KERNELIZED SET-MEMBERSHIP APPROACH TO NONLINEAR ADAPTIVE FILTERING

Amaresh V. Malipatil[1], *Yih-Fang Huang*[1]

[1]Department of Electrical Engineering
University of Notre Dame
{amalipat,huang}@nd.edu

Srinivas Andra[2], *Kristin Bennett*[3]

[2]Department of ECSE
[3]Department of Mathematical Sciences
Rensselaer Polytechnic Institute
{andras,bennek}@rpi.edu

ABSTRACT

In linear filtering, the set-membership normalized least mean squares (SM-NLMS) algorithm has been shown to exhibit desirable features of selective update and optimized variable step size. In this paper, a kernel approach to the SM-NLMS algorithm is presented that makes it feasible to address nonlinear problems. An online greedy approximation technique to achieve sparsity is discussed. Simulation results are presented for two practical problems: equalization of nonlinear inter-symbol interference (ISI) channels and predistortion of nonlinear high power amplifiers (HPA).

1. INTRODUCTION

Nonlinear distortion, either memoryless or with memory, when encountered in a system, deteriorates the system performance considerably. Volterra systems have been used for the analysis and compensation of such nonlinear distortions [1]. Recently, support vector machines (SVM) have been gaining popularity due to their improved generalization performance and sparsity of the solution for classification and regression [2], [3]. However, both of these approaches have high computational requirements resulting from identification of Volterra kernels and solution to quadratic program (QP) respectively.

The set-membership (SM) approach for linear adaptive filtering has been well studied recently [4], [5]. A kernel-based approach to SM-NLMS [6] is presented here to address nonlinear problems. The attractive features of optimized variable step-size and selective update of the conventional SM algorithms carry over to the nonlinear case as well. Apart from having lower training complexity than the SVM, the kernel SM-NLMS algorithm has an added advantage of online adaptation of the filter coefficients while the classical SVM approach follows a batch-learning implementation. This paper is organized as follows. Section 2 describes the conventional set-membership filtering (SMF) approach. Section 3 discusses the SM-NLMS algorithm and presents its two kernel versions for nonlinear problems along with a complexity reduction method. In Section 4, the kernel SM-NLMS algorithm is applied to two specific nonlinear problems to assess its performace. Section 5 concludes the paper.

This work has been supported, in part, by the Naitonal Science Foundation under Grant EEC02-03366, by the U.S. Department of the Army under Contract DAAD 16-02-C-0057-P1, and by the Indiana 21st Century Fund for Research and Technology.

2. SET MEMBERSHIP FILTERING

In set-membership filtering, the objective is to find a set of feasible filter coefficients such that the resulting estimation errors are bounded in magnitude over a model space $\mathcal{S}$, that consists of all the input vector-desired output pairs $(\mathbf{x}, d)$, where $\mathbf{x} \in \mathbb{C}^N$ and d is a complex scalar. If $\theta \in \mathbb{C}^N$ represents the linear-in-parameter vector, then the filter output is given by $y(\theta) = \theta^T \mathbf{x}$ and the filter error is $e(\theta) = d - y(\theta)$. Hence the SMF criterion is to find θ such that

$$|e(\theta)|^2 \leq \gamma^2 \quad \forall (\mathbf{x}, d) \in \mathcal{S} \tag{1}$$

where γ is an upper bound on the filter error. The SMF criterion results in a region estimate called the *feasibility set* Θ, which is given by

$$\Theta \triangleq \bigcap_{(\mathbf{x},d)\in\mathcal{S}} \{\theta \in \mathbb{C}^N : |d - \theta^T \mathbf{x}|^2 \leq \gamma^2\}$$

Properly chosen error bounds result in a non-empty *feasibility set*, which is assumed to be so throughout this paper. The *constraint set* $\mathcal{H}_n$, for input vector-desired output pairs $(\mathbf{x}_n, d_n)$ at a time instant n, refers to the set of all parameter vectors that are consistent with (1), i.e.,

$$\mathcal{H}_n = \{\theta \in \mathbb{C}^N : |d_n - \theta^T \mathbf{x}_n| \leq \gamma\}$$

The minimal set estimate for Θ at time instant *n*, referred to as the *exact membership set*, is given by $\psi_n = \cap_{i=1}^{n} \mathcal{H}_i$. In the next section, the SM-NLMS algorithm is reviewed and two kernel versions of the same are presented along with a complexity reduction method.

