SVM Multiregression for Nonlinear Channel Estimation in Multiple-Input Multiple-Output Systems

Matilde Sánchez-Fernández, *Member, IEEE*, Mario de-Prado-Cumplido, *Student Member, IEEE*, Jerónimo Arenas-García, *Student Member, IEEE*, and Fernando Pérez-Cruz, *Member, IEEE*

***Abstract*—This paper addresses the problem of multiple-input multiple-output (MIMO) frequency nonselective channel estimation. We develop a new method for multiple variable regression estimation based on Support Vector Machines (SVMs): a state-of-the-art technique within the machine learning community for regression estimation. We show how this new method, which we call M-SVR, can be efficiently applied. The proposed regression method is evaluated in a MIMO system under a channel estimation scenario, showing its benefits in comparison to previous proposals when nonlinearities are present in either the transmitter or the receiver sides of the MIMO system.**

***Index Terms*—Channel estimation, MIMO systems, multivariate regression, support vector machine.**

I. Introduction

THE aim for increasing capacity and quality of service in wireless systems is drawing considerable attention toward multiple-input multiple-output (MIMO) systems that exploit the spatial dimension and the scattering properties of most radio channels. Although noncoherent MIMO techniques have been addressed [1], channel estimation and coherent techniques are the way to achieve the capacity gain claimed. Thus, channel compensation issues, including intersymbol interference (ISI) and fading, have been addressed through different approaches in several recent publications [2]–[6], either assuming perfect channel estimation or data-aided or blind solutions.

Support vector machines (SVMs) are state-of-the-art tools for linear and nonlinear input–output knowledge discovery [7], [8]. SVMs were first devised for binary classification problems [9], and they were later extended for regression estimation problems [10], [11], among others. Although the first schemes to solve SVMs used quadratic programming, iterative reweighted least square (IRWLS) solutions are generally faster [12], allow the introduction of arbitrary cost functions in the SVM functional [13], and are straightforward extensible to adaptive schemes [14].

Nonlinear channel compensation techniques [15]–[18] and, particularly, SVM-based methods [19]–[22], have been undertaken in previous works, in most cases addressing the channel estimation problem within a single-input single-output (SISO) perspective.

Previous data-aided solutions for channel estimation issues in MIMO systems are mainly developed for flat fading channels and are based either on maximum likelihood (ML) [5] or minimum mean square error (MMSE) [4] channel estimation. In this paper, we propose a new data-aided solution that takes advantage of the MIMO channel multidimensionality by means of a regression tool, which has its roots in SVMs [7]. In the proposed solution, two assumptions are made: No ISI is present in the channel model, and fading is slow enough in order to consider that the channel remains stationary during the estimation interval. Any other consideration regarding the channel variability would lead to an adaptive extension of the method, similar to the one described for classification problems in [14]. In addition to the previous assumptions, channel nonlinearities [21], [23] might be considered either in transmission or reception. When facing nonlinear problems, SVMs show their benefits.

Thus, we propose an IRWLS-based approach for the regression of multiple variables [SVM multiregressor (M-SVR)], which is then applied to the data-aided MIMO channel estimation problem. We show that the proposed technique gives some advantages in nonlinear channels, regarding bit error rate (BER) and complexity in comparison with a radial basis function networks (RBFN) based approach and an unidimensional support vector regressor (SVR), respectively. Still, when applied to linear channels with white Gaussian noise, M-SVR converges to the MMSE solution, which is optimal in this case.

The rest of the paper is organized as follows: Section II addresses separately a general model for a MIMO system with channel nonlinearities and nonlinear channel estimation issues. Section III introduces the multidimensional regression approach based on SVM. Section IV is devoted to some computer experiments where MIMO channels with nonlinearities in transmission or reception will be the studied scenarios in order to provide a fair environment for comparison of nonlinear methods. A linear channel model will also be considered to compare M-SVR to the optimal MMSE solution. Finally, in Section V, we conclude the paper with some discussion about the obtained results and propose some lines for further work. In

Manuscript received July 1, 2003; revised March 1, 2004. This work has been supported in part by Grants CAM 07T/0016/2003 and CYCIT TIC2002-03498. F. Pérez-Cruz is supported by the Spanish Ministry of Education Postdoctoral fellowship EX2003-0437. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hans-Andrea Loeliger.

M. Sánchez-Fernández, M. de-Prado-Cumplido, and J. Arenas-García are with the Department Teoría de la Señal y Comunicaciones, University Carlos III de Madrid, 28911 Leganés-Madrid, Spain (e-mail: mati@tsc.uc3m.es; mprado@tsc.uc3m.es; jarenas@tsc.uc3m.es).

F. Pérez-Cruz is with the Department Teorí de la Señal y Comunicaciones, University Carlos III de Madrid, A28911 Leganés-Madrid, Spain, and also with the Gatsby Computational Neuroscience Unit, University College London, London, WC1N 3AR U.K. (e-mail: fernando@tsc.uc3m.es; fernando@gatsby.ucl.ac.uk).

Digital Object Identifier 10.1109/TSP.2004.831028

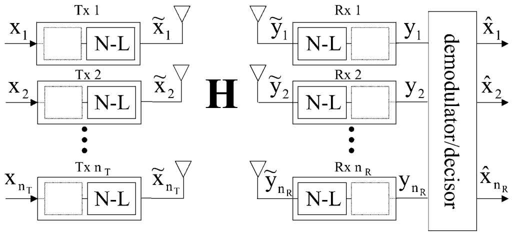


Fig. 1. MIMO system with n_T transmitting and n_R receiving antennas. The channel is modeled with a matrix $\mathbf{H}$ of size $n_R \times n_T$, where element h_{ij} corresponds to the attenuation between the jth transmitting and the ith receiving antennas.

the Appendix, we give a proof of convergence for the proposed algorithm.

II. Nonlinear Channel Estimation for MIMO Systems

A. Nonlinear Channel Models

Nonlinearities might be present at the transmission-reception chain at two points: in the front-end transmitter and receiver, leading to an equivalent nonlinear channel model even if the channel propagation model is linear. In previous works [21], [23], channel nonlinearities have been considered in the study of nonlinear channel estimation for single-input single-output (SISO) systems. This nonlinear SISO model will be generalized to a MIMO model, as detailed next.

The propagation channel model we focus in this paper is based on a linear MIMO system with n_T transmitting antennas and n_R receiving antennas. We use a matrix of independent complex Gaussian coefficients $\mathbf{H}$ to model the frequency nonselective Rayleigh channel in a baseband equivalent model (see Fig. 1). The general hypothesis for channel coefficients is a set of independent and identically distributed (i.i.d.) variables [24]. Complex white Gaussian noise is assumed in the channel, being modeled through a noise vector $\mathbf{n}$ of dimension n_R.

Within this propagation model, and without loss of generality, nonlinearities will be considered identically affecting either to each transmitter or to each receiver module from the MIMO system. Thus, both cases will be treated separately, leading to two different channel models.

Assuming that $\mathbf{x} = [x_1, \ldots, x_{n_T}]^T$ is the information signal in each time sample modeled by a quadrature phase shift keying (QPSK) baseband equivalent, nonlinearities in transmission lead to a system equation where transmitted symbols follow the nonlinear rules in [21]:

$$\widetilde{\mathbf{x}} = \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \\ \vdots \\ \widetilde{x}_{n_T} \end{bmatrix} = \begin{bmatrix} x_1 + \alpha_1 x_1^2 + \alpha_2 x_1^3 \\ x_2 + \alpha_1 x_2^2 + \alpha_2 x_2^3 \\ \vdots \\ x_{n_T} + \alpha_1 x_{n_T}^2 + \alpha_2 x_{n_T}^3 \end{bmatrix}. \tag{1}$$

The received symbols $\widetilde{\mathbf{y}}$ are linear mixtures of the transmitted symbols by means of the linear channel propagation model of the following system equation:

$$\widetilde{\mathbf{y}} = \frac{1}{n_T}\mathbf{H}\widetilde{\mathbf{x}} + \mathbf{n}. \tag{2}$$

Within this first channel model with nonlinearities in transmission, the reception procedure is considered all linear, and this implies that $\mathbf{y} = \widetilde{\mathbf{y}}$.

