{x}_j)\rangle$ in (2) without having to explicitly obtain $\psi(\mathbf{x}_i)$ and $\psi(\mathbf{x}_j)$ first. This is achieved by using a *kernel* function $K(\cdot,\cdot)$ such that

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle\psi(\mathbf{x}_i), \psi(\mathbf{x}_j)\rangle. \tag{3}$$

For example, the $d$-order polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j)^d$ corresponds to a map $\psi$ into the space spanned by all products of exactly $d$ order of $\Re^n$ [24]. More generally, it can be shown that any function satisfying Mercer's theorem can be used as kernel, and each will have an associated map $\psi$ such that (3) holds [24]. Computationally, $\epsilon$-SVR (and kernel methods in general) also has the important advantage that only quadratic programming[1], and not nonlinear optimization, is involved. Thus, the use of kernels provides an elegant nonlinear generalization of many existing linear algorithms [2], [3], [10], [16].

## III. Linear Dependency Between $\epsilon$ and the Input Noise Parameter

In order to derive a linear dependency between $\epsilon$ and the scale parameter of the input noise model, Smola *et al.* [19] considered the simpler problem of estimating the univariate location parameter $w$ from a set of data points

$$y_i = w + \eta_i, \quad i = 1, \ldots, N.$$

Here, $\eta_i$'s are i.i.d. noise belonging to some distribution $\phi(\cdot)$. Using the Cramer-Rao information inequality for unbiased estimators, the maximum likelihood estimator of $w$ with an "optimal" value of $\epsilon$ was then obtained by maximizing its statistical efficiency.

In this paper, instead of working on the location parameter estimation problem, we study the regression problem of estimating the (possibly multivariate) weight parameter $\tilde{\mathbf{w}}$ given a set $D$ of $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, with

$$y_i = \tilde{\mathbf{w}}^T\mathbf{x}_i + \eta_i, \quad i = 1, \ldots, N. \tag{4}$$

Here, $\mathbf{x}_i \in \Omega$ follows distribution $p(\cdot)$, and $\eta_i$ follows distribution $\phi(\cdot)$. The corresponding density function on $y$ is denoted $p(y|\mathbf{x}) = \phi(y - \tilde{\mathbf{w}}^T\mathbf{x})$. Notice that the bias term has been dropped here for simplicity, and this is equivalent to assuming that the $\mathbf{x}_i$'s have zero mean. Moreover, using the notation in Section II, in general one can replace $\mathbf{x}_i$ in (4) by $\mathbf{z}_i = \psi(\mathbf{x}_i)$ in feature space $\mathcal{F}$ and thus recovers the original $\epsilon$-SVR setting.

Besides, while the work in [19] is based on maximum likelihood estimation, it is now well-known that $\epsilon$-SVR is related instead to *maximum a posteriori* (MAP) estimation. As discussed in [14], [20], [21], the $\epsilon$-insensitive loss function leads to the following probability density function on $y$,

$$p(y_i|\mathbf{x}_i, \mathbf{w}, \beta, \epsilon) = \frac{\beta}{2(1+\epsilon\beta)}\exp(-\beta|y_i - \mathbf{w}^T\mathbf{x}_i|_\epsilon). \tag{5}$$

Notice that [14], [20], [21] do not have the factor $\beta$ in (5), but is introduced here to play the role of controlling the noise level[2] [6]. With the Gaussian prior[3] on $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \sqrt{\frac{\alpha}{2\pi}}\exp(-\frac{\alpha}{2}\|\mathbf{w}\|^2),$$

and on applying the Bayes rule, $p(\mathbf{w}|D, \beta, \epsilon) \propto p(D|\mathbf{w}, \beta, \epsilon)p(\mathbf{w})$, we obtain

$$\begin{aligned}\log p(\mathbf{w}|D, \beta, \epsilon) &= -\frac{\alpha}{2}\|\mathbf{w}\|^2 - \beta\sum_{i=1}^N |y_i - \mathbf{w}^T\mathbf{x}_i|_\epsilon \\ &\quad + N\log\frac{\beta}{2(1+\epsilon\beta)} + \text{const.}\end{aligned} \tag{6}$$

On setting $C = \beta/\alpha$, the optimization problem in (1) can be interpreted as finding the MAP estimate of $\mathbf{w}$ at given values of $\beta$ and $\epsilon$.

### A. Estimating the Optimal Values of $\beta$ and $\epsilon$

In general, the MAP estimate $\hat{\mathbf{w}}$ in (6) depends on the particular training set, and a closed-form solution is difficult to obtain. To simplify the analysis, we replace $\frac{1}{N}\sum_{i=1}^N |y_i - \mathbf{w}^T\mathbf{x}_i|_\epsilon$ in (6) by its expectation

$$\begin{aligned}&E_{XY}(|y - \mathbf{w}^T\mathbf{x}|_\epsilon) \\ &= \int_\Omega \int_{-\infty}^{\infty} |y - \mathbf{w}^T\mathbf{x}|_\epsilon \, p(y|\mathbf{x})p(\mathbf{x})\, dy\, d\mathbf{x} \\ &= \int_\Omega \left( \int_{-\infty}^{\mathbf{w}^T\mathbf{x}-\epsilon} (\mathbf{w}^T\mathbf{x} - \epsilon - y)p(y|\mathbf{x})dy \right. \\ &\quad \left. + \int_{\mathbf{w}^T\mathbf{x}+\epsilon}^{\infty} (y - \mathbf{w}^T\mathbf{x} - \epsilon)p(y|\mathbf{x})dy \right) p(\mathbf{x})d\mathbf{x}.\end{aligned}$$

(6) thus becomes

$$\begin{aligned}M(\mathbf{w}, \beta, \epsilon) &= -\frac{\alpha}{2}\|\mathbf{w}\|^2 - \beta N E_{XY}(|y - \mathbf{w}^T\mathbf{x}|_\epsilon) \\ &\quad + N\log\frac{\beta}{2(1+\epsilon\beta)} + \text{const.}\end{aligned} \tag{7}$$

On setting its partial derivative w.r.t. $\mathbf{w}$ to zero, it can be shown that $\hat{\mathbf{w}}$ has to satisfy

$$\begin{aligned}\alpha\hat{\mathbf{w}} + \beta N \int_\Omega \mathbf{x} &\left( \int_{-\infty}^{\hat{\mathbf{w}}^T\mathbf{x}-\epsilon} p(y|\mathbf{x})dy \right. \\ &\left. - \int_{\hat{\mathbf{w}}^T x+\epsilon}^{\infty} p(y|\mathbf{x})dy \right) p(\mathbf{x})d\mathbf{x} = 0.\end{aligned} \tag{8}$$

[1]The SVM problem can also be formulated as a linear programming problem [9], [18] instead of a quadratic programming problem.

