Department of Electrical
and
Computer Systems Engineering

Technical Report
MECSE-24-2005

A monomial ν-SV method for Regression

A. Shilton, D.Lai and M. Palaniswami

MONASH
UNIVERSITY

A Monomial ν-SV Method For Regression

A. Shilton, D. Lai, M. Palaniswami, *Senior Member, IEEE,*

Abstract

In the present paper we describe a new formulation for Support Vector regression (SVR), namely monomial ν-SVR. Like the standard ν-SVR, the monomial ν-SVR method automatically adjusts the radius of insensitivity (the tube width, ϵ) to suit the training data. However, by replacing Vapnik's ϵ-insensitive cost with a more general monomial ϵ-insensitive cost (and likewise replacing the linear tube shrinking term with a monomial tube shrinking term), the performance of the monomial ν-SVR is improved for data corrupted by a wider range of noise distributions. We focus on the quadric form of monomial ν-SVR and show that the dual form of this is simpler than the standard ν-SVR. We show that, like Suykens' Least-Squares SVR (LS-SVR) method (and unlike standard ν-SVR), the quadric ν-SVR dual has a unique global solution. Comparisons are made between the asymptotic efficiency of our method and that of standard ν-SVR and LS-SVR which demonstrate the superiority of our method for the special case of higher order polynomial noise. These theoretical predictions are validated using experimental comparisons with the alternative approaches of standard ν-SVR, LS-SVR and weighted LS-SVR.

I. Introduction

Support Vector regressors (SVRs) [1] [2] [3] are a class of non-linear regressors inspired by Vapnik's SV methods for pattern classification [4] [5]. Like Vapnik's method, SVRs first implicitly map all data into a (usually) higher dimensional feature space. In this feature space, the SVR attempts to construct a linear function of position that mimics the relationship between input (position in feature space) and output observed in the training data by minimising a measure of the empirical risk. To prevent overfitting a regularisation term is included to bias the result toward functions with smaller gradient in feature space.

Two major advantages that SVRs have over competing methods (unregularised least-squares methods, for example) are sparseness and simplicity [3] [6]. SVRs are able to give accurate results based only on a sparse subset of the complete training set, making them ideal for problems with large training sets. Moreover, such results are achievable without excessive algorithmic complexity, and use of the kernel "trick" makes the dual form of the SVR problem particularly simple.

Roughly speaking, SVR methods may be broken into ϵ-SVR [1] [2] and ν-SVR methods [7] [8], both of which require a-priori selection of certain parameters. Of particular interest is the ϵ (or ν in ν-SVR methods) parameter, which controls the sensitivity of the SVR to presence of noise in the training data. In both cases, this parameter controls the threshold ϵ (directly for ϵ-SVR, indirectly for ν-SVR) of insensitivity of the cost function to noise through use of Vapnik's ϵ-insensitive loss function.

The standard ϵ-SVR approach is associated with a simple dual problem, but unfortunately selection of ϵ requires knowledge of the noise present in the training data (and its variance in particular) which may not be available [9]. Conversely, the standard ν-SVR method has a more complex dual form, but has the advantage that selection of ν requires less knowledge of the noise process [9] (only the form of the noise is required, not the variance). Thus both forms have certain difficulties associated with them.

Yet another approach is that of Suykens' LS-SVR [10], which uses the normal least-squares cost function with an added regularisation term inspired by Vapnik's original SV method. The two main advantages of this approach are the simplicity of the resulting dual cost function, which is even simpler than ϵ-SVR; and having one less constant to choose a-priori. The disadvantages include loss of sparsity and robustness in the solution. These problems may be ameliorated somewhat through use of a weighted LS-SVR scheme [11]. However, while this method is noticeably superior when extreme outliers are present in the training data, in our experience the performance of the weighted LS-SVR may not be significantly better than the standard LS-SVR if such outliers are not present.

In view of the shortcomings of these approaches, we present a modification of Smola's ν-SVR method (monomial ν-SVR). Our approach retains the feature that ν may be selected without knowledge of the variance of the noise present in the training data. For the special case of quadric ν-SVR (second order monomial ν-SVR [12]), the associated dual optimisation problem is simpler than the standard ν-SVR method. Furthermore, we show that quadric ν-SVR method is able to out-perform both standard ν-SVR and LS-SVR (weighted or otherwise) in several cases (for example in the presence of higher order polynomial noise).

We begin in section II by reviewing the standard ϵ-SVR method and its properties. Next, using the theory of maximum-likelihood estimation as motivation, we present a modification of this method using a new monomial ϵ-insensitive cost function (monomial ϵ-SVR). Concentrating on the quadric form of this new cost function, we form the dual and show that it is no more complex than the standard ϵ-SVR dual. In subsection II-C we consider the asymptotic efficiency of our method in comparison

A. Shilton and M. Palaniswami are with the Centre of Expertise on Networked Decision and Sensor Systems, Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria 3010, Australia ({apsh,swami}@ee.mu.oz.au).

D. Lai is currently with the Department of Electrical and Computer Systems Engineering, Monash University, Victoria 3168, Australia (daniel.lai@eng.monash.edu.au)

to the standard ϵ-SVR and LS-SVR. We also address the issue of selecting ϵ to maximise this efficiency. Finally, in subsection II-D, we analyze the sparsity of the various methods.

In section III we further explore the properties of the standard ν-SVR method, and in particular its property of insensitivity to the variance of noise present in the training data. We then apply the obvious extension of this approach to the monomial ϵ-SVR formulation introduced previously to produce a monomial ν-SVR method, and show that the dual form of the quadric ν-SVR is actually less complex than the dual form of the standard ν-SVR. We then consider the problem of optimal ν selection for both our method and standard ν-SVR, and show that the property of noise variance insensitivity when selecting this constant carries over from the standard ν-SVR to monomial ν-SVR. Finally, in subsection III-C, we consider the issue of sparsity for the standard ν-SVR, monomial ν-SVR and LS-SVR methodologies.

In order to gain a better "feel" for the problem, in section IV we consider the particular case of training data affected by polynomial noise of degree $1 \leq p \leq 6$. In particular, we compare the theoretical efficiencies and the optimal values of ϵ and ν for our method against other approaches. Finally, in section V, we consider a model problem where various orders of polynomial noise are added to the training data, and compare the results achieved here with our theoretical predictions. We show that the results fit the prediction within tolerable accuracy. Moreover, we show that the predictions for the parameter ν provide a worthwhile "first guess" of the actual optimal value.

II. ϵ-SV Regression

The regression problem may be formulated as follows. Given a training set:

$$\begin{array}{rcl} \Theta & = & \left\{ \left(\mathbf{x}_1, z_1 \right), \left(\mathbf{x}_2, z_2 \right), \ldots, \left(\mathbf{x}_N, z_N \right) \right\} \\ \mathbf{x}_i & \in & \Re^{d_L} \\ z_i & \in & \Re \end{array}$$

where $z_i = \hat{g}\left(\mathbf{x}_i\right) + \text{noise}$ for some $\hat{g} : \Re^{d_L} \rightarrow \Re$; and $\mathbf{x}_i$ is drawn i.i.d. manner from an unknown distribution, construct an approximation $g : \Re^{d_L} \rightarrow \Re$ of $\hat{g}$. An approximation g constructed for a given training set Θ is called a trained machine, and the construction process training. We assume that all noise sources (eg. measurement noise, system noise etc.) are smooth, i.i.d and zero mean.

In the SV approach [1], it is usual to define (implicitly, as will be seen later) a set of functions $\varphi_j : \Re^{d_L} \rightarrow \Re$, $1 \leq j \leq d_H$, which collectively form a map from input space to feature space, $\boldsymbol{\varphi} : \Re^{d_L} \rightarrow \Re^{d_H}$, where $\boldsymbol{\varphi}\left(\mathbf{x}\right) = \left(\varphi_1\left(\mathbf{x}\right), \varphi_2\left(\mathbf{x}\right), \ldots, \varphi_{d_H}\left(\mathbf{x}\right)\right)$. Using this map, the trained machine is defined to be:

$$g\left(\mathbf{x}\right) = \mathbf{w}^T \boldsymbol{\varphi}\left(\mathbf{x}\right) + b$$

which is a linear function of position in feature space. In the ϵ-SVR framework, $\mathbf{w}$ and b are selected to minimise the regularised risk functional:

$$R_1\left(\mathbf{w}, b\right) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{N} \sum_{\left(\mathbf{x}_i, z_i\right) \in \Theta} \left| g\left(\mathbf{x}_i\right) - z_i \right|_\epsilon \tag{1}$$

where $\left| \, . \, \right|_\epsilon = \max\left(\left| \, . \, \right| - \epsilon, 0 \right)$ is Vapnik's ϵ-insensitive loss function ($\epsilon \geq 0$ is a constant). In this expression, the first term ($\frac{1}{2}\mathbf{w}^T\mathbf{w}$) characterises the *complexity* of the model[1] while the second term is a measure of empirical risk associated with the training set when this model is applied. The constant $C > 0$ controls the trade-off between empirical risk minimisation (and potential overfitting) if C is large and complexity minimisation (and potential underfitting) if C is small.

An important property of (1) is that errors of magnitude less than ϵ do not contribute to the cost, $R_1\left(\mathbf{w}, b\right)$. Assuming ϵ is well matched to the noise present in the training data (an issue we will return to shortly), this should lend a degree of noise insensitivity to the cost function [1].

For convenience, (1) is usually expressed in terms of non-negative slack variables $\boldsymbol{\xi}$, $\boldsymbol{\xi}^*$. Using this notation, the primal form of the ϵ-SVR training problem is:

$$\begin{array}{ll} \min\limits_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} & R_1\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*\right) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{N}\sum\limits_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \\ \text{such that:} & \left(\mathbf{w}^T\boldsymbol{\varphi}\left(\mathbf{x}_i\right) + b\right) \geq z_i - \epsilon - \xi_i \\ & -\left(\mathbf{w}^T\boldsymbol{\varphi}\left(\mathbf{x}_i\right) + b\right) \geq -z_i - \epsilon - \xi_i^* \\ & \boldsymbol{\xi}, \boldsymbol{\xi}^* \geq \mathbf{0} \end{array} \tag{2}$$

For reasons of mathematical tractability, it is usual to deal with the dual of (2), which may be constructed as follows. For each of the inequality constraints in (2) we associate a non-negative Lagrange-multiplier, respectively α_i, α_i^*, γ_i and γ_i^* ($1 \leq i \leq N$), noting that this gives a 1-1 correspondence between the training pair $\left(\mathbf{x}_i, z_i\right)$ and the Lagrange multipliers α_i

[1] the larger $\frac{1}{2}\mathbf{w}^T\mathbf{w}$, the larger the gradient of $g\left(\mathbf{x}\right)$ in feature space, and hence the more $g\left(\mathbf{x}\right)$ may vary for a given variation in input, $\mathbf{x}$

and α_i^* for all $1 \leq i \leq N$, and furthermore that at least one of α_i and α_i^* will be zero for all $1 \leq i \leq N$. Using the usual techniques it is straightforward to show that the dual form of (2) is:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & L_1(\boldsymbol{\alpha},\boldsymbol{\alpha}^*) = \tfrac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \mathbf{G}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) - \\ & \qquad (\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \mathbf{z} + \epsilon(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*)^T \mathbf{1} \\ \text{such that:} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \tfrac{C}{N}\mathbf{1} \\ & \mathbf{0} \leq \boldsymbol{\alpha}^* \leq \tfrac{C}{N}\mathbf{1} \\ & \mathbf{1}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) = 0 \end{aligned} \tag{3}$$

where $\mathbf{G}_{i,j} = K(\mathbf{x}_i,\mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j)$, and $K : \Re^{d_L} \times \Re^{d_L} \to \Re$ is the kernel function associated with the map $\boldsymbol{\varphi} : \Re^{d_L} \to \Re^{d_H}$. The trained machine $g(\mathbf{x})$ may be written in terms of the kernel function:

$$g(\mathbf{y}) = \sum_i (\alpha_i - \alpha_i^*) K(\mathbf{x}_i,\mathbf{y}) + b \tag{4}$$

The bias b is calculated indirectly, using the fact that $g(\mathbf{x}_i) = z_i - \epsilon$ for all i such that $0 < \alpha_i < C$ (and likewise $g(\mathbf{x}_i) = z_i + \epsilon$ for all i such that $0 < \alpha_i^* < C$).

At no point during training or use of the trained machine is knowledge of the exact form of map $\boldsymbol{\varphi} : \Re^{d_L} \to \Re^{d_H}$ required. Only the kernel function $K : \Re^{d_L} \times \Re^{d_L} \to \Re$ is required, and any symmetric function $K : \Re^{d_L} \times \Re^{d_L} \to \Re$ satisfying Mercer's condition [13] can be shown to be sufficient for the task [4].

Noting that each training vector corresponds to one pair α_i, α_i^*, and that one of these will be zero for all i, we may divide our training vectors into three distinct classes, namely:

- Non-support vectors: $\alpha_i = \alpha_i^* = 0$, $\xi_i = \xi_i^* = 0$.
- Boundary vectors: $\alpha_i - \alpha_i^* \in \left[-\frac{C}{N},\frac{C}{N}\right] \backslash \{0\}$, $\xi_i = \xi_i^* = 0$.
- Error vectors: $\alpha_i = \frac{C}{N}$, $\xi_i > 0$ or $\alpha_i^* = \frac{C}{N}$, $\xi_i^* > 0$.

Support vectors are any vectors which contribute to (4) (i.e. both boundary and error vectors). We define N_S to be the number of support vectors in the training set, N_E the number of error vectors and N_B the number of boundary vectors; so $N_S = N_B + N_E$. Non-support vectors are said to lie inside the ϵ-tube, boundary vectors on the edge of the ϵ-tube and error vectors outside the ϵ-tube.