The second channel proposed, with nonlinearities in reception, leads to a different channel model. Information symbols are now transmitted $\widetilde{\mathbf{x}} = \mathbf{x}$ and linearly mixed by means of the following equation:

$$\widetilde{\mathbf{y}} = \frac{1}{n_T}\mathbf{H}\mathbf{x} + \mathbf{n}. \tag{3}$$

In this channel model, the nonlinearities modify the received symbols as follows:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_R} \end{bmatrix} = \begin{bmatrix} \widetilde{y}_1 + \alpha_1 \widetilde{y}_1^2 + \alpha_2 \widetilde{y}_1^3 \\ \widetilde{y}_2 + \alpha_1 \widetilde{y}_2^2 + \alpha_2 \widetilde{y}_2^3 \\ \vdots \\ \widetilde{y}_{n_R} + \alpha_1 \widetilde{y}_{n_R}^2 + \alpha_2 \widetilde{y}_{n_R}^3 \end{bmatrix}. \tag{4}$$

The two channel models proposed address two different problems in complexity. In the first channel model, at the receiver end, there is a linear mixture of the transmitted symbols, even though each of them is a nonlinear function of the information symbols; noise considered here is Gaussian and white. On the other hand, the second channel model leads to nonlinear mixtures of all the transmitted symbols at the receiver, and the noise can no longer be considered Gaussian.

B. Nonlinear Techniques for Channel Estimation

The MIMO pilot-based channel estimation we consider in this paper can be faced with many tools for data regression. In general, the regression problem consists of estimating an unknown function $y = f(\mathbf{x})$ from some given occurrences $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{x}_i \in \mathbb{R}^d$ and their corresponding targets $\{y_1, \ldots, y_n\}$ and $y_i \in \mathbb{R}$. This way, a MIMO channel, for n_T transmitting and n_R receiving antennas, can be simply estimated from the regression of the n_R received symbols corresponding to each transmitted word in the pilot sequence. Usually, some parametric form is assumed for the estimation $(\hat{f}(\mathbf{x}))$, and its parameters are adjusted according to some selected criteria or cost function (for instance, the MSE).

The most widely used method for data-aided channel estimation simply uses a linear combination of the components in $\mathbf{x}$, selecting the weight of each component in order to minimize the quadratic error:

$$E = \sum_{i=1}^{n} \left(\hat{f}(\mathbf{x}_i) - y_i\right)^2. \tag{5}$$

Although the MMSE method is optimal (and quite efficient if short pilot sequences are used) for estimating linear channels with additive Gaussian noise, its performance can be very poor when either of these two hypothesis are not satisfied.

Some nonlinear techniques can also be found in the communications literature. In [15], an adaptive version of RBFNs is used to estimate the time-variant channel of an orthogonal frequency-division multiplexing (OFDM) system. RBFN implements a function of the form

$$\hat{f}(\mathbf{x}) = y = \sum_{m=1}^{M} w_m h_m \left(\|\mathbf{x} - \mathbf{c}_m\|_2\right) \tag{6}$$

where M is the number of nodes in the network, $h_m(x)$ are radial functions (i.e., their values only depend on the distance between $\mathbf{x}$ and a centroid or prototype $\mathbf{c}_m$), and w_m are the weights assigned to each of these functions.

Different forms of the radial basis functions $h_m(x)$ can be used, but we will only consider the typical Gaussian RBF:

$$h_m(x) = \exp\left(-\frac{x^2}{2\sigma_m^2}\right). \quad (7)$$

Once M is selected, $\mathbf{c}_m$, σ_m, and w_m are free parameters that must be estimated from the training set. This phase usually includes a minimization of the quadratic error using a gradient scheme. In [15], a stochastic algorithm is used to provide the network with adaption capabilities. It is also advisable to include a regularization term to the cost function:

$$E = \sum_{i=1}^{n}\left(\hat{f}(\mathbf{x}_i) - y_i\right)^2 + \frac{\lambda}{\omega}(\widehat{f}). \quad (8)$$

where $\Omega(\widehat{f})$ is a regularizer [25] that penalizes large values in $\widehat{f}$, and λ is a parameter that controls the tradeoff between error and smoothness of the solution.

RBFNs are powerful architectures that can approximate to an arbitrary precision any function [25], [26]. Thus, we have chosen them as a reference with which to compare the performance of our proposal in this paper. We have used the excellent implementation of RBFNs by Rätsch (see, for instance, [27]).

As we have previously stated, the MIMO channel estimation problem can be solved with a collection of n_R unidimensional regressors. Therefore, as a second method to perform nonlinear channel estimation, we will consider the case in which each of these regressors is implemented using an SVR [28], which we briefly describe in Section III.

Next, we will propose a new method for nonlinear channel estimation relying on SVM technology. Instead of building a different regressor for each variable in the receiver, it considers all of them at the same time, which leads to simpler solutions (whose complexity does not increase with the number of receiving antennas, n_R), while keeping the high performance of SVM methods.

III. Multiregression SVM

In this paper, we introduce a generalization of SVR to solve the problem of regression estimation for multiple variables. Thus, we refer to our proposal, which is based on a previous contribution [29], as M-SVR. Here, M-SVR is considered for discovering the dependencies between transmitted and received signals in a MIMO system.

Although under a pure Gaussian perspective the estimation of each component can be individually addressed without loss, the use of a multidimensional regression tool will help to exploit the dependencies in the channel and will make each estimate less vulnerable to the added noise. Treating all the channel paths together will allow to accurately estimate each of them when only scarce data is available, and the ε-insensitive cost function, which will be introduced in short, will improve the scheme robustness when different kinds of noise and nonlinearities appear in the system.

As stated in Section II, the unidimensional regression estimation problem is regarded as finding the mapping between an incoming vector $\mathbf{x} \in \mathbb{R}^d$ and an observable output $y \in \mathbb{R}$ from a given set of i.i.d. samples $\{(\mathbf{x}_i, y_i)\}_{i=0}^n$. The standard SVR [8], [28] solves this problem by finding the regressor $\mathbf{w}$ and b that minimizes $\|\mathbf{w}\|^2/2 + C\sum_{i=1}^n L_v(y_i - (\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b))$, where $\boldsymbol{\phi}(\cdot)$ is a nonlinear transformation to a higher dimensional space, which is also known as the feature space ($\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{\mathcal{H}}$ and $\mathcal{H} \geq d$). The SVR can be solved using only inner products between $\boldsymbol{\phi}(\cdot)$, where we do not need to know the nonlinear mapping; therefore, we only need to specify a kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}^T(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)$ that has to fulfill Mercer Theorem [8]. $L_v(\cdot)$ is known as the Vapnik ε-insensitive loss-function, which is equal to 0 for $|y_i - (\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b)| < \varepsilon$ and equal to $|y_i - (\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b)| - \varepsilon$ for $|y_i - (\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b)| \geq \varepsilon$. The solution ($\mathbf{w}$ and b) is formed by a linear combination of the training samples in the transformed space that presents an absolute error equal or greater than ε.

In the case the observable output is a vector $\mathbf{y} \in \mathbb{R}^Q$, we need to solve a multidimensional regression estimation problem in which we have to find a regressor $\mathbf{w}^j$ and b^j $(j = 1, \ldots, Q)$ for every output. We can directly generalize the one-dimensional SVR to solve the multidimensional case, leading to the minimization of

$$L_P(\mathbf{W}, \mathbf{b}) = \frac{1}{2}\sum_{j=1}^{Q}\|\mathbf{w}^j\|^2 + C\sum_{i=1}^{n} L(u_i) \quad (9)$$

where $\mathbf{W}$, $\mathbf{b}$, and u_i will be defined shortly.