[2]To be more precise, we will show in Section III-B that the optimal value of $\beta$ is inversely proportional to the noise level of the input noise model.

[3]In this paper, $\beta$ and $\epsilon$ are regarded as hyperparameters while $\alpha$ is treated as a constant. Thus, dependence on $\alpha$ will not be explicitly mentioned in the sequel.

To find suitable values for the hyperparameters $\beta$ and $\epsilon$, we use the method of type II maximum likelihood prior[4] (ML-II) in Bayesian statistics [1]. The ML-II solutions of $\beta$ and $\epsilon$ are those that maximize $p(D|\beta,\epsilon)$, which can be obtained by integrating out $\mathbf{w}$:

$$\begin{aligned} p(D|\beta,\epsilon) &= \int p(\mathbf{w},D|\beta,\epsilon)d\mathbf{w} \\ &= \int p(D|\mathbf{w},\beta,\epsilon)p(\mathbf{w})d\mathbf{w} \\ &\simeq p(D|\hat{\mathbf{w}},\beta,\epsilon)p(\hat{\mathbf{w}})\Delta\mathbf{w}, \end{aligned}$$

where $\Delta\mathbf{w}$ is the width of the integrand $p(D|\mathbf{w},\beta,\epsilon)p(\mathbf{w})$ and measures the posterior uncertainty of $\mathbf{w}$. Assuming that the contributions due to $\Delta\mathbf{w}$ are comparable at different values of $\beta$ and $\epsilon$, maximizing $p(D|\beta,\epsilon)$ w.r.t. $\beta$ and $\epsilon$ thus becomes maximizing $p(D|\hat{\mathbf{w}},\beta,\epsilon)p(\hat{\mathbf{w}})$. This is the same as maximizing $\log p(\hat{\mathbf{w}}|D,\beta,\epsilon)$, or $M(\hat{\mathbf{w}},\beta,\epsilon)$ approximately. Differentiating (7) w.r.t. $\epsilon$ and $\beta$ and then using (8), we have

$$\begin{aligned} &\frac{\partial M(\hat{\mathbf{w}},\beta,\epsilon)}{\partial\epsilon} \\ &= -\Bigg(\alpha\hat{\mathbf{w}} + N\beta\int_\Omega \mathbf{x}\Bigg(\int_{-\infty}^{\hat{\mathbf{w}}^T\mathbf{x}-\epsilon} p(y|\mathbf{x})dy \\ &\quad -\int_{\hat{\mathbf{w}}^T\mathbf{x}+\epsilon}^{\infty} p(y|\mathbf{x})dy\Bigg)p(\mathbf{x})d\mathbf{x}\Bigg)^T \frac{\partial\hat{\mathbf{w}}}{\partial\epsilon} - \frac{N\beta}{1+\epsilon\beta} \\ &\quad +N\beta\int_\Omega\Bigg(\int_{-\infty}^{\hat{\mathbf{w}}^T\mathbf{x}-\epsilon} p(y|\mathbf{x})dy \\ &\quad +\int_{\hat{\mathbf{w}}^T\mathbf{x}+\epsilon}^{\infty} p(y|\mathbf{x})dy\Bigg)p(\mathbf{x})d\mathbf{x} \\ &= -\frac{N\beta}{1+\epsilon\beta} + N\beta\int_\Omega\Bigg(\int_{-\infty}^{\hat{\mathbf{w}}^T\mathbf{x}-\epsilon} p(y|\mathbf{x})dy \\ &\quad +\int_{\hat{\mathbf{w}}^T\mathbf{x}+\epsilon}^{\infty} p(y|\mathbf{x})dy\Bigg)p(\mathbf{x})d\mathbf{x}, \end{aligned} \tag{9}$$

and

$$\begin{aligned} &\frac{\partial M(\hat{\mathbf{w}},\beta,\epsilon)}{\partial\beta} \\ &= -\Bigg(\alpha\hat{\mathbf{w}} + N\beta\int_\Omega \mathbf{x}\Bigg(\int_{-\infty}^{\hat{\mathbf{w}}^T\mathbf{x}-\epsilon} p(y|\mathbf{x})dy \\ &\quad -\int_{\hat{\mathbf{w}}^T\mathbf{x}+\epsilon}^{\infty} p(y|\mathbf{x})dy\Bigg)p(\mathbf{x})d\mathbf{x}\Bigg)^T \frac{\partial\hat{\mathbf{w}}}{\partial\beta} \\ &\quad +\frac{N}{\beta(1+\epsilon\beta)} - NE_{XY}(|y-\hat{\mathbf{w}}^T\mathbf{x}|_\epsilon) \\ &= \frac{N}{\beta(1+\epsilon\beta)} - NE_{XY}(|y-\hat{\mathbf{w}}^T\mathbf{x}|_\epsilon), \end{aligned} \tag{10}$$

[4]Notice that MacKay's evidence framework [8], which has been popularly used in the neural networks community, is computationally equivalent to the ML-II method.

respectively. Setting (9) and (10) to zero, we obtain

$$\begin{aligned} &\frac{1}{1+\epsilon\beta} \\ &= \int_\Omega\left(\int_{-\infty}^{\hat{\mathbf{w}}^T\mathbf{x}-\epsilon} p(y|\mathbf{x})dy + \int_{\hat{\mathbf{w}}^T\mathbf{x}+\epsilon}^{\infty} p(y|\mathbf{x})dy\right)p(\mathbf{x})d\mathbf{x} \\ &= \int_\Omega\int_{y\notin[\hat{\mathbf{w}}^T\mathbf{x}-\epsilon,\hat{\mathbf{w}}^T\mathbf{x}+\epsilon]} p(\mathbf{x},y)\,dy\,d\mathbf{x}, \end{aligned} \tag{11}$$

and

$$\frac{1}{\beta(1+\epsilon\beta)} = E_{XY}(|y-\hat{\mathbf{w}}^T\mathbf{x}|_\epsilon). \tag{12}$$

These can then be used to solve for $\beta$ and $\epsilon$.