We will require the following theorem later:

Theorem 1: (Theorem 3.20, [14]): If the distribution from which the measured outputs $\{z_1, z_2, \ldots, z_N\}$ are drawn is smooth then:

$$\lim_{N\to\infty} \tfrac{N_S}{N} = \lim_{N\to\infty} \tfrac{N_E}{N}$$

A. Monomial ϵ-SV Regression

Consider (1) when $\epsilon = 0$. If C is sufficiently large, and assuming that the empirical risk is non-zero, the second term (the empirical risk term) will be much larger than the first term (the regularisation term). Hence, in this case:

$$R_1(\mathbf{w},b) \approx \tfrac{C}{N} \sum_{(\mathbf{x}_i,z_i)\in\Theta} |g(\mathbf{x}_i) - z_i|$$

But this is just the Max-(log-)likelihood (ML) cost function for data affected by Laplacian noise [14]. It has been shown [14] that, under certain assumptions given in section II-C, the optimal value ϵ_{opt} for ϵ when the output is affected by Laplacian noise is 0. For other types of noise it is often found that $\epsilon_{\text{opt}} \neq 0$, as the empirical risk component of the cost function does not correspond to the ML cost in such cases. The presence of ϵ allows us to achieve a degree of noise insensitivity even though the cost function does not correspond the ML cost function.

The question raised by these observations is whether one may achieve better performance in the SVR for non-Laplacian noise by modifying the primal cost function to match the ML cost function when C is large. However, when doing so we must be mindful of the effect any changes may have on the mathematical tractability of the problem, as mathematical simplicity is one of the major strengths of the SV approach.

One variant of standard SVR which is known to have a particularly simple dual form is Suykens' least squares (LS) SV method [10]. This uses the following modified primal cost function:

$$R_{\text{LS}}(\mathbf{w},b) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{C}{2N} \sum_{(\mathbf{x}_i,z_i)\in\Theta} (g(\mathbf{x}_i) - z_i)^2 \tag{5}$$

If C is sufficiently large the second (empirical risk) term in this expression will be dominant. But the empirical risk term in (5) is just the ML cost function for training data affected by Gaussian noise, so if C is sufficiently large then $R_{\text{LS}}(\mathbf{w},b)$ will correspond approximately to the ML cost function for training data affected by Gaussian noise. Hence one would expect the

LS-SVR to perform well in the presence of Gaussian noise. However, if the noise is not Gaussian, the lack of ϵ-insensitivity in the primal is likely to make the LS-SVR excessively sensitive to noise.[2]

Motivated by this, we propose the following modification of the standard ϵ-SVR primal:

$$R_q(\mathbf{w},b) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{C}{qN}\sum_{(\mathbf{x}_i,z_i)\in\Theta}|g(\mathbf{x}_i)-z_i|_\epsilon^q \tag{6}$$

where $q\in\mathbb{Z}^+$ is a constant. If C is large and $\epsilon=0$, the second (empirical risk) term will be dominant and hence $R_q(\mathbf{w},b)$ will correspond approximately to the ML cost function for degree q polynomial noise, where polynomial noise of degree q is characterised by the density function $p(\tau)=ce^{-d'|\tau|^q}$, where $c,d>0$ are constants.

If $q=1$, (6) reduces to the standard ϵ-SV cost function (1). Similarly, if $q=2$ and $\epsilon=0$, (6) reduces to primal form of Suykens' LS-SVR. However, like the standard ϵ-SV cost function (1) (and unlike the LS-SVR cost function (5)), (6) incorporates ϵ-insensitivity to achieve noise insensitivity when the empirical risk component of the cost function does not match the noise process effecting the training data.

In terms of the usual slack variables, (6) may be written:

$$\begin{aligned}\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} & R_q(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{C}{qN}\sum_{i=1}^N\left(\xi_i^q+\xi_i^{*q}\right)\\ \text{such that:}\quad & \left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right)\geq z_i-\epsilon-\xi_i\\ & -\left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right)\geq -z_i-\epsilon-\xi_i^*\\ & \boldsymbol{\xi},\boldsymbol{\xi}^*\geq\mathbf{0}\end{aligned} \tag{7}$$

We shall refer to a regressor of form (7) as a monomial ϵ-SVR (as the empirical risk term is a monomial function of Vapnik's ϵ-insensitive cost). Unfortunately, for the general case $q>2$ the dual form of (7) is rather complicated [15]. For this reason we will restrict ourselves to the special case $q=2$ (quadric ϵ-SVR), in which case it turns out that the dual problem is mathematically "nice".

Before we construct the dual form of (7) we show that if $q=2$ the positivity constraints $\boldsymbol{\xi},\boldsymbol{\xi}^*\geq\mathbf{0}$ in (7) are superfluous, giving us the simplified primal problem:

$$\begin{aligned}\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} & R_2(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{C}{2N}\sum_{i=1}^N\left(\xi_i^2+\xi_i^{*2}\right)\\ \text{such that:}\quad & \left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right)\geq z_i-\epsilon-\xi_i\\ & -\left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right)\geq -z_i-\epsilon-\xi_i^*\end{aligned} \tag{8}$$

We have the following theorems:

Theorem 2: For every solution $(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*)$ of (8), $\boldsymbol{\xi},\boldsymbol{\xi}^*\geq\mathbf{0}$.

Proof: Suppose there exists a solution $\left(\bar{\mathbf{w}},\bar{b},\bar{\boldsymbol{\xi}},\bar{\boldsymbol{\xi}}^*\right)$ of (8) such that $\bar{\xi}_i<0$ for some $1\leq i\leq N$. Then for all other $(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*)$ satisfying the constraints contained in (8), $R_2(\mathbf{w},b,\xi,\xi^*)\geq R_2\left(\bar{\mathbf{w}},\bar{b},\bar{\boldsymbol{\xi}},\bar{\boldsymbol{\xi}}^*\right)$ by definition.

Consider $(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*)$, where $\mathbf{w}=\bar{\mathbf{w}}$, $b=\bar{b}$, $\boldsymbol{\xi}^*=\bar{\boldsymbol{\xi}}^*$ and:

$$\xi_j=\begin{cases}\bar{\xi}_j & \text{if } i\neq j\\ 0 & \text{otherwise}\end{cases}$$

First, note that as $\left(\bar{\mathbf{w}}^T\boldsymbol{\varphi}(\mathbf{x}_i)+\bar{b}\right)\geq z_i-\epsilon-\bar{\xi}_i$, $\mathbf{w}=\bar{\mathbf{w}}$, $b=\bar{b}$, $\bar{\xi}_i<0$ and $\xi_i=0$, it follows that $\left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right)\geq z_i-\epsilon-\xi_i$. Hence $(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*)$ satisfies the constraints in (8).

Second, note that:

$$\begin{aligned}&R_2\left(\bar{\mathbf{w}},\bar{b},\bar{\boldsymbol{\xi}},\bar{\boldsymbol{\xi}}^*\right) = R_2(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*) + \tfrac{C}{2}\bar{\xi}_i^2\\ &\therefore R_2(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*) < R_2\left(\bar{\mathbf{w}},\bar{b},\bar{\boldsymbol{\xi}},\bar{\boldsymbol{\xi}}^*\right)\end{aligned}$$

These two observations contradict the original assertion that $\left(\bar{\mathbf{w}},\bar{b},\bar{\boldsymbol{\xi}},\bar{\boldsymbol{\xi}}^*\right)$ with $\bar{\xi}_i<0$ for some $1\leq i\leq N$ was a solution of (8). Hence, for all solutions $(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*)$ of (8), $\boldsymbol{\xi}\geq\mathbf{0}$.

The proof of the non-negativity of $\boldsymbol{\xi}^*$ follows from an analogous argument for the elements of this vector. ■

Theorem 3: Any solution $(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*)$ of (8) will also be a solution of (7) when $q=2$, and vice-versa.

Proof: This follows trivially from theorem 2. ■

Using (8) it is straightforward to construct the dual form of (7) when $q=2$ via the usual method. The dual is:

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & L_2(\boldsymbol{\alpha},\boldsymbol{\alpha}^*) = \tfrac{1}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\right)^T\mathbf{G}\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\right)+\\ & \qquad \tfrac{N}{C}\boldsymbol{\alpha}^T\boldsymbol{\alpha}+\tfrac{N}{C}\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*-\\ & \qquad \left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\right)^T\mathbf{z}+\epsilon\left(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*\right)^T\mathbf{1}\\ \text{such that:}\quad & \boldsymbol{\alpha},\boldsymbol{\alpha}^*\geq\mathbf{0}\\ & \mathbf{1}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)=0\end{aligned} \tag{9}$$

[2]As an alternative to the approach presented here, this problem may be tackled by using a *weighted* LS-SVR method [11]. This involves using a two-step process. First, a standard LS-SVR is constructed. Based on this, weights are calculated for each training pair. These weights are subsequently used in a second (weighted) LS-SVR, the training of which results in the trained machine.

where $\mathbf{G}$ is as before. The trained machine takes the same form as (4). Note that the only constraints in (9) are positivity constraints on the dual variables, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, and a single equality constraint. Furthermore, $\boldsymbol{\xi}^{(*)} = \frac{N}{C}\boldsymbol{\alpha}^{(*)}$.

B. Training Issues

To clearly see the structure of the monomial ϵ-SVR dual problem (9), it is instructive to re-express it as follows:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} L_q(\boldsymbol{\alpha},\boldsymbol{\alpha}^*) = & \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{bmatrix}^T \mathbf{Q} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{bmatrix}^T \mathbf{s} \\ \text{such that:} \quad & \boldsymbol{\alpha},\boldsymbol{\alpha}^* \geq \mathbf{0} \\ & \mathbf{1}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) = 0 \end{aligned} \tag{10}$$

where:

$$\begin{aligned} \mathbf{Q} &= \begin{bmatrix} \mathbf{G} & -\mathbf{G} \\ -\mathbf{G} & \mathbf{G} \end{bmatrix} + \tfrac{N}{C}\mathbf{I} \\ \mathbf{s} &= \begin{bmatrix} \epsilon\mathbf{1} - \mathbf{z} \\ \epsilon\mathbf{1} + \mathbf{z} \end{bmatrix} \end{aligned}$$

Once we convert the problem into this form, it is essentially trivial to apply standard SVM training techniques (for example [16], [17]) to the problem. We also have the following useful property (which our formulation shares with Suykens' LS-SVR method):

Theorem 4: The dual problem (10) has a unique global solution.

Proof: It follows from page 79 of [18] that in order to prove that (10) has a unique solution it is sufficient to prove that $\mathbf{Q}$ is positive definite. Since we are using a Mercer kernel, $\mathbf{G}$ is positive semidefinite. Given that $C > 0$, it follows that $\frac{N}{C}\mathbf{I}$ is positive definite. We also know that, for $\mathbf{G}$ positive semidefinite, $\begin{bmatrix} \mathbf{G} & -\mathbf{G} \\ -\mathbf{G} & \mathbf{G} \end{bmatrix}$ must be positive semidefinite. Hence $\mathbf{Q}$ must be positive definite, and (10) must have a unique global solution. ■

As an aside, note that if $\mathbf{G}$ is not positive semidefinite to working precision, $\mathbf{Q}$ may still be made positive definite using the Levenberg-Marquardt [19] [20] method by choosing C appropriately.

C. Asymptotically Optimal Selection of ϵ

In [14], Smola describes how, based on certain assumptions, the parameter ϵ may be selected in an "optimal" fashion for the standard ϵ-SVR. While the assumptions made in this paper are not met by the SVR, experimental results suggest that regardless of this the predictions made are reasonably accurate, and certainly provide a useful "first guess" of the optimal value for ϵ. More importantly, Smola derives a relationship between the variance of the measurement noise and the efficiency of the training machine if the machine is trained using a given ϵ.

Smola's (and our) assumptions are:

1) The training set is infinitely large.
2) The function g estimated by the SVR converges to the actual relationship $\hat{g}$.
3) The general SVR model is replaced by an unregularised location parameter estimator.

Mathematically, the third assumption is met in the limit $K \to 0$, $C \to \infty$, and implies that $g(\mathbf{x}) = b$. Throughout the present paper, these will be referred to as "the usual assumptions".

Following [14], assume $\mathbf{Z} = \{z_1, z_2, \ldots, z_N\}$ is drawn in an i.i.d. manner from some probability density function $p(z|\theta)$ with mean θ and variance σ. Let $\hat{\theta}(\mathbf{Z})$ be an unbiased estimator for θ. The efficiency e of the estimator is defined to be [14]:

$$e = \det(\mathbf{IB})^{-1}$$

where $\mathbf{I}$ is the Fisher information matrix and $\mathbf{B}$ is the covariance matrix [14]. Loosely speaking, the higher the efficiency, the better the estimator (as there is less variance in the estimate of the mean θ). By optimal selection of ϵ, we mean choosing $\epsilon \geq 0$ to maximise the efficiency, e. The value which achieves this maxima will be written ϵ_{opt}. The Cramer-Rao bound [21] states that $e \leq 1$.

For a single parameter estimator of the form:

$$\hat{\theta}(\mathbf{Z}) = \arg\min_{\hat{\theta}} \sum_{z \in \mathbf{Z}} d\left(z, \hat{\theta}\right) \tag{11}$$

where $d\left(z,\hat{\theta}\right)$ is a piecewise twice differentiable function of $\hat{\theta}$ then, asymptotically ([22], lemma 3):

$$e = \frac{Q^2}{IG}$$

where [14]:

$$\begin{aligned} I &= N\int_z \left(\frac{\partial \ln p(z|\theta)}{\partial\theta}\right)^2 p(z|\theta)dz \\ G &= N\int_z \left(\frac{\partial d(z,\theta)}{\partial\theta}\right)^2 p(z|\theta)dz \\ Q &= N\int_z \frac{\partial^2 d(z,\theta)}{\partial\theta^2} p(z|\theta)dz \end{aligned}$$

Making the usual assumptions, the monomial ν-SVR has form (11). That is:

$$d\left(z,\hat{\theta}\right) = \left|z-\hat{\theta}\right|_\epsilon^q$$

where $\hat{\theta} = b$ and $q \in \mathbb{Z}^+$.