The Vapnik ε-insensitive loss function can be extended to multiple dimensions, but being based on an L_1 norm, it will need to account for each dimension independently, which will make the solution complexity grow linearly with the number of dimensions. If, instead, we use a L_2-based norm, all dimensions can be considered in an unique restriction, yielding a single support vector for all dimensions. Therefore, we propose to use

$$L(u) = \begin{cases} 0, & u < \varepsilon \\ u^2 - 2u\varepsilon + \varepsilon^2, & u \geq \varepsilon \end{cases} \quad (10)$$

which is a differentiable version of the loss function proposed in [29]. In the above expression, $u_i = \|\mathbf{e}_i\| = \sqrt{\mathbf{e}_i^T\mathbf{e}_i}$, $\mathbf{e}_i^T = \mathbf{y}_i^T - \boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{W} - \mathbf{b}^T$, $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^Q]$, $\mathbf{b} = [b^1, \ldots, b^Q]^T$, and $\boldsymbol{\phi}(\cdot)$ is a nonlinear transformation to the feature space.

For $\varepsilon = 0$, this problem reduces to an independent regularized kernel least square regression for each component, but for a nonzero ε, the solution will take into account all outputs to construct each individual regressor and will be able to obtain more robust predictions. The price to be paid is that the resolution of the proposed problem cannot be done straightforwardly, and we will have to rely on an iterative procedure to obtain the desired solution. We have devised a quasi-Newton approach in which each iteration has at most the same computational complexity as a least square procedure for each component. It is a weighted least square problem, and the number of iterations needed to obtain the final result is small, making the procedure only slightly more computationally demanding than the least square regression for each component. Therefore, we refer to it as an IRWLS procedure [12], [30].

A. Resolution of M-SVR

Optimization problems are solved using iterative procedures that rely in each iteration on the previous solution ($\mathbf{W}^k$ and $\mathbf{b}^k$, in our case) to obtain the following one, until the optimal solution is reached. To construct the IRWLS procedure, we modify (9) using a first-order Taylor expansion of $L(u)$ over the previous solution, leading to

$$L'_P(\mathbf{W},\mathbf{b}) = \frac{1}{2}\sum_{j=1}^{Q}\|\mathbf{w}^j\|^2 + C\left(\sum_{i=1}^{n} L\left(u_i^k\right) + \left.\frac{dL(u)}{du}\right|_{u_i^k} \frac{\left(\mathbf{e}_i^k\right)^T}{u_i^k}\left[\mathbf{e}_i - \mathbf{e}_i^k\right]\right) \quad (11)$$

where $u_i^k = \|\mathbf{e}_i^k\| = \sqrt{(\mathbf{e}_i^k)^T\mathbf{e}_i^k}$, $(\mathbf{e}_i^k)^T = \mathbf{y}_i^T - \boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{W}^k - (\mathbf{b}^k)^T$, which presents the same value and gradient as $L_P(\mathbf{W},\mathbf{b})$ for $\mathbf{W} = \mathbf{W}^k$ and $\mathbf{b} = \mathbf{b}^k$ (i.e., $L'_P(\mathbf{W}^k,\mathbf{b}^k) = L_P(\mathbf{W}^k,\mathbf{b}^k)$ and $\nabla L'_P(\mathbf{W}^k,\mathbf{b}^k) = \nabla L_P(\mathbf{W}^k,\mathbf{b}^k)$). $L'_P(\mathbf{W}^k,\mathbf{b}^k)$ is a lower bound of $L_P(\mathbf{W},\mathbf{b})$ (i.e., $L_P(\mathbf{W},\mathbf{b}) \geq L'_P(\mathbf{W},\mathbf{b})$, $\forall\ \mathbf{W} \in \mathbb{R}^{\mathcal{H}} \times \mathbb{R}^Q$) and $\forall\ \mathbf{b} \in \mathbb{R}^Q$ because $L'_P(\mathbf{W},\mathbf{b})$ is a first-order Taylor expansion of a convex function.

Now, we are going to construct a quadratic approximation from (11):

$$\begin{aligned} L''_P(\mathbf{W},\mathbf{b}) &= \frac{1}{2}\sum_{j=1}^{Q}\|\mathbf{w}^j\|^2 + C\left(\sum_{i=1}^{n} L(u_i^k) + \left.\frac{dL(u)}{du}\right|_{u_i^k}\frac{u_i^2 - (u_i^k)^2}{2u_i^k}\right) \\ &= \frac{1}{2}\sum_{j=1}^{Q}\|\mathbf{w}^j\|^2 + \frac{1}{2}\sum_{i=1}^{n} a_i u_i^2 + CT \end{aligned} \quad (12)$$

where

$$a_i = \frac{C}{u_i^k}\left.\frac{dL(u)}{du}\right|_{u_i^k} = \begin{cases} 0, & u_i^k < \varepsilon \\ \frac{2C(u_i^k - \varepsilon)}{u_i^k}, & u_i^k \geq \varepsilon \end{cases} \quad (13)$$

and CT is a sum of constant terms that do not depend either on $\mathbf{W}$ or $\mathbf{b}$, which also presents the same value and gradient as $L_P(\mathbf{W},\mathbf{b})$ for $\mathbf{W} = \mathbf{W}^k$ and $\mathbf{b} = \mathbf{b}^k$. It can be seen that (12) is a weighted least square problem in which the weights depend on the previous solution, incorporating the knowledge of all the components of each $\mathbf{y}_i$. To optimize (9), we will construct a descending direction using the optimal solution of (12), and then, we will compute the next step solution using a line search algorithm [31]. The IRWLS procedure can be summarized in the following steps.

1) Initialization: Set $k = 0$, $\mathbf{W}^k = \mathbf{0}$, $\mathbf{b}^k = \mathbf{0}$, and compute u_i^k and a_i.
2) Compute the solution to (12), and label it as $\mathbf{W}^s$ and $\mathbf{b}^s$. Define a descending direction for (9) as $\mathbf{P}^k = \begin{bmatrix} \mathbf{W}^s - \mathbf{W}^k \\ (\mathbf{b}^s - \mathbf{b}^k)^T \end{bmatrix}$.
3) Obtain the next step solution $\begin{bmatrix} \mathbf{W}^{k+1} \\ (\mathbf{b}^{k+1})^T \end{bmatrix} = \begin{bmatrix} \mathbf{W}^k \\ (\mathbf{b}^k)^T \end{bmatrix} + \eta^k\mathbf{P}^k$, computing the step size η^k using a backtracking algorithm.
4) Compute u_i^{k+1} and a_i, set $k = k+1$, and go back to step 2 until convergence.

Before actually computing $\mathbf{W}^s$ and $\mathbf{b}^s$, we would like to explicitly say that $\mathbf{P}^k$ is not a vector but a matrix. Each column of $\mathbf{P}^k$ is a descending direction for each regressor; therefore, one should see it as an aggregate of descending directions for each component to be estimated. The value of η^k is computed using a backtracking algorithm [31], in which we initially set $\eta^k = 1$ (the initial choice for η^k will become apparent in the proof of convergence given in Appendix) and check if $L_P(\mathbf{W}^{k+1},\mathbf{b}^{k+1}) < L_P(\mathbf{W}^k,\mathbf{b}^k)$. If not, we multiply η_k by a positive constant less than one and repeat the procedure until a decrease is achieved in the minimizing functional.