3. KERNEL SET MEMBERSHIP ALGORITHMS

This section presents two kernel set membership algorithms for nonlinear regression and classification. A method to reduce the complexity of kernel SM algorithm solutions is also discussed.

3.1. Kernel SM-NLMS algorithm

The SM-NLMS algorithm is a supervised learning algorithm, belonging to the set-membership adaptive recursive techniques (SMART) family [6]. The SM-NLMS is formulated by the following equations [6]:

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \alpha_n \delta_n \mathbf{x}_n^* / \mathbf{x}_n^H \mathbf{x}_n \quad (2)$$
$$\sigma_n^2 = \sigma_{n-1}^2 - \alpha_n^2 |\delta_n|^2 / \mathbf{x}_n^H \mathbf{x}_n \quad (3)$$

where $\mathbf{x}_n^*$ and $\mathbf{x}_n^H$ denote, respectively, the complex conjugate and Hermitian of $\mathbf{x}_n$. The prediction error δ_n and the gain α_n are given by

$$\delta_n = d_n - \hat{\theta}_{n-1}^T \mathbf{x}_n = d_n - y(\mathbf{x}_n) \quad (4)$$
$$\alpha_n = \begin{cases} 1 - \gamma/|\delta_n|, & \text{if } |\delta_n| > \gamma \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Rewriting (2) in nonrecursive form,

$$\hat{\theta}_n = \sum_{i=0}^{n} \alpha_i \delta_i \mathbf{x}_i^* / \mathbf{x}_i^H \mathbf{x}_i \quad (6)$$

Thus, for an input test vector $\mathbf{x}$, the filter output is

$$y(\mathbf{x}) = \sum_{i=0}^{n} \alpha_i \delta_i \mathbf{x}_i^H \mathbf{x} / \mathbf{x}_i^H \mathbf{x}_i \quad (7)$$

The kernel SM-NLMS algorithm performs linear regression in the feature space $\mathcal{F}$ of higher dimension (possibly infinite). This is accomplished by a nonlinear transformation from $\mathbf{x} \in \mathbb{C}^N$ to $\phi(\mathbf{x}) \in \mathcal{F}$. The motive behind this transformation comes from Cover's theorem on the separability of patterns [7]. It is clear from (7) that only the inner product between the input vectors needs to be computed. Hence, the inner product can be replaced by the Mercer kernel function i.e., $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$; see [3] for more details and choices for the Mercer kernel. Thus, the "kernel" trick eliminates the need for explicit computation of vectors $\phi(\mathbf{x})$ in the feature space, thereby smartly avoiding the *curse of dimensionality* [2]. The filter output is now given by,

$$y(\mathbf{x}) = \sum_{i=0}^{n} \alpha_i \delta_i k(\mathbf{x}_i, \mathbf{x}) / k(\mathbf{x}_i, \mathbf{x}_i) = \sum_{i=0}^{n} c_i k(\mathbf{x}_i, \mathbf{x}) \quad (8)$$

where $c_i = \alpha_i \delta_i / k(\mathbf{x}_i, \mathbf{x}_i)$. Sparsity is introduced in the solution because $\alpha_n = 0$ when $|\delta| < \gamma$. Sparsity is desired since it reduces the test phase complexity. A kernel set-membership binary classification (SM-BC) algorithm suitable for binary classification of nonlinearly separable patterns is presented in the next subsection.