Define $p_{\rm std}(\tau)$ and $q_{\rm std}(\tau)$ to be, respectively, the normalised (zero mean, unit variance) and symmetrised normalised distribution functions. That is:

$$\begin{aligned} q_{\rm std}(\tau) &= \tfrac{1}{2}\left(p_{\rm std}(\tau) + p_{\rm std}(-\tau)\right) \\ p_{\rm std}(z) &= \sigma p(\sigma z - \theta|\theta) \end{aligned} \tag{12}$$

Then, denoting $\omega = \frac{\epsilon}{\sigma}$:

$$I = \tfrac{N}{\sigma^2}\int_{-\infty}^{\infty}\left(\frac{\partial \ln p_{\rm std}(\tau)}{\partial\tau}\right)^2 p_{\rm std}(\tau)\,d\tau$$

$$\begin{aligned} G &= N\int_{-\infty}^{\infty}\left(\frac{\partial d(z,\theta)}{\partial\theta}\right)^2 p(z|\theta)\,dz \\ &= N\int\limits_{z\in\Re\setminus[\theta-\epsilon,\theta+\epsilon]}\left(\frac{\partial d\left(|z-\theta|_\epsilon\right)}{\partial|z-\theta|_\epsilon}\right)^2 \tfrac{1}{\sigma}p_{\rm std}\left(\tfrac{z-\theta}{\sigma}\right)dz \\ &= N\sigma^{2q-2}\int\limits_{\tau\in\Re\setminus[-\omega,\omega]}\left(\frac{\partial d\left(|\tau|_\omega\right)}{\partial|\tau|_\omega}\right)^2 p_{\rm std}(\tau)\,d\tau \\ &= 2N\sigma^{2q-2}\int_\omega^\infty(\tau-\omega)^{2q-2}\,q_{\rm std}(\tau)\,d\tau \end{aligned}$$

$$\begin{aligned} Q &= N\int_{-\infty}^{\infty}\frac{\partial^2 d(z,\theta)}{\partial\theta^2}p(z|\theta)\,dz \\ &= N\int\limits_{z\in\Re\setminus[\theta-\epsilon,\theta+\epsilon]}\frac{\partial^2 d\left(|z-\theta|_\epsilon\right)}{\partial|z-\theta|_\epsilon^2}\tfrac{1}{\sigma}p_{\rm std}\left(\tfrac{y-\theta}{\sigma}\right)dz \\ &= N\sigma^{q-2}\int\limits_{\tau\in\Re\setminus[-\omega,\omega]}\frac{\partial^2 d\left(|\tau|_\omega\right)}{\partial|\tau|_\omega^2}p_{\rm std}(\tau)\,d\tau \\ &= 2N\sigma^{q-2}\begin{cases} q_{\rm std}(\omega) & \text{if } q=1 \\ (q-1)\int_\omega^\infty(\tau-\omega)^{q-2}\,q_{\rm std}(\tau)\,d\tau & \text{if } q\geq 2 \end{cases} \end{aligned}$$

Consequently:

$$e(\omega) = \frac{2}{I_{\rm std}}\begin{cases} \frac{q_{\rm std}^2(\omega)}{\left(\frac{1}{2}-\int_0^\omega q_{\rm std}(\tau)d\tau\right)} & \text{if } q=1 \\ (q-1)^2\frac{\left(\int_\omega^\infty(\tau-\omega)^{q-2}q_{\rm std}(\tau)d\tau\right)^2}{\left(\int_\omega^\infty(\tau-\omega)^{2q-2}q_{\rm std}(\tau)d\tau\right)} & \text{if } q\geq 2 \end{cases} \tag{13}$$

where:

$$I_{\rm std} = \int_{-\infty}^{\infty}\left(\frac{\partial \ln p_{\rm std}(\tau)}{\partial\tau}\right)^2 p_{\rm std}(\tau)\,d\tau$$

is the efficiency of the monomial ϵ-SVR under the usual assumptions. The *optimal* value $\epsilon_{\rm opt}$ of the parameter $\epsilon \geq 0$ is defined to be the value which maximizes this efficiency (thereby minimising the variance in the estimate of b).

Defining:

$$\omega_{\rm opt} = \arg\min_\omega \tfrac{1}{e(\omega)} \tag{14}$$

it is clear that maximum efficiency will be achieved by setting $\epsilon = \epsilon_{\rm opt} = \omega_{\rm opt}\sigma$. This implies that $\epsilon_{\rm opt}$ is directly proportional to the noise variance σ, where the constant of proportionality is dependent on the type of noise. If the type and amount (variance) of the noise are both known, (14) provides a reasonable basis for selecting ϵ (or at least a reasonable "first guess").

If $q = 2$, (13) can be used to obtain the theoretical efficiency of Suykens' LS-SVR, under the usual assumptions, by setting $\epsilon = 0$. For later reference, we define $e_{\rm LS}$ to be this efficiency, i.e.:

$$e_{\rm LS} = \frac{2}{I_{\rm std}}\frac{\left(\int_0^\infty q_{\rm std}(\tau)d\tau\right)^2}{\left(\int_0^\infty \tau^2 q_{\rm std}(\tau)d\tau\right)} \tag{15}$$

D. Sparsity of the monomial ϵ-SVR

As discussed previously, part of our motivation for developing the framework for monomial ϵ-SVR is Suykens' LS-SVR technique. Indeed, if C is sufficiently large and the noise affecting measurements Gaussian, LS-SVR corresponds approximately to the ML estimator for the parameters $\mathbf{w}$ and b. However, a downside of the LS-SVR approach is the lack of sparsity in the solution [10], where by sparsity we are referring here to the number of non-zero elements in the vector $\boldsymbol{\alpha}-\boldsymbol{\alpha}^*$ or, equivalently as $N \rightarrow \infty$ (see theorem 1), the fraction of training vectors that are also support vectors. In the LS-SVR case, all training vectors are support vectors, i.e. $N_S = N$. While the sparsity problems of the LS-SVR may be overcome to a degree using the weighted LS-SVR approach [11], this approach is somewhat heuristic in nature, requiring some external arbiter (either human or machine) to decide when to cease pruning the dataset. However, by maintaining the ϵ-insensitive component of the cost function, the need for such heuristics (at least at this level of abstraction) is removed[3].

So long as $\epsilon > 0$, we would expect that the solution $\boldsymbol{\alpha} - \boldsymbol{\alpha}^*$ to the monomial ϵ-SVR dual problem will contain some non-zero fraction of non-support vectors. Under the usual assumptions, we have the following theorem:

Theorem 5: The fraction of support vectors found by the ϵ-SVR under the usual assumptions will, asymptotically, be:

$$\lim_{N\rightarrow\infty} \tfrac{N_S}{N} = 2\int_{\omega}^{\infty} q_{\mathrm{std}}\left(\tau\right) d\tau$$

where $\omega = \frac{\epsilon}{\sigma}$.

Proof: Given that errors vectors are those for which $\left|g\left(\mathbf{x}_i\right) - z_i\right|_{\epsilon} > 0$ (i.e. those lying outside the ϵ-tube), using theorem 1 it can be seen that:

$$\begin{aligned}\lim_{N\rightarrow\infty} \tfrac{N_S}{N} = \lim_{N\rightarrow\infty} \tfrac{N_E}{N} &= \Pr\left(\left|g\left(\mathbf{x}\right) - \hat{g}\left(\mathbf{x}\right)\right| > \epsilon\right) \\ &= \int_{z\in\Re\setminus[\theta-\epsilon,\theta+\epsilon]} p\left(z|\,\theta\right) dz \\ &= \int_{z\in\Re\setminus[-\omega,\omega]} p_{\mathrm{std}}\left(\tau\right) d\tau \\ &= 2\int_{\omega}^{\infty} q_{\mathrm{std}}\left(\tau\right) d\tau\end{aligned}$$

where $\omega = \frac{\epsilon}{\sigma}$. ∎

Note that this implies that:

$$\lim_{\substack{N\rightarrow\infty \\ \epsilon\rightarrow 0}} \tfrac{N_S}{N} = 1$$

as expected for LS-SVR. It also implies that any decrease in ϵ is likely to lead to a decrease in the sparsity of the solution. So, in general, if the training set is large then ϵ should be chosen to be as large as possible while still maintaining acceptable performance to maximise the sparsity of the solution.

III. ν-SV Regression

The major drawback of ϵ-SVR is apparent in the relation $\epsilon_{\mathrm{opt}} = \omega_{\mathrm{opt}}\sigma$. Specifically, selection of ϵ requires knowledge of what type of and how much noise is present in the training set (or alternatively we must have a specific intended accuracy to use as a basis for selecting ϵ). However, while we may have some idea of the type of noise we can expect to be dealing with for a given problem (and hence be able to calculate ω_{opt}), we are unlikely to know how much noise will be present, leaving σ (and therefore ϵ_{opt}) uncertain.

To overcome this problem, Scholkopf et. al. [7] introduced the ν-SVR formulation of the problem, which includes an additional term in the primal problem to trade-off the tube size (ϵ, no longer a constant) against model complexity and empirical risk.

From [7], the primal formulation of the standard ν-SVR training problem is:

$$\begin{aligned}\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*,\epsilon} \quad & R_{1,1}\left(\ldots\right) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\nu\epsilon + \tfrac{C}{N}\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \\ \text{such that:} \quad & \left(\mathbf{w}^T\boldsymbol{\varphi}\left(\mathbf{x}_i\right) + b\right) \geq z_i - \epsilon - \xi_i \\ & -\left(\mathbf{w}^T\boldsymbol{\varphi}\left(\mathbf{x}_i\right) + b\right) \geq -z_i - \epsilon - \xi_i^* \\ & \boldsymbol{\xi}, \boldsymbol{\xi}^* \geq \mathbf{0} \\ & \epsilon \geq 0\end{aligned} \tag{16}$$

[3]Of course, some heuristic input will still be required for each approach, either for ϵ selection or when deciding how much compromise is acceptable during pruning. The advantage of the former is that there exist alternative criteria which may be used to select ϵ (i.e. optimal performance for a given noise model), with sparsity properties being just a useful side affect of this choice. If the dataset is very large, however, sparsity may be of primary importance, in which case neither approach will have a clear advantage.

where $\nu > 0$ is a constant. The associated dual is [7]:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} L_{1,1}(\boldsymbol{\alpha},\boldsymbol{\alpha}^*) &= \tfrac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \mathbf{G}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) - \\ &\qquad (\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \mathbf{z} \\ \text{such that:}\quad & \mathbf{0} \le \boldsymbol{\alpha} \le \tfrac{C}{N}\mathbf{1} \\ & \mathbf{0} \le \boldsymbol{\alpha}^* \le \tfrac{C}{N}\mathbf{1} \\ & \mathbf{1}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) = 0 \\ & \mathbf{1}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) \le C\nu \end{aligned} \tag{17}$$

where $\mathbf{G}$ is as before.

We require the following theorem:

Theorem 6: (Theorem 3.20, [14]): For a ν-SVR trained with any training set of size N:

1) ν is an upper bound on the fraction of error vectors.
2) ν is a lower bound on the fraction of support vectors.

From this theorem we can immediately see that:

$$\nu = \lim_{N\to\infty} \tfrac{N_S}{N}$$

A. Monomial ν-SV Regression

We shall demonstrate shortly that there is a direct functional relationship between ν and the theoretical efficiency e (under the usual assumptions) that is independent of σ. This means that it is possible to find ν_{opt} to achieve maximum theoretical efficiency without knowing the noise variance σ (although the type of noise is still required). However, the price we pay for this is increased complexity in the dual formulation (17), specifically the presence of a new constraint, $\mathbf{1}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) \le C\nu$. Also, like standard ϵ-SVR, the empirical risk component of the standard ν-SVR primal corresponds to ML only for Laplacian noise. To deal with these issues, we introduce the following formulation:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*,\epsilon} R_{q,r}(\ldots) &= \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{C\nu}{r}\epsilon^r + \tfrac{C}{qN}\sum_{i=1}^{N}\left(\xi_i^q + \xi_i^{*q}\right) \\ \text{such that:}\quad & \left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right) \ge z_i - \epsilon - \xi_i \\ & -\left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right) \ge -z_i - \epsilon - \xi_i^* \\ & \boldsymbol{\xi},\boldsymbol{\xi}^* \ge \mathbf{0} \\ & \epsilon \ge 0 \end{aligned} \tag{18}$$

where $q, r \in \mathbb{Z}^+$ are constants. We shall refer to this as monomial ν-SVR when $q = r$. This reduces to the standard ν-SVR primal (16) if we choose $q = r = 1$. We will be concentrating on the case $q = r = 2$ (quadric ν-SVR).

We have already shown in theorem 2 (which will hold here also) that the positivity constraints $\boldsymbol{\xi}, \boldsymbol{\xi}^* \ge \mathbf{0}$ are superfluous if $q = 2$. We now show that the constraint $\epsilon \ge 0$ is also superfluous if $r = 2$, giving us the simplified primal problem when $q = r = 2$:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*,\epsilon} R_{2,2}(\ldots) &= \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{C\nu}{2}\epsilon^2 + \tfrac{C}{2N}\sum_{i=1}^{N}\left(\xi_i^2 + \xi_i^{*2}\right) \\ \text{such that:}\quad & \left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right) \ge z_i - \epsilon - \xi_i \\ & -\left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right) \ge -z_i - \epsilon - \xi_i^* \end{aligned} \tag{19}$$

We have the following theorems:

Theorem 7: For every solution $(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \epsilon)$ of (19), $\epsilon \ge 0$.