To obtain $\mathbf{W}^s$ and $\mathbf{b}^s$, we need to solve the weighted least square problem in (12), in which each component is decoupled. Therefore, we can solve independently for each component by equating to zero its gradient

$$\nabla_{\mathbf{w}^j} L''_P = \mathbf{w}^j - \sum_i \boldsymbol{\phi}(\mathbf{x}_i)a_i(y_{ij} - \boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w}^j - b^j) = \mathbf{0} \qquad j = 1,\ldots,Q \quad (14)$$

$$\nabla_{b^j} L''_P = -\sum_i a_i(y_{ij} - \boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w}^j - b^j) = 0 \qquad j = 1,\ldots,Q \quad (15)$$

which can be expressed as a linear system of equations:

$$\begin{bmatrix} \boldsymbol{\Phi}^T\mathbf{D}_a\boldsymbol{\Phi} + \mathbf{I} & \boldsymbol{\Phi}^T\mathbf{a} \\ \mathbf{a}^T\boldsymbol{\Phi} & \mathbf{1}^T\mathbf{a} \end{bmatrix}\begin{bmatrix} \mathbf{w}^j \\ b^j \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}^T\mathbf{D}_a\mathbf{y}^j \\ \mathbf{a}^T\mathbf{y}^j \end{bmatrix} \qquad j = 1,\ldots,Q \quad (16)$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1),\ldots,\boldsymbol{\phi}(\mathbf{x}_n)]^T$, $\mathbf{a} = [a_1,\ldots,a_n]^T$, $(\mathbf{D}_a)_{ij} = a_i\delta(i-j)$, and $\mathbf{y}^j = [y_{1j},\ldots,y_{nj}]^T$. It can be seen that the matrix in the previous linear system does not depend on j; therefore, it will be identical for all components, and the difference on the linear systems associated with each pair $(\mathbf{w}^j, b^j)$ will be due to the independent term in (16). Each column of $\mathbf{W}^s$ and $\mathbf{b}^s$ will be constructed with the solutions of (16) for each j.

It is usual to work with the feature space kernel (inner product of the transformed vectors $(\kappa(\mathbf{x}_i,\mathbf{x}_j) = \boldsymbol{\phi}^T(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j))$ instead of the whole nonlinear mapping [7]. We are going to make use of the Representer Theorem [7], [32], which states that the best solution, under fairly general conditions, to a learning problem can be expressed as a linear combination of the training samples in the feature space, i.e., $\mathbf{w}^j = \sum_i \boldsymbol{\phi}(\mathbf{x}_i)\beta^j = \boldsymbol{\Phi}^T\boldsymbol{\beta}^j$. If we replace this expression into (14) and (15), the linear system in (16) can be expressed as follows:

$$\begin{bmatrix} \mathbf{K} + \mathbf{D}_a{}^{-1} & \mathbf{1} \\ \mathbf{a}^T\mathbf{K} & \mathbf{1}^T\mathbf{a} \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta}^j \\ b^j \end{bmatrix} = \begin{bmatrix} \mathbf{y}^j \\ \mathbf{a}^T\mathbf{y}^j \end{bmatrix} \qquad j = 1,\ldots,Q \quad (17)$$

where $(\mathbf{K})_{ij} = \kappa(\mathbf{x}_i,\mathbf{x}_j)$ is known as the kernel matrix. The line search algorithm can be readily expressed in terms of $\boldsymbol{\beta}^j$, as it was presented for $\mathbf{w}^j$.

We can now argue why a nonzero ε will take into account all the outputs to construct each individual regressor. If $\varepsilon = 0$, $a_i = 2C$ for every sample, and (17) only depends on each particular output through $\mathbf{y}^j$, but if $\varepsilon \neq 0$, a_i is a function of u_i^k, that is, the square error between every dimension of $\mathbf{y}_i$ and all the regressors. Consequently, M-SVR bounds together all the outputs when constructing each individual regressor. A proof of convergence of the proposed algorithm is given in the Appendix.

Once the channel has been estimated (the $\boldsymbol{\beta}^j$ have been computed), we cannot calculate the channel directly because it is a function of the nonlinear transformation $\boldsymbol{\phi}(\cdot)$, but for each new vector $\mathbf{x}$, we can compute the jth output as $y^j = \boldsymbol{\phi}^T(\mathbf{x})\boldsymbol{\Phi}^T\boldsymbol{\beta}^j$. Now, if we define the matrix $\boldsymbol{\beta} = [\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \ldots, \boldsymbol{\beta}^Q]$, the Q outputs can be computed as

$$\mathbf{y} = \boldsymbol{\phi}^T(\mathbf{x})\boldsymbol{\Phi}^T\boldsymbol{\beta} = \mathbf{K}_{\mathbf{x}}\boldsymbol{\beta}$$

where $\mathbf{K}_{\mathbf{x}}$ is a vector that contains the kernel of the input vector $\mathbf{x}$ and the training points.

IV. Computer Experiments

In this section, we present a number of computer experiments to show the benefits of our M-SVR algorithm when used in MIMO nonlinear channel estimation. We compare the performance of the nonlinear and linear regression algorithms described in Sections II-B and III for different channels, signal-to-noise ratios (SNRs) measured at the receiver inputs, and training sequence lengths.

The goal in the channel estimation problem is to obtain a good approximation to the actual channel, modeling the dependence between transmitted and received signals. With the MMSE method, this relation is restricted to be linear, and the channel estimate $\widehat{\mathbf{H}}$ can be explicitly given. This holds for the M-SVR and SVR methods with a linear kernel, but it is no longer possible when using a nonlinear transformation to a higher dimensional space (i.e., when applying other kernel that is different from the linear one) because in this case, we are only able to compute the kernels. The RBFN method operates analogously to kernel regressors.

In pilot-aided channel estimation, it is necessary to use a training sequence known *a priori* by both the transmitter and receiver. Once the channel has been modeled, the expected received vector $\mathbf{y}$ without noise, corresponding to each possible transmitted QPSK codeword $\mathbf{x}$, is calculated. During operation, each received signal is decoded using the nearest neighbor criterion.

We have used a Gaussian kernel for both SVR-based methods:

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

where σ is a tunable hyperparameter. The hyperparameter C [see (9)], which controls the tradeoff between the regularization term and the error reduction term, has been set to $C = 10^3$, which is a good compromise value in most cases. The remaining parameter ε that sets the width of the insensitivity zone of the regressor cost function is also tuned in the training phase.

Regarding the RBFN technique, we have used a Gaussian function for $h_m(x)$. In this method, all $\mathbf{c}_m$ and σ_m parameters are optimized during the training phase. We have trained networks with a number of centroids M, that is 5%, 10%, and 20% of the training sequence length, keeping for each case the best setting. We have checked that the performance degrades if we increase the number of centroids above 20%, requiring, in addition, a much heavier computational load. Regarding the number of gradient descent algorithm iterations, 40 rounds are enough for the algorithm to converge. Finally, we have also explored different values for the regularization constant λ. However, we found that no regularization is needed, except in the case where the estimated channel is linear, for which we have used $\lambda = 10^{-3}$. We think that the values of M used in the simulations guarantee that the solution is not overfitted, except for the linear channel, where the regressor needs to provide a linear solution, and extra regularization is required.

In Sections IV-A–C, we present the simulation results for MIMO systems ($n_T = 4$, $n_R = 3$) with the channels proposed in Section II-A. The number of test words has been chosen to assure that at most one erroneous bit occurs for each 100 bits received, and all results have been averaged over 100 trials.

A. *MIMO System with Nonlinearities in the Transmitter*

We first present results when the nonlinearities between input and output signals of the channel are introduced by the transmitter equipment, due to, for example, amplifiers driven near its saturation zone. The channel is the one described by (1) and (2), with coefficients $\alpha_1 = 0.2$ and $\alpha_2 = 0$ [21].