### *B. Applications to Some Common Noise Models*

Recall that the $\epsilon$-insensitive loss function implicitly corresponds to the noise model in (5). Of course, in cases where the underlying noise model of a particular data set is known, the loss function should be chosen such that it has a close match with this known noise model, while at the same time ensuring that the resultant optimization problem can still be solved efficiently. It is thus possible that the $\epsilon$-insensitive loss function may not be best. For example, when the noise model is known to be Gaussian, the corresponding loss function is the squared loss function and the resultant model becomes a regularization network [5], [13][5]. However, an important advantage of $\epsilon$-SVR using the $\epsilon$-insensitive loss function, just like SVMs using the hinge loss in classification problems, is that sparseness of the dual variables can be ensured. On the contrary, sparseness will be lost if the squared loss function is used instead. Thus, even for Gaussian noise, the $\epsilon$-insensitive loss function is sometimes still desirable and the resultant performance is often very competitive. In this Section, by solving equations (11) and (12), we will show that there is always a linear relationship between the optimal value of $\epsilon$ and the noise level, even when the true noise model is different from (5). As in [19], three commonly used noise models, namely the Gaussian, Laplacian and uniform models, will be studied.

*1) Gaussian Noise:* For the Gaussian noise model,

$$\phi(\eta) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\eta^2}{2\sigma^2}),$$

where $\sigma$ is the standard deviation. Define $\delta(\mathbf{x}) = (\hat{\mathbf{w}} - \tilde{\mathbf{w}})^T\mathbf{x}, b_1(\mathbf{x}) = \frac{1}{\sqrt{2}}(\frac{\epsilon}{\sigma} - \frac{\delta(\mathbf{x})}{\sigma})$ and $b_2(\mathbf{x}) = \frac{1}{\sqrt{2}}(\frac{\epsilon}{\sigma} + \frac{\delta(\mathbf{x})}{\sigma})$. It can be shown that (11) and (12) reduce to

$$\frac{1}{1+\epsilon\beta} = \int_\Omega \frac{\mathrm{erfc}(b_1)+\mathrm{erfc}(b_2)}{2}p(\mathbf{x})d\mathbf{x}, \tag{13}$$

and

$$\begin{aligned} \frac{1}{\beta(1+\epsilon\beta)} &= \int_\Omega\Bigg(-\frac{\sigma b_1}{\sqrt{2}}\mathrm{erfc}(b_1) - \frac{\sigma b_2}{\sqrt{2}}\mathrm{erfc}(b_2) \\ &\quad +\frac{\sigma}{\sqrt{2\pi}}\exp(-b_1^2) + \frac{\sigma}{\sqrt{2\pi}}\exp(-b_2^2)\Bigg) \\ &\quad p(\mathbf{x})d\mathbf{x}, \end{aligned} \tag{14}$$

[5]This is sometimes also called the least squares support vector machine [23] in the SVM literature.

where $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty}\exp(-t^2)dt$ is the complementary error function [15]. Substituting (13) into (14), and after performing the integration[6], it can be shown that $\epsilon$ always appears as the factor $\epsilon/\sigma$ in the solution and thus there is always a linear relationship between the optimal values of $\epsilon$ and $\sigma$.

If we assume that

$$\begin{aligned} E_X(\delta^2(\mathbf{x})) &= \int_\Omega \delta^2(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= E_X((\hat{\mathbf{w}}^T\mathbf{x} - \tilde{\mathbf{w}}^T\mathbf{x})^2) \\ &\simeq E_X((y - \tilde{\mathbf{w}}^T\mathbf{x})^2) = \sigma^2, \end{aligned} \quad (15)$$

and also that the variation of $\hat{\mathbf{w}}$ with $\epsilon$ is small[7], then maximizing $M$ in (7) is effectively the same as minimizing $\beta E_{XY}(|y - \mathbf{w}^T\mathbf{x}|_\epsilon) - \log\frac{\beta}{2(1+\epsilon\beta)}$, which is the same as minimizing[8]

$$\begin{aligned} h(\frac{\epsilon}{\sigma}) &= \sqrt{2\pi}\exp(\frac{\epsilon^2}{2\sigma^2})((\frac{\epsilon}{\sigma})^2+3)^{-1} \\ &\quad \left(-\frac{\epsilon}{\sigma}\mathrm{erfc}(\frac{\epsilon}{\sqrt{2}\sigma}) + \frac{3}{\sqrt{2\pi}}\exp(-\frac{\epsilon^2}{2\sigma^2})\right) \\ &\quad -\log\left(\sqrt{2\pi}\exp(\frac{\epsilon^2}{2\sigma^2})((\frac{\epsilon}{\sigma})^2+3)^{-1}\right) \\ &\quad -\log\left(\mathrm{erfc}(\frac{\epsilon}{\sqrt{2}\sigma}) + \frac{\epsilon}{\sqrt{2\pi}\sigma}\exp(-\frac{\epsilon^2}{2\sigma^2})\right) \end{aligned} \quad (16)$$

A plot of $h(\frac{\epsilon}{\sigma})$ is shown in Figure 1. Its minimum can be obtained numerically as

$$\epsilon = 1.0043\sigma. \quad (17)$$

By substituting (15), (17) into (29), we can also obtain the optimal value of $\beta$ as $1/\beta = 0.9658\sigma$. This confirms the role of $\beta$ in controlling the noise variance, as mentioned in Section III.