Proof: Suppose there exists a solution $\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, \bar{\epsilon}\right)$ of (19) such that $\bar{\epsilon} < 0$. Then:

$$\begin{aligned} & R_{2,2}\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, \bar{\epsilon}\right) = R_{2,2}\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, 0\right) + \tfrac{C\nu}{2}\bar{\epsilon}^2 \\ & \therefore R_{2,2}\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, 0\right) < R_{2,2}\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, \bar{\epsilon}\right) \end{aligned}$$

Furthermore, if the constraints in (19) are satisfied for $\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, \bar{\epsilon}\right)$ they must also be satisfied for $\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, 0\right)$. Therefore $\left(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\xi}}^*, \bar{\epsilon}\right)$ does not minimise $R_{2,2}$ subject to the constraints and hence cannot be a solution of (19), which contradicts the original assertion. So we conclude that every solution $(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \epsilon)$ of (19) must satisfy $\epsilon \ge 0$. ■

Theorem 8: Any solution $(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \epsilon)$ of (19) will also be a solution of (18) when $q = r = 2$, and vice-versa.

Proof: This follows trivially from theorems 7 and 2. ■

It is not difficult to show that the dual form of (19) (and hence (18) with $q = r = 2$) is:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \; & L_{2,2}\left(\boldsymbol{\alpha},\boldsymbol{\alpha}^*\right) = \tfrac{1}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\right)^T \mathbf{G}\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\right) + \\ & \qquad \tfrac{1}{2}\left(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*\right)^T \tfrac{1}{C\nu}\mathbf{E}\left(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*\right) - \\ & \qquad \tfrac{N}{C}\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \tfrac{N}{C}\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^* - \\ & \qquad \left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\right)^T \mathbf{z} \\ \text{such that:} \quad & \boldsymbol{\alpha},\boldsymbol{\alpha}^* \geq \mathbf{0} \\ & \mathbf{1}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) = 0 \end{aligned} \tag{20}$$

where $\mathbf{E}$ is a matrix where all elements are 1. The trained machine takes the same form as (4). Furthermore:

$$\begin{aligned} \epsilon &= \tfrac{1}{C\nu}\mathbf{1}^T\left(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*\right) \\ \boldsymbol{\xi}^{(*)} &= \tfrac{N}{C}\boldsymbol{\alpha}^{(*)} \end{aligned}$$

from which we see that:

$$\epsilon = \tfrac{1}{\nu}\tfrac{1}{N}\sum_{(\mathbf{x}_i,z_i)\in\Theta} \left|g\left(\mathbf{x}_i\right)-z_i\right|_\epsilon \tag{21}$$

Note that the only constraints in (20) are positivity constraints on the dual variables, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, and a single equality constraint. By comparison, the standard ν-SVR dual (17) (which is otherwise of comparable complexity) has upper bounds on these variables and an upper bound on $\mathbf{1}^T\left(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*\right)$.

For training purposes, note that (20) may be expressed identically to (10) by setting:

$$\begin{aligned} \mathbf{Q} &= \begin{bmatrix} \mathbf{G} & -\mathbf{G} \\ -\mathbf{G} & \mathbf{G} \end{bmatrix} + \tfrac{1}{C\nu}\mathbf{E} + \tfrac{N}{C}\mathbf{I} \\ \mathbf{s} &= \begin{bmatrix} -\mathbf{z} \\ \mathbf{z} \end{bmatrix} \end{aligned}$$

Like monomial ϵ-SVR, this form will have a unique global minimum (to see this, note that the additional term $\frac{1}{C\nu}\mathbf{E}$ is positive semidefinite and therefore will not affect the validity of theorem 4).

Finally, note that in the limit $\nu \to \infty$ it is clear from (21) that if $q = r = 2$, $\epsilon \to 0$. In fact, it is clear from the general form of the monomial ν-SVR (18) that as $\nu \to \infty$ we must have $\epsilon \to 0$ to ensure that the primal cost $R_{2,2}$ is finite. This implies that in the limit $\nu \to \infty$ the form of the quadric ν-SVR approaches the form of the LS-SVR.

B. Asymptotically Optimal Selection of ν

In section II-C, we showed that for ϵ-SVR the theoretical efficiency e is a function of $\omega = \frac{\epsilon}{\sigma}$. We now show that e may also be expressed (somewhat indirectly) as a function of ν (making the usual assumptions), independent of σ if $q = r$. The advantage of this form is that, unlike ϵ_{opt}, calculation of ν_{opt} (the value of ν which results in maximum efficiency) does not require knowledge of σ.

Consider the regularised cost function in its primal form:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*,\epsilon} \; & R_{q,r}\left(\ldots\right) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{C\nu}{r}\epsilon^r + \tfrac{C}{qN}\sum_{i=1}^{N}\left(\xi_i^q + \xi_i^{*q}\right) \\ \text{such that:} \quad & \left(\mathbf{w}^T\boldsymbol{\varphi}\left(\mathbf{x}_i\right)+b\right) \geq z_i - \epsilon - \xi_i \\ & -\left(\mathbf{w}^T\boldsymbol{\varphi}\left(\mathbf{x}_i\right)+b\right) \geq -z_i - \epsilon - \xi_i^* \\ & \boldsymbol{\xi},\boldsymbol{\xi}^* \geq \mathbf{0} \\ & \epsilon \geq 0 \end{aligned}$$

where the positivity constraints on ϵ, $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ may or may not be superfluous; and $r, q \in \mathbb{Z}^+$. First, we note that:

$$\tfrac{\partial R_{q,r}}{\partial \epsilon} = C\left(\nu\epsilon^{r-1} + \tfrac{1}{N}\sum_{i=1}^{N}\left(\xi_i^{q-1}\tfrac{\partial \xi_i}{\partial \epsilon} + \xi_i^{*q-1}\tfrac{\partial \xi_i^*}{\partial \epsilon}\right)\right)$$

where[4]:

$$\begin{aligned} \tfrac{\partial \xi_i}{\partial \epsilon} &= \begin{cases} 0 & \text{if } g\left(\mathbf{x}_i\right) > z_i - \epsilon \text{ or } \left(g\left(\mathbf{x}_i\right) = z_i - \epsilon, \delta\epsilon > 0\right) \\ -1 & \text{otherwise} \end{cases} \\ \tfrac{\partial \xi_i^*}{\partial \epsilon} &= \begin{cases} 0 & \text{if } g\left(\mathbf{x}_i\right) < z_i + \epsilon \text{ or } \left(g\left(\mathbf{x}_i\right) = z_i + \epsilon, \delta\epsilon > 0\right) \\ -1 & \text{otherwise} \end{cases} \end{aligned}$$

[4]We have taken some liberties here when calculating this derivative, which is not actually well defined. However, the rate of change as ϵ is either increased or decreased is well defined, which is what we are indicating here.

and hence:

$$\begin{aligned}\frac{\partial R_{q,r}}{\partial \epsilon} &= C\begin{cases}\nu\epsilon^{r-1}-\frac{N_S}{N} & q=1,\delta\epsilon<0\\ \nu\epsilon^{r-1}-\frac{N_E}{N} & q=1,\delta\epsilon>0\\ \nu\epsilon^{r-1}-\frac{1}{N}\sum\limits_{i=1}^{N}\left(\xi_i^{q-1}+\xi_i^{*q-1}\right) & \text{otherwise}\end{cases}\\ &= C\begin{cases}\nu\epsilon^{r-1}-\frac{N_S}{N} & q=1,\delta\epsilon<0\\ \nu\epsilon^{r-1}-\frac{N_E}{N} & q=1,\delta\epsilon>0\\ \nu\epsilon^{r-1}-\frac{1}{N}\sum\limits_{i=1}^{N}\left|z_i-g\left(\mathbf{x}_i\right)\right|_\epsilon^{q-1} & \text{otherwise}\end{cases}\end{aligned} \tag{22}$$

We aim to select ν to arrive at a specific ϵ. In order to achieve this, the optimality condition $\frac{\partial R_{q,r}}{\partial \epsilon}=0$ must be met for this particular value of ϵ. In other words, if $q\geq 2$ we require that:

$$\begin{aligned}\nu &= \epsilon^{1-r}\tfrac{1}{N}\sum_{i=1}^{N}\left(\xi_i^{q-1}+\xi_i^{*q-1}\right)\\ &= \epsilon^{1-r}\tfrac{1}{N}\sum_{i=1}^{N}\left|z_i-\left(\mathbf{w}^T\boldsymbol{\varphi}\left(\mathbf{x}_i\right)+b\right)\right|_\epsilon^{q-1}\end{aligned}$$

Applying the usual assumptions ($g\left(\mathbf{x}\right)=b$, $N\to\infty$), we find that in all cases[5]:

$$\nu=\sigma^{q-r}\left[2\omega^{1-r}\textstyle\int_\omega^\infty\left(\tau-\omega\right)^{q-1}q_{\mathrm{std}}\left(\tau\right)d\tau\right]$$

where, as usual, $\omega=\frac{\epsilon}{\sigma}$, and $q_{\mathrm{std}}\left(\tau\right)$ is defined by (12) . If $q=r$ then:

$$\nu=2\omega^{1-q}\textstyle\int_\omega^\infty\left(\tau-\omega\right)^{q-1}q_{\mathrm{std}}\left(\tau\right)d\tau \tag{23}$$

This implies that if $q=r$ then ν may be selected to achieve a particular efficiency e_{select} (assuming said efficiency is achievable) by finding ω required to achieve this and then substituting this into the relevant expression for ν. At no point is it necessary to use σ, so only the noise type is required.

Note that if we select $q=r=1$, we can retrieve Smola's original result [14]:

$$\nu=1-\textstyle\int_{-\omega}^{\omega}p_{\mathrm{std}}\left(\tau\right)d\tau$$

from which see that for a standard ν-SVR if ν = 1 then, in the limit $N\to\infty$, $\omega=\epsilon=0$.

Generally, if $q=r$ then in order to achieve maximum efficiency we should choose:

$$\nu_{\mathrm{opt}}=2\omega_{\mathrm{opt}}^{1-q}\textstyle\int_{\omega_{\mathrm{opt}}}^\infty\left(\tau-\omega_{\mathrm{opt}}\right)^{q-1}q_{\mathrm{std}}\left(\tau\right)d\tau$$

where:

$$\omega_{\mathrm{opt}}=\arg\min_{\omega}\tfrac{1}{e(\omega)}$$

C. Sparsity of the Monomial ν-SVR

We have shown in theorem 5 that, under the usual assumptions:

$$\lim_{N\to\infty}\tfrac{N_S}{N}=2\textstyle\int_\omega^\infty q_{\mathrm{std}}\left(\tau\right)d\tau$$

which is a 1-1 function (on $\omega\geq 0$) if the distribution $q_{\mathrm{std}}\left(\tau\right)$ is positive on $\tau\geq 0$. We have also shown in the previous section that (again using these assumptions) if $q=r$ then from (23):

$$\nu=2\omega^{1-q}\textstyle\int_\omega^\infty\left(\tau-\omega\right)^{q-1}q_{\mathrm{std}}\left(\tau\right)d\tau$$

which is also a 1-1 function (on $\omega\geq 0$) if $q_{\mathrm{std}}\left(\tau\right)$ is positive on $\tau\geq 0$.

These two functions demonstrate that in general there exists a direct relation between the asymptotic value ($N\to\infty$) of the fraction of support vectors in the training set (and hence the sparsity of the solution) and the parameter ν. In general, the

[5]If $q=1$ then in the limit $N\to\infty$ (22) becomes:

$$\lim_{N\to\infty}\tfrac{\partial R_{1,r}}{\partial\epsilon}=C\begin{cases}\nu\epsilon^{r-1}-\lim\limits_{N\to\infty}\frac{N_S}{N} & \text{if } \delta\epsilon<0\\ \nu\epsilon^{r-1}-\lim\limits_{N\to\infty}\frac{N_E}{N} & \text{if } \delta\epsilon>0\end{cases}$$

which is well defined, as theorem 1 states that $\lim\limits_{N\to\infty}\frac{N_S}{N}=\lim\limits_{N\to\infty}\frac{N_E}{N}$. For optimality, $\frac{\partial R_{q,r}}{\partial\epsilon}=0$, and so using theorem 5 it follows that, if $q=1$:

$$\nu=\sigma^{1-r}\left[2\omega^{1-r}\textstyle\int_\omega^\infty q_{\mathrm{std}}\left(\tau\right)d\tau\right]$$

TABLE I

OPTIMAL $\frac{\epsilon}{\sigma}$ AND ν FOR STANDARD AND QUADRIC (LABELLED (S) AND (Q), RESPECTIVELY) ϵ-SVR AND ν-SVR METHODS WITH POLYNOMIAL ADDITIVE NOISE OF DEGREE $1 \leq p \leq 6$, AND ASYMPTOTIC SUPPORT VECTOR RATIO'S AT OPTIMALITY.

Polynomial degree	1	2	3	4	5	6
Optimal $\frac{\epsilon}{\sigma}$ (S)	0	0.61	1.12	1.36	1.48	1.56
Optimal ν (S)	1	0.54	0.29	0.19	0.14	0.11
$\lim_{N\to\infty} \frac{N_S}{N}$ (S)	1	0.54	0.29	0.19	0.14	0.11
Optimal $\frac{\epsilon}{\sigma}$ (Q)	0	0	0.61	0.97	1.17	1.30
Optimal ν (Q)	∞	∞	0.56	0.18	0.09	0.05
$\lim_{N\to\infty} \frac{N_S}{N}$ (Q)	1	1	0.59	0.39	0.30	0.23

exact nature of this relation will be dependent of the form of the noise process affecting the training data. In the special case $q = r = 1$ (i.e. standard ν-SVR), we see that:

$$\nu = 2\int_{\omega}^{\infty} q_{\mathrm{std}}(\tau) d\tau$$

and so:

$$\nu = \lim_{N\to\infty} \frac{N_S}{N}$$

which coincides with the asymptotic form of theorem 6.