3.2. Kernel SM-BC algorithm

In binary classification, for a given input vector-observation pair ($\mathbf{x}$,d), it is desired that the classifier output ($y(\mathbf{x}) = \theta^T \mathbf{x}$) satisfies the following conditions: $y(\mathbf{x}) > 0$, if $d = +1$; and $y(\mathbf{x}) < 0$, if $d = -1$. These two conditions together yield,

$$dy(\mathbf{x}) > 0 \quad (9)$$

The aim is to obtain an estimate of the optimal classifier θ that satifies the above inequality for all $(\mathbf{x}, d) \in \mathcal{S}$. To achieve better generalization performance, a margin γ_b is used as a threshold in the above inequality. Thus,

$$dy(\mathbf{x}) > \gamma_b \quad \forall (\mathbf{x}, d) \in \mathcal{S} \quad (10)$$

As in SM-NLMS algorithm, the parameter estimate is updated selectively, i.e., only when the inequality (10) is not satisfied. Using the *point-wise approach* as in [6], the new estimate is obtained by taking a projection of the old estimate onto the hyperplane represented by $d_n y(\mathbf{x}_n) - \gamma_b = 0$. The recursive updates for the estimate of the classifier is given by

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \beta_n \mathbf{x}_n^* / \mathbf{x}_n^H \mathbf{x}_n \quad (11)$$

where the *deviation* β_n is defined as follows:

$$\beta_n = \begin{cases} d_n y_n - \gamma_b & \text{if } d_n y_n < \gamma_b \\ 0, & \text{otherwise.} \end{cases}$$

The classifier output in the nonrecursive form is easily computed using $y(\mathbf{x}) = \hat{\theta}^T \mathbf{x}$ and (11):

$$y(\mathbf{x}) = \sum_{i=0}^{n} \beta_i \mathbf{x}_i^H \mathbf{x} / \mathbf{x}_i^H \mathbf{x}_i \quad (12)$$

which is of the form similar to (7). To perform classification of nonlinearly separable patterns, the kernelized form of (12) can be used as follows:

$$y(\mathbf{x}) = \sum_{i=0}^{n} \beta_i k(\mathbf{x}_i, \mathbf{x}) / k(\mathbf{x}_i, \mathbf{x}_i) = \sum_{i=0}^{n} f_i k(\mathbf{x}_i, \mathbf{x}) \quad (13)$$

where $f_i = \beta_i / k(\mathbf{x}_i, \mathbf{x}_i)$. This algorithm also exhibits the desirable features of selective update and online adaptation of classifier estimates. The computational complexity of both the algorithms in the test phase depends on the number of updates in the training phase which are equivalent to the support vectors in SVM. A complexity reduction method using an online greedy approximation technique proposed in [8] is discussed in the next subsection.

3.3. Sparse online greedy approximation

Let $\{\tilde{\mathbf{x}}_j\}_{j=1}^m$ be the m vectors retained till time instant n and $\{\tilde{c}_j\}_{j=1}^m$ be their corresponding filter coefficients so that $y(\mathbf{x}_n) = \sum_{j=1}^m \tilde{c}_j k(\tilde{\mathbf{x}}_j, \mathbf{x}_n)$. If (5) yields $\alpha_n \neq 0$, the algorithm seeks to find coefficients $\{a_{n,j}\}_{j=1}^m$ satisfying *approximate linear dependence* condition

$$\Big\| \sum_{j=1}^{m} a_{n,j}\phi(\tilde{\mathbf{x}}_j) - \phi(\mathbf{x}_n) \Big\|^2 \le \nu \tag{14}$$

where $\nu \ge 0$. Minimizing the left hand side of (14) helps to simultaenously check whether the condition is satisfied and obtain the coefficient vector $\mathbf{a}_n = [a_{n,1} a_{n,2} \cdots a_{n,m}]^T$ that best satisfies it. For numerical stability, an l_2-norm regularization term $\mu\|a\|^2$ ($\mu \ge 0$) is added to the left hand side in (14), resulting in the following minimization problem,

$$\min_{\mathbf{a}_n \in \mathbb{R}^m} \Big\{ \mathbf{a}_n^T(\tilde{K}_n + \mu I_m)\mathbf{a}_n - 2\mathbf{a}_n^T \tilde{\mathbf{k}}_n + k_{nn} \Big\} \tag{15}$$