In the following, we repeat an experiment in [19] to verify this ratio of $\epsilon/\sigma$. The target function to be learned is $f(x) = 0.9\mathrm{sinc}(10x/\pi)$. The training set consists of 200 $x$'s drawn independently from a uniform distribution on $[-1, 1]$, and the corresponding outputs have Gaussian noise $N(0, \sigma^2)$ added. Testing error is obtained by directly integrating the squared difference between the true target function and the estimated function over the range $[-1, 1]$. As in [19], the model selection problem for the regularization parameter $C$ in (1) is side-stepped by always choosing the value of $C$ that yields the smallest testing error. The whole experiment is repeated 40 times, and the spline kernel with an infinite number of knots is used.

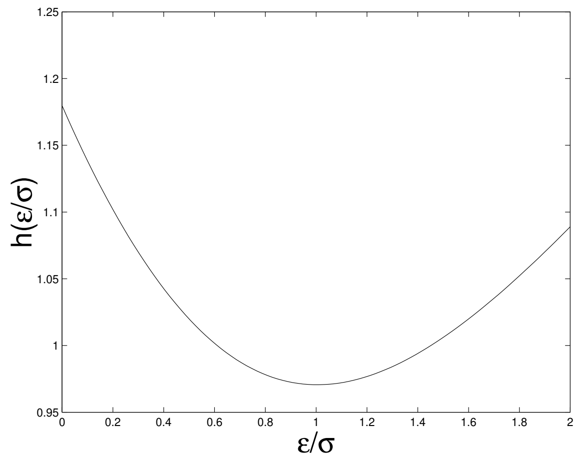



Fig. 1. The objective function $h(\frac{\epsilon}{\sigma})$ to be minimized in the case of Gaussian noise.

Figure 2 shows the linear relationship of $\epsilon = 0.9846\sigma$ obtained from the experiment. This is close to the value of $\epsilon = 0.9\sigma$ obtained in the experiment in [19], and is also in good agreement with our predicted ratio of $\epsilon = 1.0043\sigma$. In comparison, the theoretical results in [19] suggest the "optimal" choice of $\epsilon = 0.6166\sigma$.

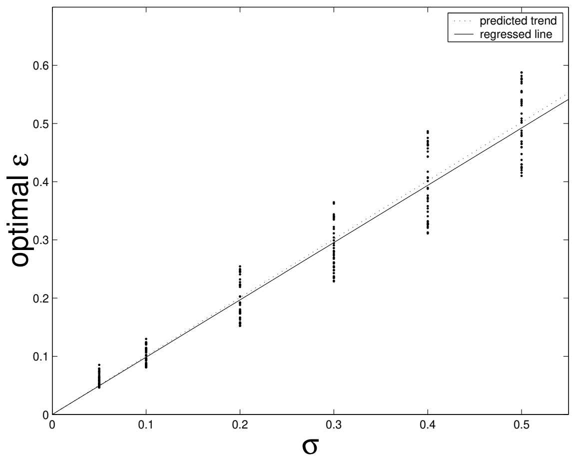



Fig. 2. Relationship between the optimal value of $\epsilon$ and the Gaussian noise level $\sigma$.

*2) Laplacian Noise:* Next, we consider the Laplacian noise model

$$\phi(\eta) = \frac{1}{2\sigma}\exp(-\frac{|\eta|}{\sigma}).$$

Here, we only consider the simpler case when $\mathbf{x}$ is one-dimensional, with uniform density over the range $[-L, L]$ (i.e., $X \sim U([-L, L])$). After tedious computation, it can be shown that (11) and (12) reduce to

$$\begin{aligned} &\frac{1}{1+\epsilon\beta} \\ &= 1 - \frac{\epsilon}{L(\tilde{w}-\hat{w})} + \frac{\sigma}{2L(\tilde{w}-\hat{w})}\exp\left(-\frac{L(\tilde{w}-\hat{w})}{\sigma}\right) \end{aligned}$$

[6]Here, the integration is performed by using the symbolic math toolbox of Matlab.

[7]Notice that from (8), we have

$$\begin{aligned} &\frac{\partial\hat{\mathbf{w}}}{\partial\epsilon} \\ &= \left(\frac{\alpha}{\beta N}\mathbf{I} + \int_\Omega \mathbf{x}\mathbf{x}^T(p(\hat{\mathbf{w}}^T\mathbf{x} - \epsilon|\mathbf{x}) + p(\hat{\mathbf{w}}^T\mathbf{x} + \epsilon|\mathbf{x}))p(\mathbf{x})d\mathbf{x}\right)^{-1} \\ &\quad \int_\Omega \mathbf{x}\left(p(\hat{\mathbf{w}}^T\mathbf{x} - \epsilon|\mathbf{x}) - p(\hat{\mathbf{w}}^T\mathbf{x} + \epsilon|\mathbf{x})\right)p(\mathbf{x})d\mathbf{x}, \end{aligned}$$

where $\mathbf{I}$ is the identity matrix. $\partial\hat{\mathbf{w}}/\partial\epsilon$ thus involves the difference of two very similar terms and will be small.

[8]More detailed derivations are in the Appendix.

$$\left(\exp(\frac{\epsilon}{\sigma}) - \exp(-\frac{\epsilon}{\sigma})\right), \tag{18}$$

and

$$\begin{aligned} \frac{1}{\beta} \quad = \quad & \frac{\sigma^2}{L(\tilde{w}-\hat{w})} + \frac{\sigma}{2L(\tilde{w}-\hat{w})} \exp\left(-\frac{L(\tilde{w}-\hat{w})}{\sigma}\right) \\ & \left((\epsilon-\sigma)\exp(\frac{\epsilon}{\sigma}) - (\epsilon+\sigma)\exp(-\frac{\epsilon}{\sigma})\right) \\ & + \frac{(\tilde{w}-\hat{w})L}{2} - \frac{\epsilon^2}{2L(\tilde{w}-\hat{w})}. \end{aligned} \tag{19}$$

Substituting (19) back into (18), we notice again that $\epsilon$ always appears in the factor $\epsilon/\sigma$ and thus there is always a linear relationship between the optimal value of $\epsilon$ and $\sigma$ under the Laplacian noise.