IV. PERFORMANCE IN THE PRESENCE OF POLYNOMIAL NOISE

To gain a better understanding of the method, we will consider a particular example where the training data is affected by additive noise that is polynomial in nature (this includes Gaussian and Laplacian noise as special cases $p = 2$ and $p = 1$ respectively). For polynomial noise, $p_{\mathrm{std}}(\tau) = c_p e^{-c'_p|\tau|^p}$, $p \in \mathbb{Z}^+$, where:

$$\begin{aligned} c_p &= \tfrac{1}{2}\frac{p}{\Gamma\left(\frac{1}{p}\right)}\sqrt{\frac{\Gamma\left(\frac{3}{p}\right)}{\Gamma\left(\frac{1}{p}\right)}} \\ c'_p &= \left(\sqrt{\frac{\Gamma\left(\frac{3}{p}\right)}{\Gamma\left(\frac{1}{p}\right)}}\right)^{p} \end{aligned}$$

We will consider the related questions of optimal parameter selection and sparsity, comparing Smola's standard ν-SVR, our monomial ν-SVR and Suykens' LS-SVR method.

A. Asymptotically Optimal Selection of ϵ and ν

Using (13) and (23) it is not difficult to show that under the usual assumptions, for a monomial ν-SVR, denoting the efficiency of an order q monomial SVR for a training set affected by polynomial noise of order p as $e_{p,q}(\omega)$ (under the usual assumptions), and likewise the connection (23) between ν and ω as $\nu_{p,q}(\sigma, \omega)$:

$$e_{p,q}(\omega) = \frac{1}{\Gamma\left(2-\frac{1}{p}\right)} \begin{cases} \frac{e^{-2c'_p\omega^p}}{\urcorner_{0,p}(\omega)} & \text{if } q = 1 \\ \left(\frac{q-1}{pc'^{\frac{1}{p}}_p}\right)^2 \frac{\urcorner^2_{q-2,p}\left(c'_p;\omega\right)}{\urcorner_{2q-2,p}\left(c'_p;\omega\right)} & \text{if } q \geq 2 \end{cases} \tag{24}$$

$$e_{\mathrm{LS}} = \frac{\Gamma^2\left(\frac{1}{p}\right)}{p^2\Gamma\left(2-\frac{1}{p}\right)\Gamma\left(\frac{3}{p}\right)} \tag{25}$$

$$\nu_{p,q}(\sigma, \omega) = \left[\frac{2c_p}{pc'^{\frac{1}{p}}_p}\frac{1}{\omega^{q-1}}\urcorner_{q-1,p}\left(c'_p;\omega\right)\right] \tag{26}$$

where we have defined:

$$\begin{aligned} \urcorner_{m,p}(\beta;\omega) &= p\beta^{\frac{1}{p}}\int_{\omega}^{\infty}(\tau-\omega)^m e^{-\beta\tau^p}d\tau \\ &= \sum_{i=0}^{m}\binom{m}{i}(-\omega)^{m-i}\beta^{-\frac{i}{p}}\Gamma\left(\frac{i+1}{p}, \beta\omega^p\right) \end{aligned} \tag{27}$$

for $m \in \mathbb{Z}\backslash\mathbb{Z}^-$ and $p \in \mathbb{Z}^+$; and $\Gamma(a, x)$ is the complementary incomplete gamma function [23]:

$$\Gamma(a, x) = \int_x^{\infty} t^{a-1}e^{-t}dt$$

Table I shows the optimal values for ϵ and ν for polynomial noise of degree $1 \leq p \leq 6$ for both standard and quadric ϵ-SVR and ν-SVR (source [14]). Unsurprisingly, the optimal value for ϵ for the quadric ϵ-SVR in the presence of Gaussian noise ($p = 2$) is 0, as in this case the quadric ϵ-SVR primal and the maximum-likelihood estimator both take approximately

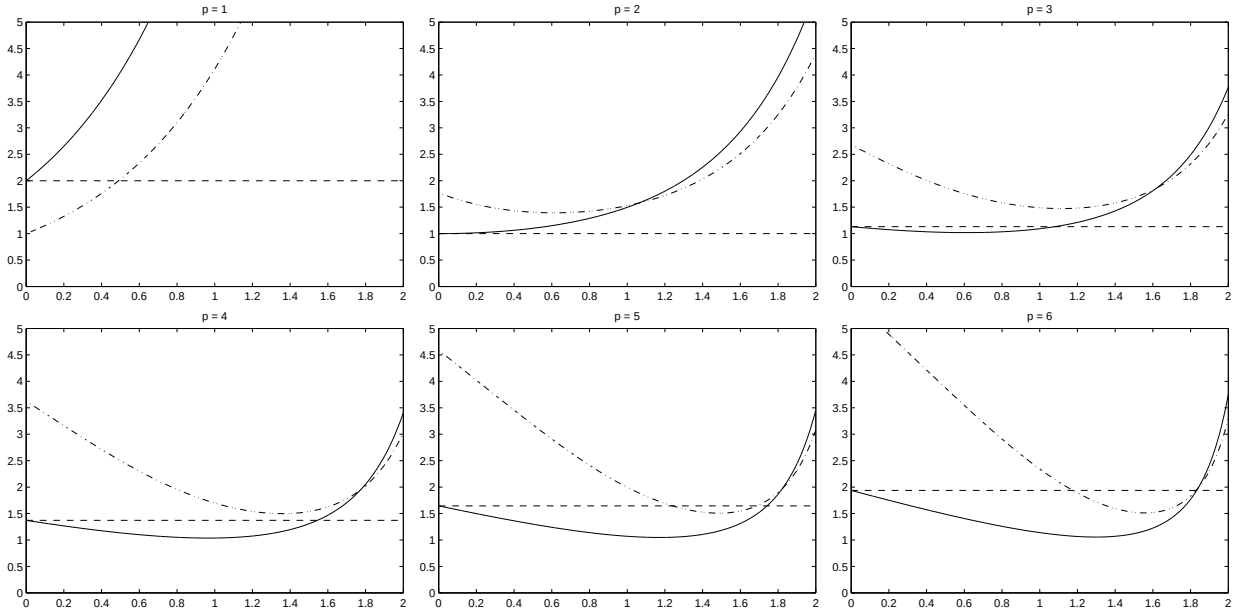

Fig. 1. Comparative inverse asymptotic efficiency versus $\frac{\epsilon}{\sigma}$ of standard ϵ-SVR, quadric ϵ-SVR and LS-SVR for polynomial noise of degrees $1 \leq p \leq 6$. In all cases, the solid line represents the efficiency of quadric ϵ-SVR, the dotted line the efficiency of standard ϵ-SVR, and the dashed line the efficiency of the LS-SVR.

the same form if $\epsilon = 0$ and C is large (just as $\epsilon_{\text{opt}} = 0$ for standard ϵ-SVR in the presence of Laplacian noise ($p = 1$)). Note also that, for the cases represented in the table, $\epsilon_{\text{opt}} > 0$ for all $p > q$ and $\epsilon_{\text{opt}} = 0$ for all $p \leq q$. Generally:

Theorem 9: Under the usual assumptions, given a set of training data affected by polynomial noise of degree $p = q$ with non-zero variance, the optimal value (ϵ_{opt} which maximizes e) for the parameter ϵ as defined by (24) will be zero.

Theorem 10: Under the usual assumptions, given a set of training data affected by polynomial noise of degree $p > q$ with non-zero variance, the optimal value (ϵ_{opt} which maximizes e) for the parameter ϵ as defined by (24) will be positive.

Conjecture 11: Under the usual assumptions, given a set of training data affected by polynomial noise of degree $p < q$ with non-zero variance, the optimal value (ϵ_{opt} which maximizes e) for the parameter ϵ as defined by (24) will be zero.

The proofs for theorems 9 and 10, and a partial proof of conjecture 11 (which we have observed experimentally to be try for all $1 \leq q \leq 1000$) may be found in appendix I.

Figure 1 shows the inverse asymptotic efficiency versus $\frac{\epsilon}{\sigma}$ for both standard and quadric ϵ-SVRs, as well as LS-SVR. Note that with the exception of Laplacian noise ($p = 1$) the optimal theoretical efficiency of the quadric ϵ-SVR exceeds the optimal theoretical efficiency of the standard ϵ-SVR. Also note that the efficiency of monomial ϵ-SVR methods exceeds that of LS-SVR for all $p > 2$.

Finally, figure 2 shows the inverse asymptotic efficiency versus ν for both standard and quadric ν-SVRs, as well as LS-SVR. The important thing to note from these graphs is that, although the range of ν is much larger for quadric ν-SVR methods than for standard ν-SVR methods, the efficiency itself quickly flattens out. In particular, although the theoretically optimal value for Gaussian noise is $\nu \to \infty$, it can be seen that even when $\nu = 1$ the efficiency is very close to it's maximum, $e = 1$. Also note the comparative flatness of the efficiency curves for quadric ν-SVR.

B. Sparsity Issues

In the case of polynomial noise, theorem 5 implies that for monomial SVRs:

$$\lim_{N\to\infty} \frac{N_S}{N} = \frac{2c_p}{pc_p'^{\frac{1}{p}}} \beth_{0,p}\left(c_p'; \omega\right)$$

We have already observed that for monomial ν-SVRs:

$$\nu_{p,q}\left(\omega\right) = \frac{2c_p}{pc_p{}'^{\frac{1}{p}}} \frac{1}{\omega^{q-1}} \beth_{q-1,p}\left(c_p'; \omega\right)$$

Or, in the two cases of particular interest (standard ($q = 1$) and quadric ($q = 2$) ν-SVR):

$$\begin{aligned} \nu_{p,1}\left(\omega\right) &= \frac{2c_p}{pc_p{}'^{\frac{1}{p}}} \beth_{0,p}\left(c_p'; \omega\right) \\ \nu_{p,2}\left(\omega\right) &= \frac{2c_p}{pc_p{}'^{\frac{1}{p}}} \frac{1}{\omega} \beth_{1,p}\left(c_p'; \omega\right) \end{aligned}$$

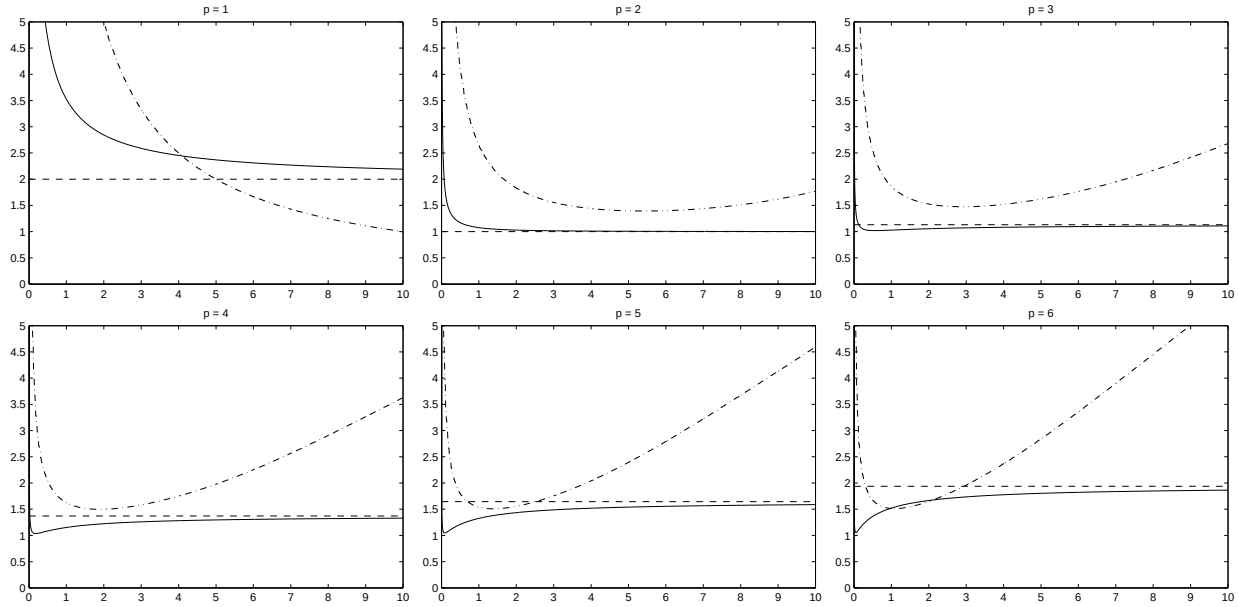

Fig. 2. Comparative inverse asymptotic efficiency versus ν of standard ν-SVR, quadric ν-SVR and LS-SVR for polynomial noise of degrees $1 \leq p \leq 6$. In all cases, the solid line represents the efficiency of quadric ν-SVR, the dotted line the efficiency of standard ν-SVR and the dashed line the efficiency of the LS-SVR. Note that for all graphs the ν-scale differs for standard and quadric ν-SVRs. For standard ν-SVRs, the actual setting for ν is one tenth of that indicated on the x axis (for quadric ν-SVRs, ν is as indicated).

In the case $q = 1$, as expected, this implies:

$$\lim_{N\to\infty} \frac{N_S}{N} = \nu$$

The case $q = 2$ is slightly more complex, and best illustrated graphically. We do this in figure 3, which shows our predictions for the fraction of support vectors found as a function of ν for both standard and quadric ν-SVR methods for polynomial noise of degree $1 \leq p \leq 6$. Note that the general shape of the curves is essentially identical in all cases. Generally, the fraction of support vectors found by the quadric ν-SVR will increase quickly while ν is small and then level out, approaching 1 as $\nu \to \infty$ (as expected, given that the LS case corresponds with $\nu \to \infty$ and treats all vectors as support vectors).