M-SVR is able to parameterize nonlinearities effectively, as it is seen in Fig. 2(a) and (b), and obtains lower BER than the RBFN for variable SNR. The improvement of our method is specially representative for short training lengths, although the difference is only slightly reduced for the longest training sets. The saturation point of the curves, for which the BER is no longer improved, increases as the SNR grows. In any case, this point is reached in first place by the M-SVR. For the sake of clearness in the reading of the figure, we have split the results in two plots, grouping in each one alternative SNRs.

We have also included the results for the SVR method, which use an L_1 cost function, instead of the L_2 used by the M-SVR. These results are in general slightly worse than the M-SVR solution. In spite of the fact that SVR algorithm performance is similar to the M-SVR method, its computational burden is much more intensive. While M-SVR requires just a few iterations of the IRWLS to converge (about five steps), SVR needs approximately two orders of magnitude more iterations. Besides, the complexity of SVR increases both with n_R and the length of the training sequence, whereas that of M-SVR does not depend on n_R.

B. *MIMO System with Nonlinearities in the Receiver*

For the test of channels with nonlinearities in the receiver, which are in general harder to tackle, we have run simulations for two scenarios.

First, coefficients of (3) and (4) are set to $\alpha_1 = 0$ and $\alpha_2 = -0.9$. In Fig. 3, we give BER curves versus pilot sequence lengths. For this highly nonlinear channel, the estimate given by the RBFN is clearly and consistently worse than the M-SVR solution, which seems to model the problem more accurately. The plotted curves correspond to an SNR of 10.5 dB; experiments with other noise levels present a similar performance.

The second channel uses $\alpha_1 = 0.2$ and $\alpha_2 = 0$. As shown in Fig. 4(a) and (b), results in this case are similar to those obtained in Section IV-A. However, they present a slightly higher BER

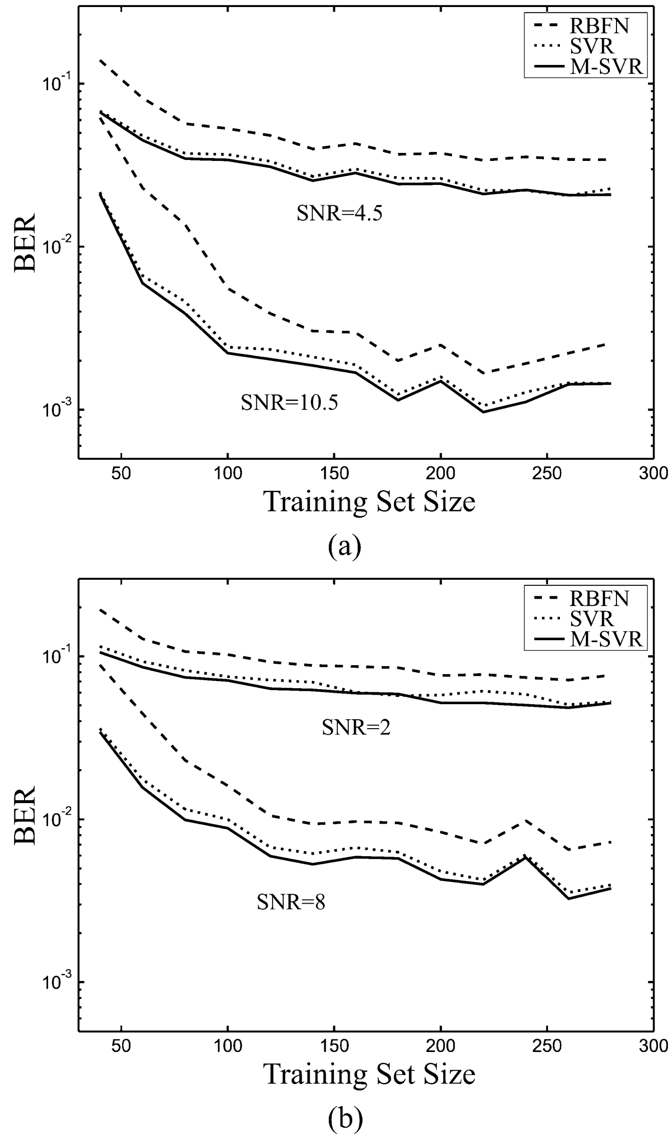


Fig. 2. M-SVR, SVR, and RBFN BERs as a function of the length of the training sequence used during the training phase; the curves depicted are generated with (a) SNR= 4.5 and 10.5 dB and (b) SNR = 2 and 8 dB. The nonlinearity phenomenon occurs in the transmitter, affecting then solely the input signals. Improvement of the SVR-based methods is evident.

because this case has a higher grade of nonlinearity, as discussed in Section II-A.

C. Linear MIMO System

Finally, we have carried out experiments for a linear channel with added Gaussian noise in order to check how M-SVR performs in comparison with MMSE, which is known to be optimal in this case (see Fig. 5). The analytical expression for the linear channel can be obtained by setting $\alpha_1 = \alpha_2 = 0$ in any of the previous nonlinear models. Results for both methods are almost identical, the slight advantage of MMSE being due to the fact that M-SVR makes no *a priori* assumption about the linearity of the channel. If we made use of this information, we could employ a linear kernel instead. We have carried out experiments (not included in the figures) that show that results obtained by M-SVR with a linear kernel are identical to those of

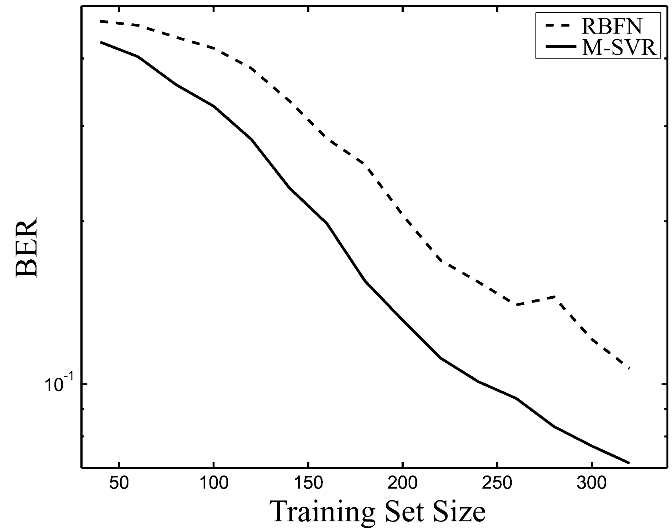


Fig. 3. M-SVR and RBFN BERs as a function of the length of the training sequence used during the training phase for a fixed SNR of 10.5 dB. The channel presents a very high nonlinear behavior.

MMSE. Again, RBFN exhibits an increase in BER with respect to M-SVR performance.

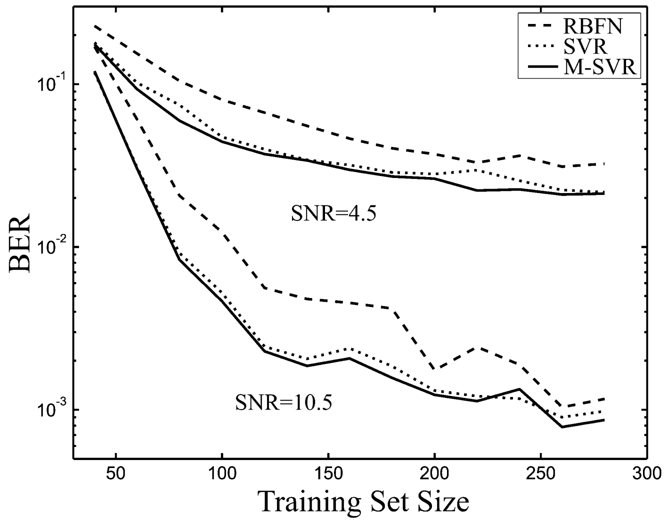
V. Discussion and Further Work

In this work, we have tackled channel estimation for MIMO systems. We have presented a new multivariate regression algorithm based on the machine learning state-of-the-art Support Vector Machines to solve this problem. The M-SVR algorithm takes advantage of the MIMO spatial diversity, and it is capable of discovering the dependencies between the transmitted and received signals. M-SVR can be used with nonlinear kernels, such as the Gaussian kernel, in order to effectively address nonlinear channel estimation.