Recall that the variance for the uniform distribution $U(a,b)$ is $(b-a)^2/12$ and that for the exponential distribution[9] $Expon(\beta)$ is $1/\beta^2$. As in Section III-B.1, we consider $\delta(x) = (\hat{w}-\tilde{w})x$ and assume that $E_X(\delta^2(x)) \simeq E_X((y-\hat{w}x)^2) = 2\sigma^2$. With $X \sim U([-L,L])$, we also obtain $var(X) = L^2/3$. Consequently

$$E_X(\delta^2(x)) \simeq var((\tilde{w}-\hat{w})X) \simeq L^2(\tilde{w}-\hat{w})^2/3, \tag{20}$$

and hence

$$\sigma = L(\tilde{w}-\hat{w})/\sqrt{6}. \tag{21}$$

Analogous to Section III-B.1, by plugging in (18), (19) and (21), it can be shown that the problem reduces to minimizing (Figure 3)

$$\begin{aligned} & h(\frac{\epsilon}{\sigma}) \\ = \quad & \frac{-\frac{\epsilon}{\sigma} + \frac{1}{\sqrt{6}}((\frac{\epsilon}{\sigma})^2+1) - \frac{\exp(-\sqrt{6})}{2\sqrt{6}}(\exp(\frac{\epsilon}{\sigma}) + \exp(-\frac{\epsilon}{\sigma})) + \frac{\sqrt{6}}{2} - \frac{1}{2\sqrt{6}}(\frac{\epsilon}{\sigma})^2}{\frac{1}{\sqrt{6}} + \frac{\exp(-\sqrt{6})}{2\sqrt{6}}((\frac{\epsilon}{\sigma}-1)\exp(\frac{\epsilon}{\sigma}) - (\frac{\epsilon}{\sigma}+1)\exp(-\frac{\epsilon}{\sigma})) + \frac{\sqrt{6}}{2} - \frac{1}{2\sqrt{6}}(\frac{\epsilon}{\sigma})^2} \\ & + \log\left(\frac{1}{\sqrt{6}} + \frac{\exp(-\sqrt{6})}{2\sqrt{6}}((\frac{\epsilon}{\sigma}-1)\exp(\frac{\epsilon}{\sigma}) - (\frac{\epsilon}{\sigma}+1)\exp(-\frac{\epsilon}{\sigma})) + \frac{\sqrt{6}}{2} - \frac{1}{2\sqrt{6}}(\frac{\epsilon}{\sigma})^2\right) \\ & - \log\left(1 - \frac{1}{\sqrt{6}}\frac{\epsilon}{\sigma} + \frac{\exp(-\sqrt{6})}{2\sqrt{6}}(\exp(\frac{\epsilon}{\sigma}) - \exp(-\frac{\epsilon}{\sigma}))\right). \end{aligned}$$

Again, the solution can be obtained numerically as $\epsilon = 0$, which was also obtained in [19]. Notice that this is intuitively correct as the density function in (5) degenerates to the Laplacian density when $\epsilon = 0$. Moreover, we can also obtain the optimal value of $\beta$ as $1/\beta = 1.5977\sigma$, again showing that $\beta$ is inversely proportional to the noise level of the Laplacian noise model.

To verify the ratio of $\epsilon/\sigma$, we repeat the experiment in Section III-B.1 but with the Laplacian noise $L(0,\sigma)$ instead. Moreover, as our analysis applies only when $x$ is one-dimensional, we also investigate experimentally the optimal value of $\epsilon$ in the two-dimensional case. The ratios obtained for the one- and two-dimensional cases are $\epsilon = 0.0565\sigma$ and $\epsilon = 0.0559\sigma$ respectively (Figure 4), which are close to our prediction of $\epsilon = 0$.

*3) Uniform Noise:* Finally, we consider the uniform noise model

$$\phi(\eta) = \frac{1}{2\sigma} I_{[-\sigma,\sigma]}(\eta), \tag{22}$$

[9]The density function of the exponential distribution $\theta \sim Expon(\beta)$ is $p(\theta) = \beta\exp(-\beta\theta)$ (where $\beta > 0, \theta \geq 0$).

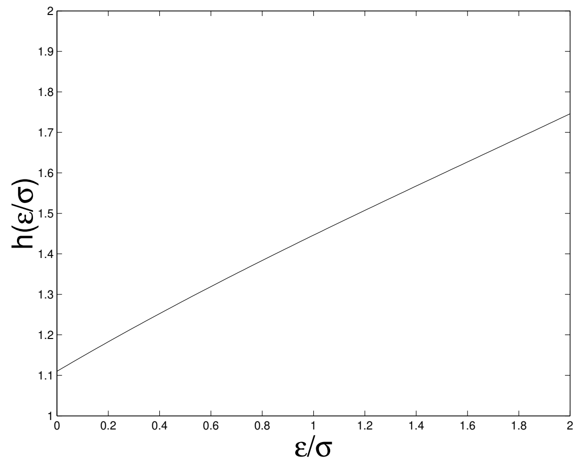
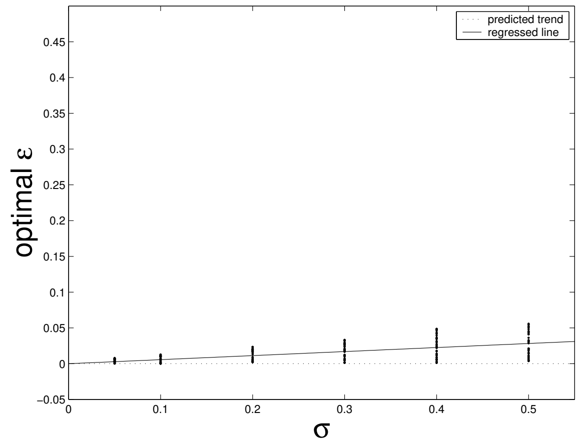
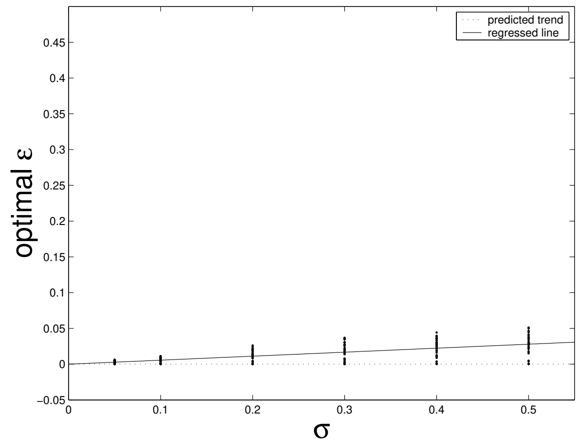