Table I gives the expected asymptotic ratio of support vectors to training vectors when ν is optimally selected for the usual degrees of polynomial noise. On average, the results given in the table imply that quadric ν-SVRs may require approximately twice as many support vectors as standard ν-SVRs to achieve optimal accuracy on the same dataset. This may be understood by realising that the act of extracting support vectors is essentially a form of lossy compression. The modified ν-SVR is (theoretically) able to achieve more accurate results than standard ν-SVR because it can handle more information (by using less compression or, equivalently, finding more support vectors) before over-fitting (and subsequent degradation in performance) begins.

V. Experimental Results

Due to the restrictive assumptions made when deriving the results for theoretical efficiency given in the preceding sections (which will, in the strictest sense, never be satisfied for any SVR), it is important that we seek some form of experimental confirmation of these results. This is our aim in the present section.

For ease of comparison our experimental procedure is modelled on that of [24]. We have numerically computed the risk in the form of root mean squared error (RMSE) as a function of ν for both standard and quadric ν-SVR methods, allowing clear comparison between the two methods. We also compute the RMSE for Suykens' LS-SVR, which is the limiting case of quadric ν-SVR as $\nu \to \infty$, and (non-sparse) weighted LS-SVR [11]. Plots of risk versus ν are given for polynomial noise of degree $1 \leq p \leq 6$ to compare the theoretical and experimental results, and some relevant results are given for the effect of other parameters on the ν curves.

Finally, we compare the sparsity of standard and quadric ν-SVR for different orders of polynomial noise $1 \leq p \leq 6$, and compare these results with the theoretical predictions.

As in [24], the training set consisted of 100 examples (x_i, z_i) where x_i is drawn uniformly from the range $[-3, 3]$ and z_i is given by the noisy sinc function:

$$\begin{aligned} z_i &= \operatorname{sinc}(x_i) + \zeta_i \\ &= \frac{\sin(\pi x_i)}{\pi x_i} + \zeta_i \end{aligned}$$

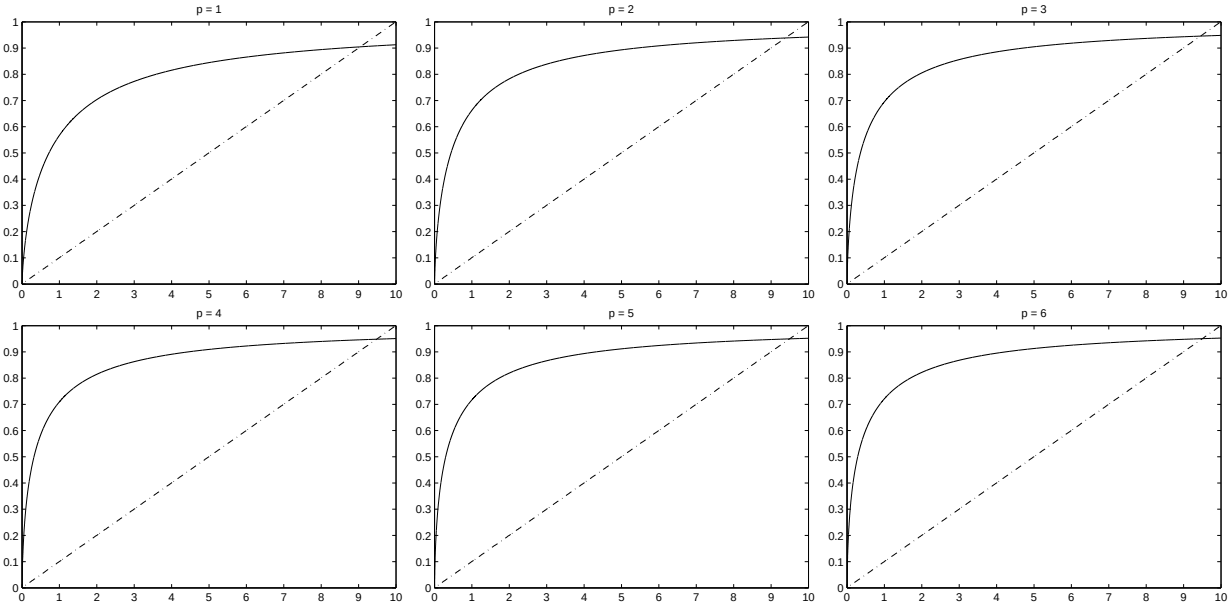

Fig. 3. Comparative asymptotic fraction of support vectors versus ν of standard and quadric ν-SVR for polynomial noise of degrees $1 \leq p \leq 6$. In all cases, the solid line represents the efficiency of modified ν-SVR and the dotted line standard ν-SVR. Note that for all graphs the ν-scale differs for standard and quadric ν-SVRs. For standard ν-SVRs, the actual setting for ν is one tenth of that indicated on the x axis (for quadric ν-SVRs, ν is as indicated).

where ζ_i represents additive polynomial noise. The test set consisted of 500 pairs (x_i, z_i) where the x_i's were equally spaced over the interval $[-3, 3]$ and z_i was given by the noiseless sinc function. 500 trials were carried out for each result.

Quadric ν-SVR, standard LS-SVR and weighted LS-SVR code for this experiment was written in C++ and compiled using DJGPP on a 1GHz Pentium III with 512MB of memory, running Windows 2000[6]. Experiments using standard ν-SVR methods were done using LibSVM [16].

By default, we set the parameter $C = 100$ and noise variance $\sigma = 0.5$. In all experiments we use the Gaussian RBF kernel function $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{2\sigma_{\text{kernel}}^2}\|\mathbf{x}-\mathbf{y}\|^2}$ where $2\sigma_{\text{kernel}}^2 = 1$ by default.

A. Additive Polynomial Noise

In our first experiment, we have used training data affected by polynomial noise of degree $1 \leq p \leq 6$. Plots of RMSE for standard ν-SVR, quadric ν-SVR, LS-SVR and weighted LS-SVR are shown in figure 4. With the exception of Laplacian noise ($p = 1$), these results resemble the theoretical predictions shown in figure 2. Note that the quadric ν-SVR outperforms both standard ν-SVR and (weighted and unweighted) LS-SVR in all cases except Laplacian noise ($p = 1$).

Whilst the general shape of the RMSE curves closely resembles the shape of the predicted efficiency curves, it is important to note that the sharp optimum predicted by the theory for $p \geq 4$ (see figure 2) is not present in the experimental results. Instead, the region of optimality is somewhat to the right of this and also somewhat blunter.

From an application point of view, this is actually an advantage, as it means that results are far less sensitive to ν than expected. Indeed, from these results we can empirically say that the "sweet spot" for ν lies between 0.5 and 1 for polynomial noise of degree $p \geq 3$ and anything above 2 otherwise. But, roughly speaking, selecting $\nu = 1$ will in all cases presented give results superior to the standard ν-SVR method and at least comparable to the LS-SVR methods (better if $p \geq 3$).

In the case of Laplacian noise the actual performance (in terms of RMSE) of both the quadric ν-SVR and the LS-SVRs are better than that of the standard ν-SVR, whereas theory would suggest that this should not be the case. We are unsure as to why this anomaly occurs.

Figure 5 shows the ratio of support vectors to training vectors for both standard and quadric ν-SVRs as a function of ν. These curves closely match the predictions given in figure 3. It is clear from figures 4 and 5 that, as expected, the number of support vectors found by the quadric ν-SVR is substantially larger than the number found by the standard ν-SVR. However, this is still substantially less than the number of support vectors found by the LS-SVR (which uses all training vectors as support vectors).

B. Parameter Variation with Additive Gaussian Noise

In our second experiment, we consider the performance of the standard and quadric ν-SVR in the presence of additive Gaussian noise as other parameters of the problem (particularly σ (the noise variance), C and σ_{kernel}). In particular, we wish

[6]code available at http://www2.ee.mu.oz.au/pgrad/apsh/svm/

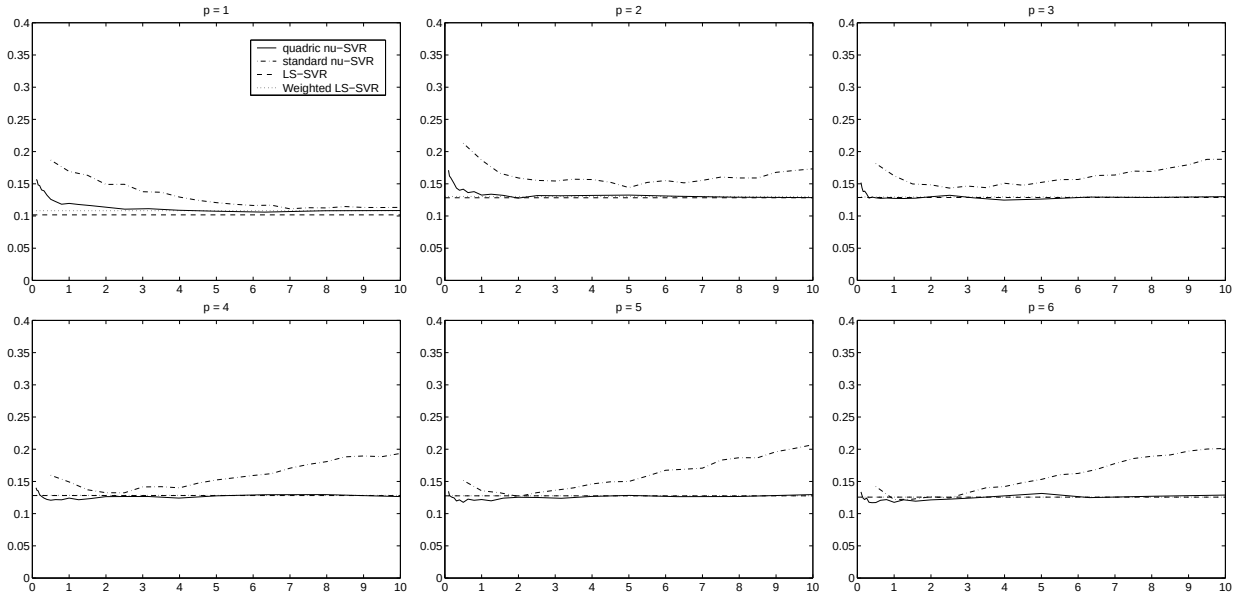

Fig. 4. Risk (RMSE) versus ν for sinc data with polynomial noise of degree $p \in \{1, 2, 3, 4, 5, 6\}$ working left to right, top to bottom (c.f. [24], figure 2). In all cases, $\sigma = 0.5$, $C = 100$ and $2\sigma^2_{\text{kernel}} = 1$. Note that for all graphs the ν-scale differs for standard and modified ν-SVRs. For standard ν-SVRs, the actual setting for ν is one tenth of that indicated on the x axis (for modified ν-SVRs, ν is as indicated).

to see if the RMSE versus ν curve retains the same general form as these parameters are varied.

The top row of figure 6 shows the form of the RMSE versus ν curve for a range of noise levels. In all cases the form of the curves remains essentially unchanged. It will be noted, however, that for the lowest noise case ($\sigma = 0.1$) the standard ν-SVR is able to out-perform the quadric ν-SVR. We are unsure as to why this occurs. Once again, selecting $\nu > 1$ gives reasonable performance in all cases (c.f. Smola's result for standard ν-SVR [9], where the "optimal area" is $\nu \in [0.3, 0.8]$).

The middle row of figure 6 shows the same curve with the noise variance σ fixed for different values of C, namely $C \in \{10, 100, 1000\}$. Once again, the RMSE curve for quadric ν-SVR is roughly as predicted, and $\nu > 1$ gives reasonable results. In this case, $C = 10$ provides an anomalous result.

Finally, in the bottom row of figure 6 , we give the RMSE versus ν curves when the kernel parameter $2\sigma^2_{\text{kernel}}$ is varied. These results follow the same pattern as for variation of σ and C.

It is interesting to note here that, while the RMSE versus ν curve obtained using the quadric ν-SVR fits more closely the predicted curve than does the same curve for standard ν-SVR when parameters are chosen badly, this does not imply that the performance of the quadric ν-SVR will necessarily be better than the standard ν-SVR in this case. Indeed, the two cases where other SV parameters (i.e. not ν) have been chosen badly (i.e. $C = 10$ and $2\sigma^2_{\text{kernel}} = 0.1$ in figure 6) are the two cases where the standard ν-SVR most outperforms the quadric ν-SVR. However, as one should continue to search until appropriate parameters are found, this should not be too much of a problem in practical situations.

VI. Conclusion

In this paper we have reviewed the standard SVR techniques of ϵ-SVR and ν-SVR. Motivated by the relative merits and demerits of these approaches, we then introduced a new, more general form of SVR (monomial ν-SVR). We proceeded to investigate a special case of monomial ν-SVR (quadric ν-SVR) in some detail and showed that the dual form of the quadric ν-SVR dual problem is significantly simpler than the standard ν-SVR dual. We have compared the theoretical asymptotic efficiencies of our proposed formulation against other SVR methods under certain restrictive assumptions. Our investigation indicates that the quadric ν-SVR method not only shares the standard ν-SVR method's property of (theoretical) noise variance insensitivity, but is also more efficient in many cases (in particular, when the training data is affected by polynomial noise of degree $p \geq 2$). These predictions have been experimentally tested, and comparisons have been made between the performances of quadric ν-SVR, standard ν-SVR, standard LS-SVR and weighted LS-SVR methods in the presence of higher order polynomial noise. Based on these results, we conclude that the theoretical predictions give a useful insight into the characteristics of the various methods. Both theoretical and experimental results indicate that performance of the quadric ν-SVR method in many cases exceeds that of both standard ν-SVR and LS-SVR.