where $[\tilde{K}_n]_{i,j} = k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$, $(\tilde{\mathbf{k}}_n)_i = k(\tilde{\mathbf{x}}_i, \mathbf{x}_n)$, $k_{nn} = k(\mathbf{x}_n, \mathbf{x}_n)$, $i, j = 1, 2 \cdots m$ and I_m is an $m \times m$ identity matrix.
Solving (15) yields $\mathbf{a}_n = \hat{K}_n^{-1}\tilde{\mathbf{k}}_n$, where $\hat{K}_n = \tilde{K}_n + \mu I_m$, and the condition (14) becomes,

$$k_{nn} - (\tilde{\mathbf{k}}_n + \mu \mathbf{a}_n)^T \mathbf{a}_n \le \nu \tag{16}$$

If (16) is satisfied, then $\mathbf{x}_n$ is discarded, $\hat{K}_n$ is unaltered and the filter coefficients $\tilde{c}_i$ are adjusted as follows,

$$\begin{aligned} \hat{K}_{n+1} &= \hat{K}_n \\ \tilde{c}_i &= \tilde{c}_i + c_n a_{n,i} \quad i = 1, 2, \cdots m \end{aligned}$$

Otherwise, the algorithm performs the following updates.

$$\begin{aligned} \tilde{c}_{m+1} &= c_n \\ \tilde{\mathbf{x}}_{m+1} &= \mathbf{x}_n \\ \hat{K}_{n+1} &= [\hat{K}_n \quad \tilde{\mathbf{k}}_n; \tilde{\mathbf{k}}_n^T \quad k_{nn} + \mu] \\ m &= m + 1 \end{aligned}$$

Computing the coefficient vector $\mathbf{a}_{n+1}$ requires inversion of $\hat{K}_{n+1}$ which can be efficiently computed in a recursive manner using the Sherman-Woodbury formula [9],

$$\hat{K}_{n+1}^{-1} = [P \quad \mathbf{q}; \mathbf{q}^T \quad s]$$

where $P = \hat{K}_n^{-1} + s\hat{K}_n^{-1}\tilde{\mathbf{k}}_n\tilde{\mathbf{k}}_n^T\hat{K}_n^{-1}$, $\mathbf{q} = -s\hat{K}_n^{-1}\tilde{\mathbf{k}}_n$ and $s = 1/(k_{nn} + \mu - \tilde{\mathbf{k}}_n^T\hat{K}_n^{-1}\tilde{\mathbf{k}})$. The sparse greedy approximation can be applied to the kernel SM-BC algorithm as well, as explained above.

4. APPLICATIONS AND SIMULATION RESULTS

In this section, two specific applications of the algorithms presented in the previous section are considered.

4.1. Adaptive Equalization of Nonlinear ISI channels

In practice, nonlinear ISI channels could be encountered in digital satellite communications [10] or digital magnetic recording [11]. A nonlinear channel with memory is usually modeled as a linear channel followed by a memoryless nonlinearity. The linear channel considered here has an impulse response $[1, \; -0.5]$, followed by a memoryless nonlinearity $x + 0.2x^2 - 0.9x^3$, where x is the output of the linear channel. The transmitter sends binary signals