Fig. 3. The objective function $h(\frac{\epsilon}{\sigma})$ to be minimized in the case of Laplacian noise.




(a) one-dimensional $x$.




(b) two-dimensional $\mathbf{x}$.

Fig. 4. Relationship between the optimal value of $\epsilon$ and the Laplacian noise level $\sigma$.

where $I_A(\eta) = \begin{cases} 1 & \text{if } \eta \in A \\ 0 & \text{otherwise.} \end{cases}$ is the indicator function. Again, we only consider the simpler case when $\mathbf{x}$ is one-dimensional, with uniform density over the range $[-L, L]$. Moreover, only the case when $\frac{\sigma-\epsilon}{\tilde{w}-\hat{w}} \leq L \leq \frac{\sigma+\epsilon}{\tilde{w}-\hat{w}}$ will be discussed here. Derivation for the case when $L < \frac{\sigma-\epsilon}{\tilde{w}-\hat{w}}$ is similar. Whereas for $L > \frac{\sigma+\epsilon}{\tilde{w}-\hat{w}}$ (i.e., $\tilde{w}L - \sigma > \hat{w}L + \epsilon$), the corresponding SVR solution will be very poor (Figure 5) and thus usually not the case of interest.

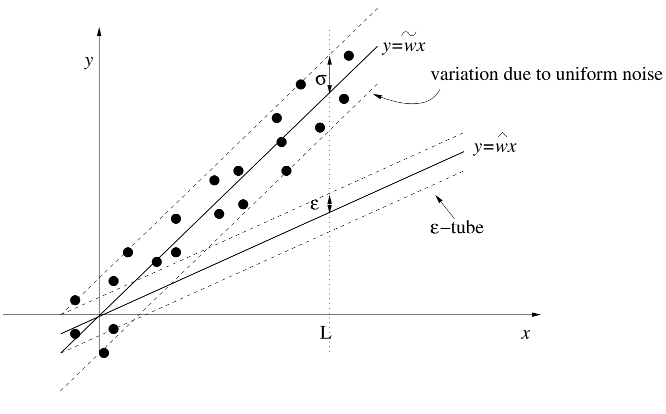



Fig. 5. Scenario when $L > \frac{\sigma+\epsilon}{\tilde{w}-\hat{w}}$ in the case of uniform noise.

It can be shown that (11) and (12) reduce to

$$\frac{1}{1+\epsilon\beta} = \frac{((\sigma-\epsilon)+(\tilde{w}-\hat{w})L)^2}{4\sigma(\tilde{w}-\hat{w})L}, \tag{23}$$

and

$$\frac{1}{\beta(1+\epsilon\beta)} = \frac{((\tilde{w}-\hat{w})L+(\sigma-\epsilon))^3}{12\sigma(\tilde{w}-\hat{w})L}.$$

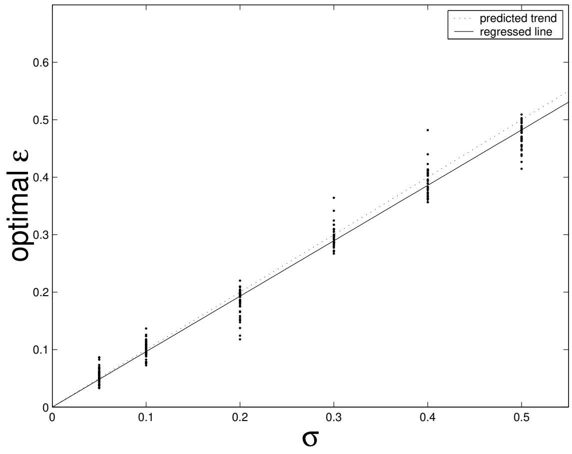
Combining and simplifying, we have $\frac{1}{\beta} = \frac{(\tilde{w}-\hat{w})L+(\sigma-\epsilon)}{3}$. Substituting back into (23), we obtain

$$\frac{L}{\sigma}(\tilde{w}-\hat{w})\left(2-\frac{\epsilon}{\sigma}-(\tilde{w}-\hat{w})\frac{L}{\sigma}\right) = 1-2(\frac{\epsilon}{\sigma})^2+\frac{\epsilon}{\sigma}. \tag{24}$$

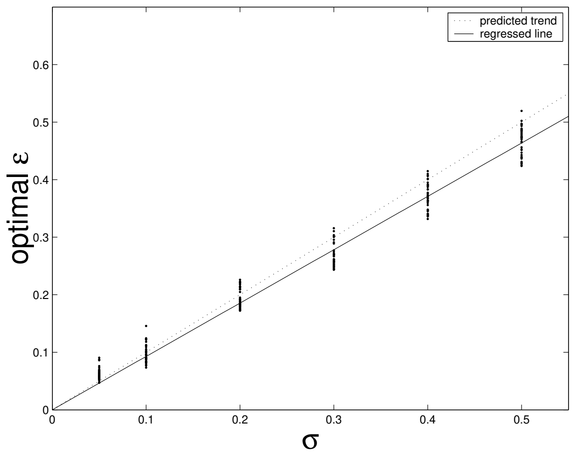
Again, we notice that $\epsilon$ always appears in the factor $\epsilon/\sigma$ and thus there is also a linear relationship between the optimal value of $\epsilon$ and $\sigma$ in the case of uniform noise.