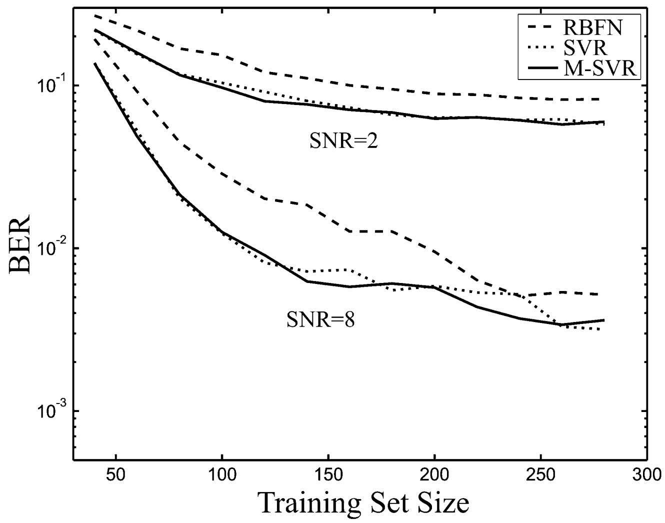
The theoretical aspects of M-SVR are fully developed, and a proof of its convergence is given. M-SVR requires a computational load that is comparable with that of other well-known methods such as the MMSE estimator. Moreover, M-SVR resolution lays in the IRWLS algorithm, which can be easily modified to use different cost functions or to confer it adaptive properties.

Simulation examples have been used to test our method and to favorably compare it with the standard SVR and to an RBFN method, which is applied independently over each dimension, employing nonlinear channel models. We have also compared it with the optimal MMSE strategy for linear channel models, yielding almost equivalent results.

Channels with ISI and the inclusion of decoding stages are logical further research lines, as well as simulations with different kinds of noises or the implementation of specific kernels suitable for MIMO communications. As mentioned before, the introduction of other cost functions in the learning algorithm and its modification into adaptive schemes for time-variant channels are also interesting possibilities. Finally, we believe it is relevant to mention that the proposed M-SVR algorithm can be extended to other signal processing problems such as sample imputation [33], device modeling [34] or chaotic systems [35], among others.



(a)



(b)

Fig. 4. M-SVR, SVR, and RBFN BERs as a function of the length of the pilot sequence for (a) SNR = 4.5 and 10.5 dB and (b) SNR = 2 and 8 dB. In this case, it is the receiver where the nonlinear effects appear in the system, affecting the input signals, the channel, and the noise.

APPENDIX
M-SVR PROOF OF CONVERGENCE

To prove the convergence of the above algorithm, we can rely on the Wolfe Conditions [31] that state the necessary and sufficient conditions for a line search algorithm to find a stationary point. As the proposed problem in (9) is convex, the unique stationary point is the global optimum. Therefore, the Wolfe conditions that ensure the convergence of the algorithm are

$$L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) < L_P(\mathbf{z}^k) + c_1 \nabla L_P(\mathbf{z}^k)^T \mathbf{p}^k \quad (18)$$

$$\nabla L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k)^T \mathbf{p}^k > c_2 \nabla L_P(\mathbf{z}^k)^T \mathbf{p}^k \quad (19)$$

for $0 \le c_1 < c_2 \le 1$, where, in our case, $(\mathbf{z}^k)^T = [((\mathbf{w}^1)^k)^T (b^1)^k \ldots ((\mathbf{w}^Q)^k)^T (b^Q)^k]$, and $\mathbf{p}^k$ is a vector formed by all the columns of $\mathbf{P}^k$ written one after another in a column vector. The first condition is also known as the strictly decreasing property and the second as the sufficient decreasing property. As their names show, they guarantee that in each step, we advance toward the solution and that the taken step is sufficiently large to reach the optimal solution with any required precision in a finite number of steps. We will now prove that the proposed procedure fulfills both conditions.

The first condition can be easily proved. We will set $c_1 = 0$, and we need to show that $L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) < L_P(\mathbf{z}^k)$. We will first show that $L''_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) < L''_P(\mathbf{z}^k)$, which can be readily seen because $\mathbf{z}^{k+1} = \mathbf{z}^k + \eta^k \mathbf{p}^k$ is constructed as a convex combination of $\mathbf{W}^k$ and $\mathbf{b}^k$ and the optimal solution for (12), $\mathbf{W}^s$, and $\mathbf{b}^s$. Therefore, being that the problem in (12) is convex, for any $\eta^k \in (0, 1]$, we will know that $L''_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) < L''_P(\mathbf{z}^k)$. By construction, $L''_P(\mathbf{z}^k) = L_P(\mathbf{z}^k)$, and we made the gradient of both equal; therefore, for sufficiently small η^k, the value of the function in $\mathbf{z}^k + \eta^k \mathbf{p}^k$ can be expressed by the first-order Taylor expansion around $\mathbf{z}^k$; consequently, $L''_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) \approx L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k)$, and the first condition will hold. There exists an $\eta^k > 0$ for which $L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k)$ is less than $L_P(\mathbf{z}^k)$ and the backtracking algorithm is devised for finding it.

Let us rewrite the second condition as follows so that the proof can be more clearly explained:

$$\sum_{j=1}^{Q} \nabla_{\mathbf{z}^j} L_P \left((\mathbf{z}^j)^k + \eta^k (\mathbf{p}^j)^k \right)^T (\mathbf{p}^j)^k > c_2 \sum_{j=1}^{Q} \nabla_{\mathbf{z}^j} L_P((\mathbf{z}^j)^k)^T (\mathbf{p}^j)^k \quad (20)$$

where $\mathbf{z}^j = \begin{bmatrix} \mathbf{w}^j \\ b^j \end{bmatrix}$, and $(\mathbf{p}^j)^k$ is the j th column of $\mathbf{P}^k$, $(\mathbf{p}^j)^k = \begin{bmatrix} (\mathbf{w}^j)^s - (\mathbf{w}^j)^k \\ (b^j)^s - (b^j)^k \end{bmatrix}$. After some algebraic manip-

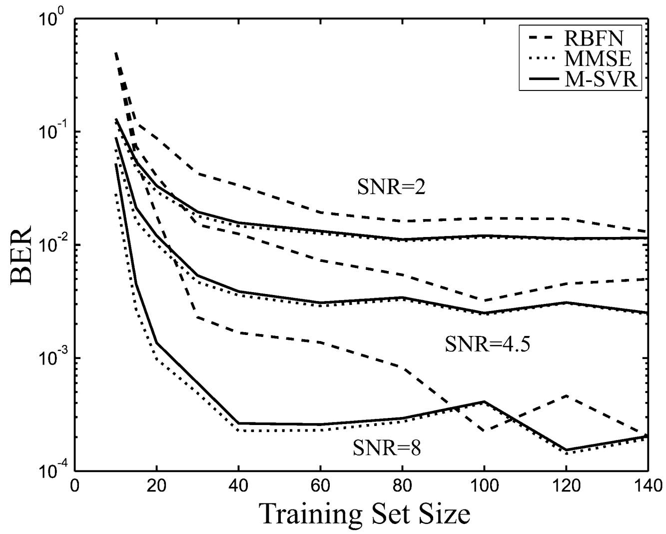


Fig. 5. M-SVR, MMSE, and RBFN BER curves in a linear MIMO channel for SNR ranging from 2 to 8 dB. MMSE solution is optimal in this case, and M-SVR, with a nonlinear kernel, approximates it with reasonable accuracy.