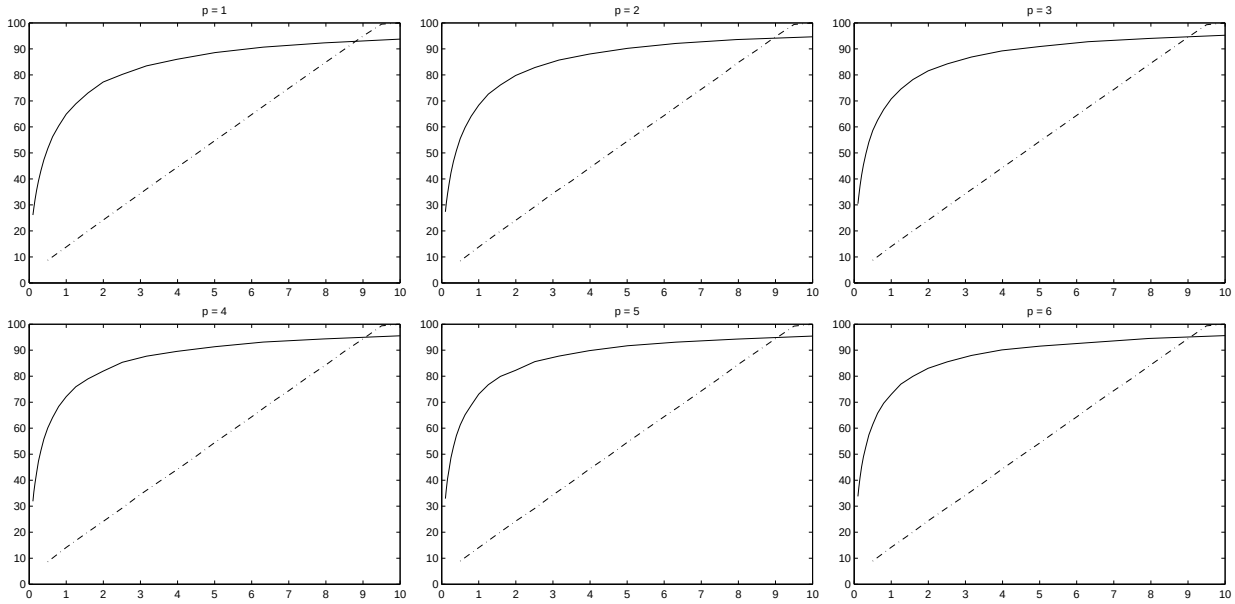

Fig. 5. Number of support vectors (out of 100 training vectors) versus ν for sinc data with polynomial noise of degree $p \in \{1, 2, 3, 4, 5, 6\}$ working left to right, top to bottom. In all cases, $\sigma = 0.5$, $C = 100$ and $2\sigma^2_{\text{kernel}} = 1$. The dotted line gives results for standard ν-SVR, while the solid line gives results for quadric ν-SVR. Note that for all graphs the ν-scale differs for standard and modified ν-SVRs. For standard ν-SVRs, the actual setting for ν is one tenth of that indicated on the x axis (for modified ν-SVRs, ν is as indicated).

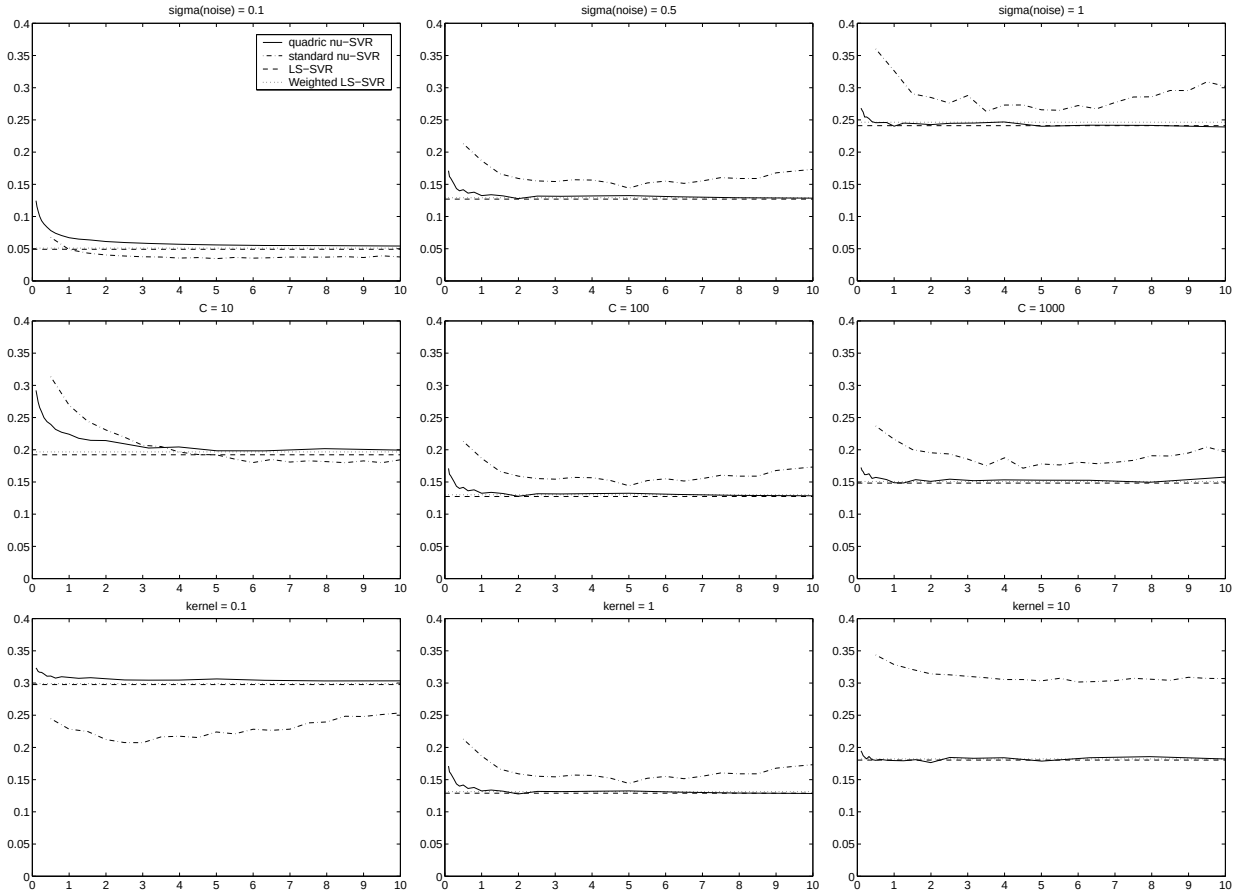

Fig. 6. Risk (RMSE) versus ν for sinc data with Gaussian noise (c.f. [24], figure 1). The top row shows performance for different noise variances, $\sigma \in \{0.1, 0.5, 1\}$ from left to right, with $C = 100$ and $2\sigma^2_{\text{kernel}} = 1$. The middle row gives performance for $C \in \{10, 100, 1000\}$, respectively, with $\sigma = 0.5$ and $2\sigma^2_{\text{kernel}} = 1$. Finally, the bottom row shows performance for $2\sigma^2_{\text{kernel}} \in \{0.1, 1, 10\}$, respectively, with $\sigma = 0.5$ and $C = 100$. Note that for all graphs the ν-scale differs for standard and quadric ν-SVRs. For standard ν-SVRs, the actual setting for ν is one tenth of that indicated on the x axis (for quadric ν-SVRs, ν is as indicated).

APPENDIX I
PROOFS OF THEOREMS

In this appendix we give proofs for theorems 10 and 9, and a partial proof of conjecture 11. Before proceeding, however, some preliminary results are required. In what follows, $B(x,y)$ is the *beta function* [25]:

$$B(x,y) = \tfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

It is straightforward to show that:

$$\tfrac{d}{d\omega}\urcorner_{m,p}(\beta;\omega) = -\begin{cases} m\urcorner_{m-1,p}(\beta;\omega) & \text{if } m>0 \\ p\beta^{\frac{1}{p}}e^{-\beta\omega^p} & \text{if } m=0 \end{cases} \tag{28}$$

$$\urcorner_{m,p}(\beta;0) = \beta^{-\frac{m}{p}}\Gamma\left(\tfrac{m+1}{p}\right) \tag{29}$$

Theorem 12: $\lim_{x\to 0^+} x\Gamma(ax) = \frac{1}{a}$ for all $a>0$.

Proof: Using the Euler limit form of $\Gamma(x)$ [26], it can be seen that:

$$x\Gamma(ax) = x\tfrac{1}{ax}\textstyle\prod_{n=1}^{\infty}\left[\left(1+\frac{1}{n}\right)^{ax}\left(1+\frac{ax}{n}\right)^{-1}\right]$$

and therefore $\lim_{x\to 0^+} x\Gamma(ax) = \frac{1}{a}$. ∎

Theorem 13: $B(a+c,b) > B(a,b+c)$ for all $a>b$, $a,b,c>0$; and $B(a+c,b) < B(a,b+c)$ for all $a<b$, $a,b,c>0$.

Proof: From [25], section 1.5, equation 5, it can be seen that:

$$\begin{aligned} B(a+c,b) &= \tfrac{B(c,b)}{B(a,c)}B(a,b+c) \\ &= \tfrac{\frac{\Gamma(b)}{\Gamma(b+c)}}{\frac{\Gamma(a)}{\Gamma(a+c)}}B(a,b+c) \end{aligned} \tag{30}$$

But the gradient $\psi_0(x)$ of $\ln(\Gamma(x))$ is a monotonically increasing function of $x>0$, and so for all $a>b$, $a,b,c>0$:

$$\ln(\Gamma(a+c)) - \ln(\Gamma(a)) > \ln(\Gamma(b+c)) - \ln(\Gamma(b))$$

and hence $\frac{\Gamma(b)}{\Gamma(b+c)} > \frac{\Gamma(a)}{\Gamma(a+c)}$ for all $a>b$, $a,b,c>0$. Using (30), it follows that $B(a+c,b) > B(a,b+c)$ for all $a>b$, $a,b,c>0$, which proves the first part of the theorem. The proof of the second part is essentially the same, except that, as $a<b$, $\frac{\Gamma(b)}{\Gamma(b+c)} < \frac{\Gamma(a)}{\Gamma(a+c)}$, and so $B(a+c,b) < B(a,b+c)$ for all $a<b$, $a,b,c>0$. ∎

We may now proceed to the proofs of theorems 9 and 10.

Proof of theorem 9: From (24), if $p=q$ then, using (29) (and noting that $x\Gamma(x) = \Gamma(x+1)$):

$$\begin{aligned} e_{p,p}(0) &= \tfrac{1}{\Gamma\left(2-\frac{1}{p}\right)}\begin{cases} \frac{1}{\urcorner_{0,p}(0)} & \text{if } p=1 \\ \left(\frac{p-1}{pc_p'^{\frac{1}{p}}}\right)^2\frac{\urcorner^2_{p-2,p}\left(c_p';0\right)}{\urcorner_{2p-2,p}\left(c_p';0\right)} & \text{if } p\geq 2 \end{cases} \\ &= \begin{cases} 1 & \text{if } p=1 \\ \frac{(p-1)^2}{p^2c_p'^{\frac{2}{p}}}\frac{c_p'^{\frac{4-2p}{p}}\Gamma^2\left(1-\frac{1}{p}\right)}{c_p'^{\frac{2-2p}{p}}\Gamma^2\left(2-\frac{1}{p}\right)} & \text{if } p\geq 2 \end{cases} \\ &= \begin{cases} 1 & \text{if } p=1 \\ \left(1-\frac{1}{p}\right)^2\frac{\Gamma^2\left(1-\frac{1}{p}\right)}{\Gamma^2\left(2-\frac{1}{p}\right)} & \text{if } p\geq 2 \end{cases} \\ &= 1 \end{aligned}$$

But $e_{p,q}(\omega) \leq 1$ by the Cramer-Rao bound [21], so optimal efficiency is achieved for $\omega = \frac{\epsilon}{\sigma} = 0$. Hence $\epsilon_{\text{opt}} = 0$ is a solution. ∎

Proof of theorem 10: Note that $e_{p,q}(\omega) \in C^2$ for $\omega \geq 0$. Note also that the range of ω is $\omega \in [0,\infty)$. If $\epsilon_{\text{opt}} = 0$, and hence $\omega_{\text{opt}} = 0$, $e_{p,q}(\omega)$ must have a (global) maxima at $\omega = 0$, which implies that the gradient $e'_{p,q}(\omega)$ of $e_{p,q}(\omega)$ must be non-positive at 0.[7] Given this, to prove the theorem, it is sufficient to prove that the gradient $e'_{p,q}(\omega)$ of $e_{p,q}(\omega)$ at $\omega = 0$ is positive for all $p>q$ (intuitively it may be seen that if the gradient is positive at zero then increasing ω must result in an increase in $e_{p,q}(\omega)$, and so $\omega = 0$ cannot be a maxima of $e_{p,q}(\omega)$, global or otherwise).