As in Section III-B.1, consider $\delta(x) = (\hat{w}-\tilde{w})x$ and assume that $E_X(\delta^2(x)) \simeq E_X((y-\hat{w}x)^2) = \sigma^2/3$. Using (20) again, we obtain $\sigma = L(\tilde{w}-\hat{w})$. On substituting back into (24) and proceed as in Section III-B.1, we obtain, after tedious computation, $\epsilon = \sigma$. This is intuitively correct since the density function in (5) becomes effectively the same as (22) when $\epsilon = \sigma$, as noise with magnitude larger than the size of the $\epsilon$-tube then becomes impossible. Moreover, as a side-product, we can also obtain the optimal value of $\beta$ as $1/\beta = \sigma/3$, showing that $\beta$ is again inversely proportional to $\sigma$.

Figure 6 shows the linear relationships obtained from the experiment, with $\epsilon = 0.9646\sigma$ and $\epsilon = 0.9275\sigma$ for the one- and two-dimensional cases respectively. These are again in close match with our prediction of $\epsilon = \sigma$.




(a) one-dimensional $x$.




(b) two-dimensional $\mathbf{x}$.

Fig. 6. Relationship between the optimal value of $\epsilon$ and the uniform noise level $\sigma$.

## IV. CONCLUSION

In this paper, we study the $\epsilon$-SVR problem and derive the optimal choice of $\epsilon$ at a given value of the input noise parameter. While the results in [19] considered only the problem of location parameter estimation using maximum likelihood, our analysis is based on the original $\epsilon$-SVR formulation and corresponds to MAP estimation. Consequently, our predicted ratio of $\epsilon/\sigma$ in the case of Gaussian noise is much closer to the experimentally observed optimal value. Besides, we accord with [19] in that $\epsilon$ scales linearly with the scale parameter under the Gaussian, Laplacian and uniform noise models.

In order to apply these linear dependency results in practical applications, one has to first arrive at an estimate of the noise level $\sigma$. One way to obtain this is by using Bayesian methods (e.g., [6]). In the future, we will investigate the integration of these two and also its applications in some real-world problems. Besides, the work here is also useful in designing simulation experiments. Typically, a researcher/practitioner may be experimenting a new technique on $\epsilon$-SVR, while not directly addressing the issue of finding the optimal value of $\epsilon$. The results here can then be used to ascertain that a suitable value/range of $\epsilon$ has been chosen in the simulation. Finally, our analyses under the Laplacian and uniform noise models are restricted to the case when the input $\mathbf{x}$ is one-dimensional, and with uniform density over a certain range. Extensions to the multivariate case and to other noise models will also be investigated in the future.

## ACKNOWLEDGMENTS

This research has been partially supported by the Research Grants Council of the Hong Kong Special Administrative Region under grants HKUST2033/00E and HKUST6195/02E.

## APPENDIX

Recall the following notations introduced in Section III-B.1,

$$\begin{aligned}\delta(\mathbf{x}) &= (\hat{\mathbf{w}} - \tilde{\mathbf{w}})^T \mathbf{x},\\ b_1(\mathbf{x}) &= \frac{1}{\sqrt{2}}(\frac{\epsilon}{\sigma} - \frac{\delta(\mathbf{x})}{\sigma}),\\ b_2(\mathbf{x}) &= \frac{1}{\sqrt{2}}(\frac{\epsilon}{\sigma} + \frac{\delta(\mathbf{x})}{\sigma}).\end{aligned}$$

In the following, we will use the second-order Taylor series expansions for $\operatorname{erfc}(x)$ and $\exp(-x^2)$,

$$\begin{aligned}\operatorname{erfc}(x+h) &= \operatorname{erfc}(x) - \frac{2}{\sqrt{\pi}}e^{-x^2}h + \frac{2}{\sqrt{\pi}}xe^{-x^2}h^2\\ &\quad +O(h^3), \qquad (25)\\ \exp(-(x+h)^2) &= \exp(-x^2)(1-2xh+(2x^2-1)h^2)\\ &\quad +O(h^3). \qquad (26)\end{aligned}$$

Using (25), (13) then becomes

$$\begin{aligned}&\frac{1}{1+\epsilon\beta}\\ &= \int_\Omega \frac{\operatorname{erfc}(b_1)+\operatorname{erfc}(b_2)}{2} p(\mathbf{x})d\mathbf{x}\\ &\simeq \int_\Omega \left(\operatorname{erfc}(\frac{\epsilon}{\sqrt{2}\sigma}) + \frac{2}{\sqrt{\pi}}\frac{\epsilon}{\sqrt{2}\sigma}(\exp(-\frac{\epsilon^2}{2\sigma^2}))\frac{\delta^2(\mathbf{x})}{2\sigma^2}\right) p(\mathbf{x})d\mathbf{x}\\ &= \operatorname{erfc}(\frac{\epsilon}{\sqrt{2}\sigma}) + \frac{\epsilon}{\sqrt{2\pi}\sigma^3}\exp(-\frac{\epsilon^2}{2\sigma^2})E_X(\delta^2(\mathbf{x})). \qquad (27)\end{aligned}$$

Similarly, using (26), (14) becomes

$$\begin{aligned}&\frac{1}{\beta(1+\epsilon\beta)}\\ &= \int_\Omega \left(-\frac{\sigma b_1}{\sqrt{2}}\operatorname{erfc}(b_1) - \frac{\sigma b_2}{\sqrt{2}}\operatorname{erfc}(b_2)\right.\\ &\quad \left. + \frac{\sigma}{\sqrt{2\pi}}\exp(-b_1^2) + \frac{\sigma}{\sqrt{2\pi}}\exp(-b_2^2)\right) p(\mathbf{x})d\mathbf{x}\\ &\simeq \int_\Omega \left[-\epsilon\left(\frac{\operatorname{erfc}(b_1)+\operatorname{erfc}(b_2)}{2}\right) + \frac{\delta(\mathbf{x})}{2}(\operatorname{erfc}(b_1) - \operatorname{erfc}(b_2))\right.\\ &\quad \left. + \frac{\sigma}{\sqrt{2\pi}}\cdot 2\exp(-\frac{\epsilon^2}{2\sigma^2})\left(1 + (\frac{\epsilon^2}{\sigma^2}-1)\frac{\delta^2(\mathbf{x})}{2\sigma^2}\right)\right] p(\mathbf{x})d\mathbf{x}.\end{aligned}$$