ulations, $(\mathbf{p}^j)^k = 1/\eta^k \begin{bmatrix} (\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k \\ (b^j)^{k+1} - (b^j)^k \end{bmatrix}$. We can now manipulate the left side of (20) as follows:

$$\sum_{j=1}^{Q} \nabla_{\mathbf{z}^j} L_P \left((\mathbf{z}^j)^k + \eta^k (\mathbf{p}^j)^k\right)^T (\mathbf{p}^j)^k$$
$$= \frac{1}{\eta^k} \sum_{j=1}^{Q} \left(\|(\mathbf{w}^j)^{k+1}\|^2 - ((\mathbf{w}^j)^k)^T (\mathbf{w}^j)^{k+1}\right)$$
$$- \frac{1}{\eta^k} \sum_{j=1}^{Q} \left(C \sum_{i=1}^{n} \left.\frac{dL(u)}{du}\right|_{u_i^{k+1}} \frac{e_{ij}^{k+1}}{u_i^{k+1}} \Big[\boldsymbol{\phi}(\mathbf{x}_i)((\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k) + ((b^j)^{k+1} - (b^j)^k) \Big] \right). \quad (21)$$

Now we add and subtract y_{ij} for all i and j to (21) and, for simplicity, we will drop the $1/\eta^k$, leading to (22), shown at the bottom of the page. We now add and subtract $C \sum_{i=1}^{k} L(u_i^{k+1})$ and $\frac{1}{2} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^k\|^2$ to (22), leading to

$$\sum_{j=1}^{Q} \left(\frac{1}{2} \|(\mathbf{w}^j)^{k+1}\|^2 - ((\mathbf{w}^j)^k)^T (\mathbf{w}^j)^{k+1} + \frac{1}{2} \|(\mathbf{w}^j)^k\|^2 \right)$$
$$+ \frac{1}{2} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^{k+1}\|^2 + C \sum_{i=1}^{k} L(u_i^{k+1})$$
$$- \frac{1}{2} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^k\|^2 - C \sum_{i=1}^{k} L(u_i^{k+1})$$
$$- C \sum_{i=1}^{n} \left.\frac{dL(u)}{du}\right|_{u_i^{k+1}} \frac{\mathbf{e}_i^{k+1}}{u_i^{k+1}} \left[\mathbf{e}_i^k - \mathbf{e}_i^{k+1}\right]$$
$$= \frac{1}{2} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k\|^2$$
$$+ L_P(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}) - L_P'''(\mathbf{W}^k, \mathbf{b}^k) \quad (23)$$

where we have defined

$$L_P'''(\mathbf{W}, \mathbf{b}) = \frac{1}{2} \sum_{j=1}^{Q} \|\mathbf{w}^j\|^2 + C \left(\sum_{i=1}^{n} L(u_i^{k+1}) + \left.\frac{dL(u)}{du}\right|_{u_i^{k+1}} \times \frac{(\mathbf{e}_i^{k+1})^T}{u_i^{k+1}} \left[\mathbf{e}_i - \mathbf{e}_i^{k+1}\right] \right) \quad (24)$$

as $L_P'(\mathbf{W}, \mathbf{b})$ in (11), but the first-order Taylor expansion is made over the actual solution instead of the previous one. $L_P'''(\mathbf{W}, \mathbf{b})$ is also a lower bound for $L_P(\mathbf{W}, \mathbf{b})$ for any $\mathbf{W}$ and $\mathbf{b}$. Therefore

$$\sum_{j=1}^{Q} \nabla_{\mathbf{z}^j} L_P \left((\mathbf{z}^j)^k + \eta^k (\mathbf{p}^j)^k\right)^T (\mathbf{p}^j)^k$$
$$= \frac{1}{\eta^k} \left(\frac{1}{2} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k\|^2 + L_P(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}) - L_P'''(\mathbf{W}^k, \mathbf{b}^k) \right) \quad (25)$$

where we have recovered the $1/\eta^k$ that we suppressed in (22).

$$\sum_{j=1}^{Q} \left(\|(\mathbf{w}^j)^{k+1}\|^2 - ((\mathbf{w}^j)^k)^T (\mathbf{w}^j)^{k+1}\right)$$
$$- \sum_{j=1}^{Q} \left(C \sum_{i=1}^{n} \left.\frac{dL(u)}{du}\right|_{u_i^{k+1}} \frac{e_{ij}^{k+1}}{u_i^{k+1}} \left[\left(y_{ij} - \boldsymbol{\phi}(\mathbf{x}_i)(\mathbf{w}^j)^k - (b^j)^k\right) - \left(y_{ij} - \boldsymbol{\phi}(\mathbf{x}_i)(\mathbf{w}^j)^{k+1} + (b^j)^{k+1}\right) \right] \right)$$
$$= \sum_{j=1}^{Q} \left(\|(\mathbf{w}^j)^{k+1}\|^2 - ((\mathbf{w}^j)^k)^T (\mathbf{w}^j)^{k+1}\right) - C \sum_{i=1}^{n} \left.\frac{dL(u)}{du}\right|_{u_i^{k+1}} \frac{\mathbf{e}_i^{k+1}}{u_i^{k+1}} \left[\mathbf{e}_i^k - \mathbf{e}_i^{k+1}\right]. \quad (22)$$

$$\frac{1}{\eta^k} \left(\frac{1}{2} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k\|^2 + L_P(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}) - L_P'''(\mathbf{W}^k, \mathbf{b}^k) \right)$$
$$- \frac{1}{\eta^k} \left(-\frac{1}{2} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k\|^2 - L_P(\mathbf{W}^k, \mathbf{b}^k) + L_P'(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}) \right)$$
$$= \frac{1}{\eta^k} \sum_{j=1}^{Q} \|(\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k\|^2 + \frac{1}{\eta^k} \left(L_P(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}) - L_P'(\mathbf{W}^{k+1}, \mathbf{b}^{k+1})\right)$$
$$+ \frac{1}{\eta^k} \left(L_P(\mathbf{W}^k, \mathbf{b}^k) - L_P'''(\mathbf{W}^k, \mathbf{b}^k)\right) > 0. \quad (27)$$

We can repeat the same procedure for the right-hand side of (20), leading to

$$c_2 \sum_{j=1}^{Q} \nabla_{\mathbf{z}^j} L_P((\mathbf{z}^j)^k)^T (\mathbf{p}^j)^k = \frac{c_2}{\eta^k}\left(-\frac{1}{2}\sum_{j=1}^{Q} \|(\mathbf{w}^j)^{k+1} - (\mathbf{w}^j)^k\|^2 - L_P(\mathbf{W}^k, \mathbf{b}^k) + L'_P(\mathbf{W}^{k+1}, \mathbf{b}^{k+1})\right). \quad (26)$$

Now, we set $c_2 = 1$, and we show that (25) minus (26) is greater than zero to proof the sufficient decreasing property see (27), shown at the bottom of the previous page. The second and third terms are greater or equal than zero by construction because it is a convex function minus its first-order Taylor expansion, and the first term is the norm of a vector; therefore, unless $\mathbf{W}^{k+1} = \mathbf{W}^k$, the condition will hold. If $\mathbf{W}^{k+1} = \mathbf{W}^k$, the algorithm has converged to the optimal solution. As η^k is always greater than zero, it does not play any role in the non-negativity proof.

REFERENCES

[1] B. M. Hochwald and T. L. Marzetta, "Space-time modulation for unknown fading," in *Proc. SPIE AeroSense Conf.*, Orlando, FL, Apr. 1999.

[2] C. Komninakis, C. Frauili, A. H. Sayed, and R. D. Wesel, "Multi-input multi-output fading channel tracking and equalization using kalman estimation," *IEEE Trans. Signal Processing*, vol. 50, pp. 1065–1075, May 2002.