Using (28), it is straightforward to show that $e'_{p,q}(\omega) = d_{p,q}(\omega)\,\bar{e}'_{p,q}(\omega)$, where:

$$\bar{e}'_{p,q}(\omega) = \begin{cases} e^{-c_p'\omega^p} - 2c_p'^{\,1-\frac{1}{p}}\omega^{p-1}\urcorner_{0,p}\left(c_p';\omega\right) & q=1 \\ \frac{1}{p}\urcorner_{0,p}\left(c_p';\omega\right) - c_p'^{\,\frac{1}{p}}e^{-c_p'\omega^p}\frac{\urcorner_{2,p}\left(c_p';\omega\right)}{\urcorner_{1,p}\left(c_p';\omega\right)} & q=2 \\ c_p'^{\,\frac{1}{p}}\frac{(q-1)}{(q-2)}\frac{\urcorner_{q-2,p}\left(c_p';\omega\right)}{\urcorner_{q-3,p}\left(c_p';\omega\right)} - c_p'^{\,\frac{1}{p}}\frac{\urcorner_{2q-2,p}\left(c_p';\omega\right)}{\urcorner_{2q-3,p}\left(c_p';\omega\right)} & q\geq 3 \end{cases} \tag{31}$$

[7] As $\omega = 0$ is at the end of the range $\omega \in [0,\infty)$ it follows that if the gradient is negative at $\omega = 0$ then there will be a local maxima at that point.

and:

$$d_{q,p}(\omega)=\begin{cases}\frac{pc_p'^{\frac{1}{p}}e^{-2c_p'\omega^p}}{\Gamma\left(2-\frac{1}{p}\right)\urcorner_{2,p}\left(c_p';\omega\right)} & q=1\\ \frac{2\Gamma\left(\frac{1}{p}\right)}{p\Gamma\left(\frac{3}{p}\right)}\frac{\urcorner_{0,p}\left(c_p';\omega\right)\urcorner_{1,p}\left(c_p';\omega\right)}{\Gamma\left(2-\frac{1}{p}\right)\urcorner_{2,p}\left(c_p';\omega\right)} & q=2\\ \frac{2(q-1)^2(q-2)\Gamma^{\frac{3}{2}}\left(\frac{1}{p}\right)}{p^2\Gamma^{\frac{3}{2}}\left(\frac{3}{p}\right)}\cdots & \\ \cdots\frac{\urcorner_{q-2,p}\left(c_p';\omega\right)\urcorner_{q-3,p}\left(c_p';\omega\right)\urcorner_{2q-3,p}\left(c_p';\omega\right)}{\Gamma\left(2-\frac{1}{p}\right)\urcorner_{2q-2,p}^2\left(c_p';\omega\right)} & q\geq 3\end{cases}$$

is a positive smooth function for $\omega\geq 0$, $p,q\in\mathbb{Z}^+$. Hence if $\bar{e}'_{p,q}(0)>0$ for $p>q\in\mathbb{Z}^+$ then $e'_{p,q}(0)>0$ for $p>q\in\mathbb{Z}^+$, which is sufficient to prove the theorem. Using (28), for all $p,q\in\mathbb{Z}^+$:

$$\bar{e}'_{p,q}(0)=\begin{cases}1 & \text{if } q=1\\ \frac{1}{p}\Gamma\left(\frac{1}{p}\right)-\frac{\Gamma\left(\frac{3}{p}\right)}{\Gamma\left(\frac{2}{p}\right)} & \text{if } q=2\\ \frac{q-1}{q-2}\frac{\Gamma\left(\frac{q-1}{p}\right)}{\Gamma\left(\frac{q-2}{p}\right)}-\frac{\Gamma\left(\frac{2q-1}{p}\right)}{\Gamma\left(\frac{2q-2}{p}\right)} & \text{if } q\geq 3\end{cases}\tag{32}$$

Which proves the theorem in the case $q=1$. Consider the case $p>q>2$. Writing $q=2+m$, $p=2+m+n$, where $m,n\in\mathbb{Z}^+$, and using the result $x\Gamma(x)=\Gamma(x+1)$, if $q\geq 3$:

$$\begin{aligned}\bar{e}'_{p,q}(0) &= \frac{\frac{q-1}{p}}{\frac{q-2}{p}}\frac{\Gamma\left(\frac{q-1}{p}\right)}{\Gamma\left(\frac{q-2}{p}\right)}-\frac{\Gamma\left(\frac{2q-1}{p}\right)}{\Gamma\left(\frac{2q-2}{p}\right)}\\ &= \frac{\Gamma\left(\frac{2m+n+3}{m+n+2}\right)}{\Gamma\left(\frac{2m+n+2}{m+n+2}\right)}-\frac{\Gamma\left(\frac{2m+3}{m+n+2}\right)}{\Gamma\left(\frac{2m+2}{m+n+2}\right)}\\ &= \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)}\left(B(a+c,b)-B(a,b+c)\right)\end{aligned}$$

where $a=\frac{2m+n+2}{m+n+2}>0$, $b=\frac{2m+2}{m+n+2}>0$ and $c=\frac{1}{m+n+2}>0$. As $a>b$, it follows from theorem 13 that $B(a+c,b)>B(a,b+c)$, and hence $\bar{e}'(0)>0$, which proves the theorem for the case $q\geq 3$.

Suppose that n (and subsequently q) is treated as a real number such that $n\geq 0$ (so $q\geq 2$). The inequality $B(a+c,b)>B(a,b+c)$ will still hold, as $a>b$ for all $n\in[0,\infty)$, $m>0$. Hence $\lim_{q\to 2^+}\bar{e}'_{p,q}(0)>0$. Furthermore, using theorem 12:

$$\lim_{q\to 2^+}\frac{q-1}{q-2}\frac{\Gamma\left(\frac{q-1}{p}\right)}{\Gamma\left(\frac{q-2}{p}\right)}-\frac{\Gamma\left(\frac{2q-1}{p}\right)}{\Gamma\left(\frac{2q-2}{p}\right)}=\frac{1}{p}\Gamma\left(\frac{1}{p}\right)-\frac{\Gamma\left(\frac{3}{p}\right)}{\Gamma\left(\frac{2}{p}\right)}$$

which implies that:

$$\bar{e}'_{p,2}(0)=\lim_{q\to 2^+}\bar{e}'_{p,q}(0)>0$$

and so $\bar{e}'_{p,q}(0)>0$ for all $p>q\in\mathbb{Z}^+$. Hence $e'_{p,q}(0)>0$ for all $p>q\in\mathbb{Z}^+$, which proves the theorem. ■

Using an analogous method, we can also give a partial proof of conjecture 11:

Partial proof of conjecture 11: To prove that $\epsilon_{\text{opt}}=0$ it is necessary (although insufficient) to prove that $e_{p,q}(\omega)$ has a local maxima at $\omega=0$. By analogy with the proof of theorem 10 it is necessary to prove that the gradient $e'_{p,q}(\omega)$ of $e_{p,q}(\omega)$ at $\omega=0$ is non-positive for all $p<q$. From the proof of theorem 10, $e'_{p,q}(\omega)=d_{p,q}(\omega)\,\bar{e}'_{p,q}(\omega)$, $d_{p,q}(\omega)>0$, for all $\omega\geq 0$, $p,q\in\mathbb{Z}^+$, where $\bar{e}'_{p,q}(\omega)$ is given by (31). Hence, if $\bar{e}'_{p,q}(0)\leq 0$, $e'_{p,q}(0)\leq 0$, and $e_{p,q}(\omega)$ will have a local maxima at $\omega=0$.

As $1<p<q$, $q\geq 2$. If $q=2$, $p=1$, and hence by (32) $\bar{e}'_{1,2}(0)=-1$. Therefore $e_{p,2}(\omega)$ has a local maxima at $\omega=0$. If $q\geq 3$, writing $q=2+m$, $p=2+m-n$, where $m\in\mathbb{Z}^+$, $n\in\{1,2,\ldots,m+1\}$, it can be seen that if $p>q>2$:

$$\bar{e}'_{p,q}(0)=\frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)}\left(B(a+c,b)-B(a,b+c)\right)$$

where $a=\frac{2m-n+2}{m-n+2}>0$, $b=\frac{2m+2}{m-n+2}>0$ and $c=\frac{1}{m-n+2}>0$. As $a<b$, it follows from theorem 13 that $B(a+c,b)<B(a,b+c)$, and hence $\bar{e}'(0)<0$. Therefore, in general, $e_{p,q}(\omega)$ has a local maxima at $\omega=0$ for all $p>q\in\mathbb{Z}^+$.

Now, if we could prove that this is a *unique* (and hence global) maxima, it would follow that $\omega_{\text{opt}}=\epsilon_{\text{opt}}=0$, which would prove the conjecture. Alternatively, proving that $e'_{p,q}(\omega)\leq 0$ for all $\omega\geq 0$ would demonstrate that the maxima at $\omega=0$ is global (although not necessarily unique) and thereby prove the conjecture. Unfortunately we have been unable to do either rigorously, and so the proof remains incomplete. ■

REFERENCES

[1] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds., vol. 9. The MIT Press, 1997, p. 155.

[2] V. Vapnik, S. Golowich, and A. Smola, *Advances in Neural Information Processing Systems*. MIT Press, 1997, vol. 9, ch. Support Vector Methods for Function Approximation, Regression Estimation, and Signal Processing, pp. 281–187.

[3] A. Smola and B. Schölkopf, "A tutorial on support vector regression, Tech. Rep. NeuroCOLT2 Technical Report Series, NC2-TR-1998-030, Oct 1998.

[4] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] I. Guyon, B. Boser, and V. Vapnik, "Automatic capacity tuning of very large VC-dimension classifiers," in *Advances in Neural Information Processing Systems*, S. Hanson, J. Cowan, and C. Giles, Eds., vol. 5. Morgan Kaufmann, 1993, pp. 147–155.

[6] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery and Data Mining*, vol. 2, pp. 121–167, 1998.

[7] B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson, "Shrinking the tube: A new support vector regression algorithm," *Advances in Neural Information Processing Systems*, vol. 11, pp. 330–336, 1999.

[8] B. Schölkopf and A. J. Smola, "New support vector algorithms, Tech. Rep. NeuroCOLT2 Technical Report Series, NC2-TR-1998-031, Nov 1998.

[9] A. Smola, N. Murata, B. Schölkopf, and K.-R. Muller, "Asymptotically optimal choice of ϵ-loss for support vector machines," in *Proceedings of the 8th International Conference on Artificial Neural Networks - Perspectives in Neural Computing*, L. Niklasson, M. Boden, and T. Ziemke, Eds. Springer Verlag, 1998, pp. 105–110.

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

[11] ——, "Weighted least squares support vector machines: robustness and sparse approximation," *Neurocomputing, Special issue on fundamental and information processing aspects of neurocomputing*, vol. 48, no. 1–4, pp. 85–105, Oct 2002.

[12] A. Shilton and M. Palaniswami, "A modified ν-SV method for simplified regression," in *Proceedings of the International Conference on Intelligent Sensing and Information Processing*, 2004, pp. 422–427.

[13] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Transactions of the Royal Society of London*, vol. 209, no. A, 1909.

[14] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, Massachusetts: MIT Press, 2001.

[15] A. Smola, B. Schölkopf, and K. Muller, "General cost functions for support vector regression," in *Proceedings of the Ninth Australian Conf. on Neural Networks*, T. Downs, M. Frean, and M. Gallagher, Eds., 1998, pp. 79 – 83.

[16] C. Chang and C. Lin, "LIBSVM: a library for support vector machines (version 2.3)," 2001.

[17] M. Palaniswami and A. Shilton, "Adaptive support vector machines for regression," in *Proceedings of the 9th International Conference on Neural Information Processing*, 2000.

[18] R. Fletcher, *Practical Methods of Optimisation*. Chichester: John Wiley and Sons, 1981.

[19] K. Levenberg, "A method for the solution of certain problems in least squares," *Quart. Appl. Math.*, vol. 2, pp. 164–168, 1944.

[20] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Appl. Math.*, vol. 11, pp. 431–441, 1963.

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

[22] N. Murata, S. Yoshizawa, and S.-I. Amari, "Network Information Criterion—determining the number of hidden units for an artificial neural network model," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 865–872, November 1994. [Online]. Available: citeseer.nj.nec.com/murata94network.html

[23] S. G. Krantz, *Handbook of Complex Variables*. Birkhusers, 1999.

[24] A. Chalimourda, B. Schölkopf, and A. Smola, "Experimentally optimal ν in support vector regression for different noise models and parameter settings," *Neural Networks*, vol. 17, no. 1, pp. 127–141, 2004.

[25] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions*. New York: Krieger, 1981, vol. 1, ch. 1.5 The Beta Function, pp. 9–13.

[26] H. Bateman, *Higher Transcendental Functions*. McGraw-Hill Book Company Inc., 1953, vol. 1.

Alistair Shilton received his combined B.Sc. / B.Eng. degree from the University of Melbourne, Melbourne, Australia, in 2000, specialising in physics, applied mathematics and electronic engineering. He is currently pursuing the Ph.D. degree in electronic engineering at the University of Melbourne. His research interests include machine learning, specializing in support vector machines; signal processing, communications and differential topology.

Daniel Lai received his B.Eng degree in Electrical and Computer Systems from Monash University, Melbourne, Australia in 2002. He is currently pursuing his Ph.D at Monash University. His research interests include optimization techniques for machine learning formulations, decomposition techniques, evolutionary optimization and dynamic programming.

M. Palaniswami received his BE(Hons) from the University of Madras, ME from the Indian Institute of science, India, MEngSc from the University of Melbourne and Ph.D from the University of Newcastle, Australia before rejoining the University of Melbourne. He has been serving the University of Melbourne for over 16 years. He has published more than 180 refereed papers and a huge proportion of them appeared in prestigious IEEE Journals and Conferences. He was given a Foreign Specialist Award by the Ministry of Education, Japan in recognition of his contributions to the field of Machine Learning. He served as associate editor for Journals/transactions including IEEE Transactions on neural Networks and Computational Intelligence for Finance.. His research interests include SVMs, Sensors and Sensor Networks, Machine Learning, Neural Network, Pattern Recognition, Signal Processing and Control. He is the Co-Director of Centre of Expertise on Networked Decision & Sensor Systems. He is the associate editor for International Journal of Computational Intelligence and Applications and serves on the editorial board of ANZ Journal on Intelligent Information processing Systems. He is also the Subject Editor for International Journal on Distributed Sensor Networks.