Using (13) and (25), this simplifies to

$$\begin{aligned}&E_{XY}(|y - \tilde{\mathbf{w}}^T\mathbf{x}|_\epsilon)\\ &= \frac{1}{\beta(1+\epsilon\beta)}\\ &= -\frac{\epsilon}{1+\epsilon\beta} + \int_\Omega \left[\delta(\mathbf{x})\cdot\frac{2}{\sqrt{\pi}}\exp(-\frac{\epsilon^2}{2\sigma^2})\cdot\frac{\delta(\mathbf{x})}{\sqrt{2}\sigma}\right.\\ &\quad \left. + \sqrt{\frac{2}{\pi}}\exp(-\frac{\epsilon^2}{2\sigma^2})\left(\sigma + (\frac{\epsilon^2}{\sigma^2}-1)\frac{\delta^2(\mathbf{x})}{2\sigma}\right)\right] p(\mathbf{x})d\mathbf{x}\\ &= -\frac{\epsilon}{1+\epsilon\beta} + \sqrt{\frac{2}{\pi}}\exp(-\frac{\epsilon^2}{2\sigma^2})\\ &\quad \left(\sigma + \int_\Omega (\frac{\epsilon^2}{\sigma^2}+1)\frac{\delta^2(\mathbf{x})}{2\sigma}p(\mathbf{x})d\mathbf{x}\right)\\ &= -\frac{\epsilon}{1+\epsilon\beta} + \sqrt{\frac{2}{\pi}}\exp(-\frac{\epsilon^2}{2\sigma^2})\\ &\quad \left(\sigma + \frac{1}{2\sigma}(\frac{\epsilon^2}{\sigma^2}+1)E_X(\delta^2(\mathbf{x}))\right), \qquad (28)\end{aligned}$$

from which we obtain

$$\beta = \sqrt{\frac{\pi}{2}}\exp(\frac{\epsilon^2}{2\sigma^2})\frac{2\sigma}{2\sigma^2 + (\frac{\epsilon^2}{\sigma^2}+1)E_X(\delta^2(\mathbf{x}))}. \qquad (29)$$

Substituting (27) into (28) and simplifying, we have,

$$\begin{aligned}&E_{XY}(|y - \tilde{\mathbf{w}}^T\mathbf{x}|_\epsilon)\\ &= -\epsilon\operatorname{erfc}(\frac{\epsilon}{\sqrt{2}\sigma})\\ &\quad + \sqrt{\frac{2}{\pi}}\exp(-\frac{\epsilon^2}{2\sigma^2})\left(\sigma + \frac{E_X(\delta^2(\mathbf{x}))}{2\sigma}\right). \qquad (30)\end{aligned}$$

Now, maximizing $M$ in (7) is thus the same as minimizing $\beta E_{XY}(|y - \mathbf{w}^T\mathbf{x}|_\epsilon) - \log\frac{\beta}{2(1+\epsilon\beta)}$. Substituting in (15), (27), (29) and (30), this becomes minimizing $h(\frac{\epsilon}{\sigma})$ in (16).

## REFERENCES

[1] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 1985.

[2] V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory and Methods*. Wiley, 1998.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[4] H.D. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161, San Mateo, CA, 1997. Morgan Kaufmann.

[5] T. Evgenious, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 1999.

[6] M.H. Law and J.T. Kwok. Bayesian support vector regression. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, pages 239–244, Key West, Florida, 2001.

[7] Y. Li, S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In *Proceedings of the the Fourth IEEE International Conference on Face and Gesture Recognition*, pages 300–305, Grenoble, 2000.

[8] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, May 1992.

[9] O.L. Mangasarian and D.R. Musicant. Robust linear and support vector regression. *IEEE Transactions on Pattern Analysis and Machine Learning*, 22(9), 2000.

[10] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–202, 2001.

[11] K.R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks*, 1997.

[12] C. Papageorgiou, F. Girosi, and T. Poggio. Sparse correlation kernel reconstruction. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 1633–1636, Phoenix, AZ, 1999.

[13] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo 1140, MIT, July 1989.

[14] M. Pontil, S. Mukherjee, and F. Girosi. On the noise model of support vector machine regression. A.I. Memo No. 1651, MIT, 1998.

[15] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, New York, 2nd edition, 1992.

[16] B. Schölkopf, C. Burges, and A. Smola, editors. *Advances in Kernel Methods – Support Vector Learning*. MIT Press, Cambridge, 1998.

[17] B. Schölkopf, A. Smola, and R.C. Williamson. New support vector algorithms. *Neural Computation*, 12(4):1207–1245, 2000.

[18] A. Smola, B. Schölkopf, and Rätsch G. Linear programs for automatic accuracy control in regression. In *Proceedings of the International Conference on Artificial Neural Networks*, 1999.

[19] A.J. Smola, N. Murata, B. Schölkopf, and K.-R. Müller. Asymptotically optimal choice of $\epsilon$-loss for support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks*, 1998.

[20] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. NeuroCOLT2 Technical Report NC2-TR-1998-030, Royal Holloway College, 1998.

[21] A.J. Smola, B. Schölkopf, and K.-R. Müller. General cost functions for support vector regression. In *Australian Congress on Neural Networks*, 1998.

[22] M. Stitson, A. Gammerman, V.N. Vapnik, V. Vovk, C. Watkins, and J. Weston. Support vector regression with ANOVA decomposition kernels. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press, Cambridge, 1999.

[23] J.A.K. Suykens. Least squares support vector machines for classification and nonlinear modelling. In *Proceedings of the Seventh International Workshop on Parallel Applications in Statistics and Economics*, 2000.

[24] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.