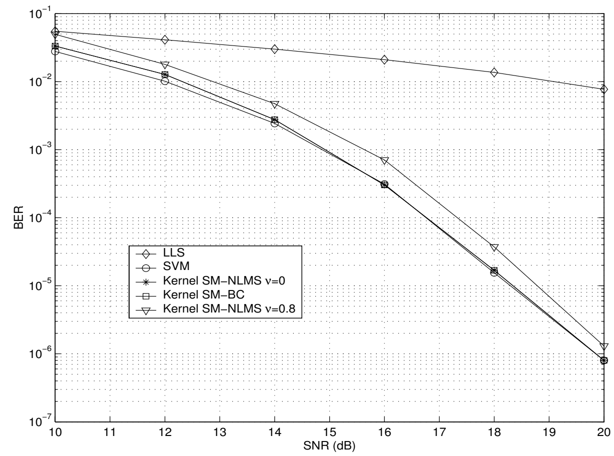

Fig. 1. BER performance of proposed algorithms

(± 1) with equal probability. The signal at receiver input is corrupted by additive white gaussian noise (AWGN). The receiver input samples are grouped to form the input vector $\mathbf{x}_n = [x(n) \; x(n-1) \; \cdots \; x(n-N+1)]^T$ and the decoded symbol is given by $\hat{d}_{n-D} = sign(y_n)$, where D is the decoding delay. Gaussian Radial Basis Function (RBF) kernel was used and the value of σ_{RBF} was tuned to achieve good generalization performace through cross validation. For SVM, the value of C, the regularization parameter was set using the method given in [12]. For the kernel SM-NLMS algorithm, the threshold was set to $3\gamma_v$, where γ_v^2 is the variance of the AWGN and $\gamma_b = 1$ for the kernel SM-BC algorithm. The size of training set is 200 symbols and test data consists of 10^6 symbols. The results are averaged over 100 independent trials with $N = 3$ and $D = 1$. The bit error rate (BER) curves show that the proposed algorithms perform significantly better than the conventional linear least squares (LLS) algorithm and comparably with the SVM. The online greedy approximation method explained in Section 3.3 is applied to the Kernel SM-NLMS algorithm with $\nu = 0$, $\nu = 0.8$ and $\mu = 0.01$. At SNR = 20 dB, the number of support vectors for SVM was found to be around 33%. The kernel SM-NLMS algorithm with $\nu = 0$ retained 49% of the vectors while the number of retained vectors with $\nu = 0.8$ was around 22%. The price paid for reduction in test phase complexity is degradation in performance as seen in Fig. 1. The curves for kernel SM-NLMS with $\nu = 0$ and kernel SM-BC are not distinguishable because they overlap.

4.2. Nonlinear High Power Amplifiers

Bandwith efficient modulation schemes like M-QAM have large amplitude fluctuations that makes them vulnerable to HPA nonlinearity, resulting in degraded BER performance and spectral regrowth that causes adjacent channel interference. A good solution is to use a predistorter (PD) that com-

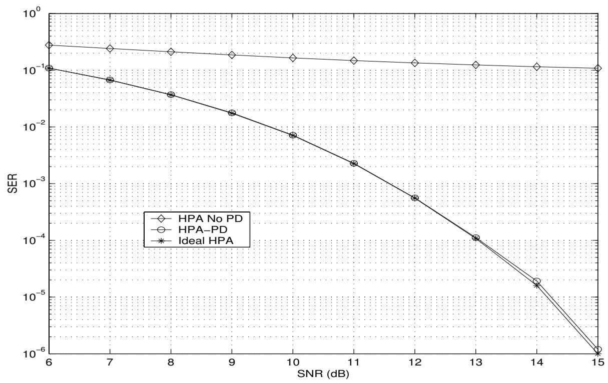

Fig. 2. Performance of HPA with and without PD

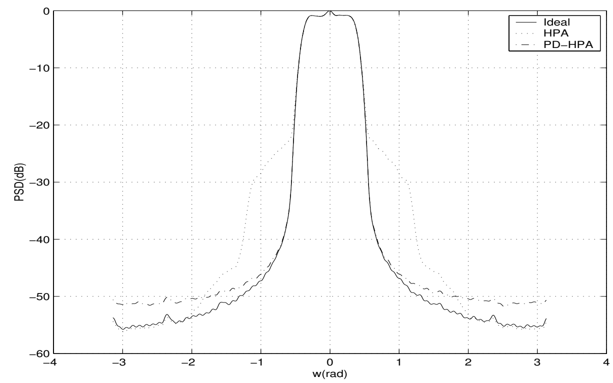

Fig. 3. Spectrum of the HPA output with and without PD

pensates for HPA nonlinearity so that the amplifier can be operated close to saturation with high power efficiency. The AM/AM $(A(r))$ and AM/PM $(\Phi(r))$ characteristics of the HPA memoryless nonlinearity are modeled using the Saleh model as in [13]. $A(r)$ and r represent, respectively, the normalized output amplitude and normalized input amplitude of the HPA; and $\Phi(r)$ represents the output phase. The digital predistorter should have AM/AM $(G(\cdot))$ and AM/PM $(\Psi(\cdot))$ characteristics that satisfy the following equations.