[3] N. Al-Dhahir and A. H. Sayed, "The finite-length multi-input multi-output MMSE-DFE," *IEEE Trans. Signal Processing*, vol. 48, pp. 2921–2936, Oct. 2000.

[4] J. Baltersee, G. Fock, and H. Meyr, "Achievable rate of MIMO channels with data-aided channel estimation and perfect interleaving," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 2358–2368, Dec. 2001.

[5] Q. Sun, D. C. Cos, H. C. Huang, and A. Lozano, "Estimation of continuous flat fading MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 549–553, Oct. 2002.

[6] Y. Li and R. Liu, "Adaptive blind source separation and equalization for multiple-input/multiple-output systems," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2864–2876, Nov. 1998.

[7] B. Schölkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2001.

[8] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. ACM Workshop COLT*, D. Haussler, Ed., Pittsburgh, PA, 1992, pp. 144–152.

[10] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[11] V. N. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, pp. 169–184.

[12] F. Pérez-Cruz, P. L. Alarcón-Diana, A. Navia-Vázquez, and A. Artés-Rodríguez, "Fast training of support vector classifiers," in *Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2000, pp. 734–740.

[13] F. Pérez-Cruz, A. Navia-Vázquez, A. R. Figueiras-Vidal, and A. Artés-Rodríguez, "Empirical risk minimization for Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 14, pp. 296–303, Mar. 2003.

[14] F. Pérez-Cruz and A. Artés-Rodríguez, "Adaptive SVC for nonlinear channel equalization," in *Proc. EUSIPCO*, Toulouse, France, Sept. 2002.

[15] X. Zhou and X. Wang, "Channel estimation for OFDM systems using adaptive radial basis function networks," *IEEE Trans. Veh. Technol.*, vol. 52, pp. 48–59, Jan. 2003.

[16] M. Martone, "Wavelet-based separating kernels for array processing of cellular DS/CDMA signals in fast fading," *IEEE Trans. Commun.*, vol. 48, pp. 979–995, June 2000.

[17] A. Sayeed and B. Aazhang, "Joint multipath-doppler diversity in mobile wireless communications," *IEEE Trans. Veh. Technol.*, vol. 47, pp. 123–132, Jan. 1999.

[18] Y. Li and N. Sollenberger, "Adaptive antenna arrays for OFDM systems with cochannel interference," *IEEE Trans. Commun.*, vol. 47, pp. 217–229, Feb. 1999.

[19] S. Chen, S. Gunn, and C. J. Harris, "Decision feedback equalizer design using Support Vector Machines," in *Proc. Inst. Elect. Eng., Vision, Image Signal Process.*, vol. 147, 2000, pp. 213–219.

[20] S. Chen, A. Samingan, and L. Hanzo, "Support vector machine multiuser receiver for DS-CDMA signals in multipath channels," *IEEE Trans. Neural Networks*, vol. 12, pp. 604–611, May 2001.

[21] D. J. Sebald and J. A. Bucklew, "Support Vector Machines techniques for nonlinear equalization," *IEEE Trans. Signal Processing*, vol. 48, pp. 3217–3226, Nov. 2000.

[22] S. Chakrabartty and G. Cauwenberghs, "Sequence estimation and channel equalization using forward decoding kernel machines," in *Proc. ICASSP*, vol. 3, May 2002, pp. 2669–2672.

[23] S. Chen, G. J. Gibson, C. F. N. Cowan, and P. M. Grant, "Adaptative equalization of finite nonlinear channels using multilayer perceptrons," in *Signal Process.*, vol. 10, 1990, pp. 107–119.

[24] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, pp. 311–335, Mar. 1998.

[25] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Second ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[26] J. Park and I. W. Sandberg, "Universal approximation using radial basis function networks," *Neural Comput.*, vol. 3, pp. 246–257, 1991.

[27] K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. N. Vapnik, "Predicting time series with Support Vector Machines," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 243–254.

[28] A. J. Smola and B. Schölkopf. (1998) A tutorial on support vector regression, Tech. Rep. NC-TR-98-030. Royal Holloway College, Univ. London, London, U.K.. [Online]. Available: ftp://www.neurocolt.com/pub/ neurocolt/~tech_reports/1998/98 030.ps.Z

[29] F. Pérez-Cruz, G. Camps, E. Soria, J. Pérez, A. R. Figueiras-Vidal, and A. Artés-Rodríguez, "Multi-dimensional function approximation and regression estimation," in *Proc. ICANN*, Madrid, Spain, 2002.

[30] F. Pérez-Cruz, A. Navia-Vázquez, P. L. Alarcón-Diana, and A. Artés-Rodríguez, "An IRWLS procedure for SVR," in *Proc. EUSIPCO*, Tampere, Finland, Sept. 2000.

[31] J. Nocedal and S. J. Wright, *Numerical Optimization* New York, 1999.

[32] G. S. Kimeldorf and G. Wahba, "Some results in Tchebycheffian spline functions," *J. Math. Anal. Applicat.*, vol. 33, pp. 82–95, 1971.

[33] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. New York: Wiley, 1987.

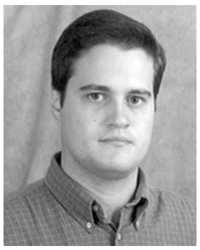
[34] S. A. Maas and D. Neilson, "Modeling MESFET's for intermodulation analysis of mixers and amplifiers," *IEEE Trans. Microwave Theory Techn.*, vol. 38, pp. 1964–1971, Dec. 1990.

[35] M. Lázaro, I. Santamaría, C. Pantaleón, J. Ibáñez, and L. Vielva, "A regularized technique for the simultaneous reconstruction of a function and its derivatives with application to nonlinear transistor modeling," *Signal Process.*, vol. 83, no. 9, pp. 1859–1870, 2003.

Matilde Sánchez-Fernández (M'03) received the Telecommunication Engineering and Ph.D. degrees from the Universidad Politécnica de Madrid, Madrid, Spain, in 1996 and 2001, respectively.

She has been an Assistant Professor at Universidad Carlos III de Madrid since April 2000. Previously, she worked for Telefónica, Madrid, as a Telecommunication Engineer. She performed research with the Information and Telecommunication Technology Center, Kansas University, Lawrence, and at Bell Labs, Holmdel, NJ. Her current research interests are dealing with MIMO techniques, turbo codes, mobile communications, simulation and modeling of communication systems, and adaptive coding and modulation.

Mario de-Prado-Cumplido (S'02) was born in Madrid, Spain, in December 1977. He received the M.Sc. degree in telecommunication engineering from the Universidad Politécnica de Madrid in 2000. He is currently pursuing the Ph.D. degree with the Universidad Carlos III de Madrid, where he is doing research on machine learning algorithms, feature selection, and probability density estimation techniques and their application to biomedical problems, with particular attention to cardiovascular pathologies.



Jerónimo Arenas-García (S'00) was born in Seville, Spain, in 1977. He received the Telecommunication Engineer degree in 2000 from Universidad Politécnica de Madrid, Madrid, Spain, where he was ranked number 1 in the National Award to graduation. He is currently pursuing the Ph.D. degree at the Department of Signal Theory and Communications, Universidad Carlos III de Madrid.

His present research interests are focused in the fields of adaptive signal processing and machine learning and their applications.

Fernando Pérez-Cruz (S'97–M'01) was born in Sevilla, Spain, in 1973. He received the Telecommunication Engineering degree in 1996 from the Universidad de Sevilla and the Ph.D. degree, also in Telecommunication Engineering, from the Universidad Politécnica de Madrid, Madrid, Spain, in 2000.

He is an Associated Professor with the Department of Signal Theory and Communication, Universidad Carlos III de Madrid. He is currently on sabbatical at the Gatsby Computational Neuroscience Unit, University College, London, U.K. His current interest lies in machine learning algorithmic and theoretical developments and its application to signal processing and financial data. He has authored over 40 contributions to international journals and conferences.