$$\begin{aligned} A(G(r)) &= k \cdot r \\ \Psi(r) + \Phi(G(r)) &= 0 \end{aligned}$$

where k is the desired gain of the linearized HPA. The pairs $(A(r)/k, r)$ and $(r, \Phi(r))$ are used as the input-desired output pairs to train the amplitude and phase PDs, respectively. Gaussian RBF kernel with σ_{RBF} tuned through cross validation is used. In the simulations, a 16-QAM modulation scheme is used. The baseband symbols are upsampled by a factor of 8 before being filtered by a square root raised-cosine (SRRC) filter with a roll-off factor 0.25 and truncated to 10 symbol duration. The sample size of training set is 100 symbols (800 samples) and that of test set is 10^6 symbols. The results are averaged over 100 independent trials with a Peak Back Off (PBO) [14] of 0.22dB. The receiver uses a SRRC filter to match the transmitter pulse shape. It is clear from Fig. 2 that, in presence of AWGN, the symbol error rate (SER) performance of the PD-HPA combination is very close that of the ideal case of a linear HPA. Also, the spectral regrowth is considerably reduced as seen in Fig. 3 thereby bringing down the adjacent channel interference.

5. CONCLUSION

A kernelized SM-NLMS algorithm was presented to solve practical nonlinear problems by implicit transformation of input vectors onto a feature space. The kernel SM-BC algorithm performs binary classification of nonlinearly separable patterns. Both algorithms have attractive features of selective update and adaptive step size. The test phase complexity can be reduced by retaining only the support vectors that are *exactly* or *approximately linearly independent* in the feature space.

6. REFERENCES

[1] M. Schetzen, *Volterra and Weiner Theories of Nonlinear Systems,* John Wiley & Sons, 1980.

[2] V. Vapnik, *Statistical Learning Theory,* Wiely, NY, 1998.

[3] B. Schölkopf and A. J. Smola, *Learning with Kernels,* MIT Press, 2001.

[4] S. Gollamudi, S. Kapoor, S. Nagaraj and Y. F. Huang, "Set-membership adaptive equalization and an updator-shared implementation for multiple channel communications systems," *IEEE Trans. Sig. Proc.*, 46(9), pp. 2372-2385, 1998.

[5] P.S.R. Diniz, S. Werner, "Set-membership binormalized data-reusing LMS algorithms," *IEEE Trans. Sig. Proc.,*, vol. 51, no. 1, pp. 124-134, 2003.

[6] S. Gollamudi, S. Nagaraj, S. Kapoor, and Y. F. Huang, "Set-membership filtering and a set-membership normalized LMS algorithm with an adaptive step size," *IEEE Signal Processing Letters*, vol. 5, no. 5, pp. 111-114, 1198.

[7] T. Cover, " Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Elect. Comp.*, 14(3), pp. 326-334, 1965.

[8] Y. Engel, S. Mannor and R. Meir, "Sparse Online Greedy Support Vector Regression," in *Proc. of the 13th European Conference on Machine Learning*, 2002.

[9] C. C. Cowen, *Linear Algebra for Engineering and Science,* West Pickle Press, 1996.

[10] G. Kechriotis, E. Zervas, and E.S. Manolakos, "Using recurrent neural networks for adaptive communication channel equalization," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 267-278, 1994.

[11] N. P. Sands and J.M. Cioffi, "Nonlinear channel models for digital magnetic recording," *IEEE Trans. Magn.*, vol. 29, no. 6, pp. 3996-3998, 1993.

[12] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, no. 1, pp. 113-126, 2004.

[13] D. S. Han and T. Hwang, "An Adaptive Pre-Distorter for the Compensation of HPA Nonlinearity," IEEE Trans. on Broadcasting, vol. 46, no. 2, pp. 152-157, 2000.

[14] J. K. Cavers, "Amplifier linearization using a digital predistorter with fast adaptation and low memory requirements," *IEEE Trans. Veh. Technol.*, vol. 39, no. 4, pp. 374-382, 1990.