A kernel-based RLS algorithm for nonlinear adaptive filtering using sparse approximation theory

Cédric Richard

Institut Charles Delaunay (ICD, FRE CNRS 2848), Laboratoire LM2S
Université de Technologie de Troyes, BP 2060, 10010 Troyes cedex - France
tel.: +33.3.25.71.58.47 fax.: +33.3.25.71.56.99 cedric.richard@utt.fr

1 Short abstract

In the last ten years, there has been an explosion of activity in the field of learning algorithms utilizing reproducing kernels, most notably for classification and regression. A common characteristic in kernel-based methods is that they deal with models whose order equals the number of input data, making them unsuitable for online applications. In this paper, we investigate a new kernel-based RLS algorithm that makes unnecessary the use of any computationally demanding sparsification procedure. The increase in the model order is controlled by the coherence parameter, a fundamental quantity that is used to characterize the behavior of dictionaries in sparse approximation problems.

2 Extended abstract

Adaptive filtering has become a topic of keen interest over the past three decades to help cope with time variations of system parameters and lack of *a priori* statistical information [11, 15]. Linear models are still routinely used because of their inherent simplicity from conceptual and implementational point of view. In many practical situations, however, nonlinear signal processing is needed. It includes items such as nonlinear system identification, prediction and control, e.g., in communications and biomedical engineering, see [9]. Following the pioneering works [1, 2, 13], there has been recent progress in function approximation methods based on reproducing kernel Hilbert spaces (RKHS) [12, 16], including, for example, support vector regression [18]. A common characteristic in kernel-based methods is that they deal with models whose order is the size of the training set, making them unsuitable for online applications. Several algorithms have been proposed to circumvent this computational burden in time series prediction problems [3, 4, 8]. Nevertheless they require excessively elaborate and costly operations such as matrix inversion.

The aim of this paper is to develop a new kernel-based RLS algorithm that makes unnecessary the use of any computationally demanding sparsification procedure. The increase in the model order is controlled by the coherence parameter, a fundamental quantity that characterizes the behavior of dictionaries in sparse approximation problems [19]. This abstract is organized as follows. In the first part, we introduce some basic principles of kernel-based optimum filtering in RKHS. Next we present our nonlinear adaptive filtering methods based on the coherence parameter. Finally simulations illustrate how efficient the proposed algorithm is.

Let $\mathcal{H}$ denote a RKHS of real-valued functions ψ on a compact $\mathcal{U} \subset \mathbb{R}^p$, and let $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$ be the inner product in $\mathcal{H}$. Let $\kappa : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ be the reproducing kernel of $\mathcal{H}$, meaning that $\psi(\boldsymbol{u}_i) = \langle \psi(\cdot), \kappa(\cdot, \boldsymbol{u}_i) \rangle_{\mathcal{H}}$ for all $\psi(\cdot) \in \mathcal{H}$ and every fixed $\boldsymbol{u}_i \in \mathcal{U}$. The problem is to determine a function $\psi(\cdot)$ of $\mathcal{H}$ that minimizes the sum of squared errors between n samples d_i of the desired response and the corresponding model output samples $\psi(\boldsymbol{u}_i)$, namely,

$$\min_{\psi \in \mathcal{H}} \sum_{i=1}^{n} |d_i - \psi(\boldsymbol{u}_i)|^2. \tag{1}$$

By virtue of the representer theorem [17], the solution to this problem can be expressed as a kernel expansion in terms of available data, that is,

$$\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \, \kappa(\cdot, \boldsymbol{u}_j). \tag{2}$$

Substituting (2) into equation (1) and using the reproducing property $\kappa(\boldsymbol{u}_i, \boldsymbol{u}_j) = \langle \kappa(\cdot, \boldsymbol{u}_i), \kappa(\cdot, \boldsymbol{u}_j) \rangle_{\mathcal{H}}$, elementary algebra shows that problem (1) reduces to

$$\min_{\boldsymbol{\alpha}} \; \|\boldsymbol{d} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 \tag{3}$$

where $\boldsymbol{K}$ is the n-by-n Gram matrix whose (i,j)-th entry is $\kappa(\boldsymbol{u}_i, \boldsymbol{u}_j)$. With the matrix $\boldsymbol{P} = (\boldsymbol{K}^t \boldsymbol{K})^{-1}$ assumed to exist, the solution to this problem is $\boldsymbol{\alpha} = \boldsymbol{P}\boldsymbol{K}^t\boldsymbol{d}$. Adaptive filtering raises the question of how to process an increasing amount of observations and update $\psi(\cdot)$ as new data is collected. Clearly the optimal approach outlined above cannot be used because it involves a n-th order model, see (2), and the n-by-n matrix $\boldsymbol{P}$ to be inverted, at each time instant n. Several attempts have been made recently to circumvent this computational burden [3, 4, 8]. Consider the m-th order model at any given time instant n

$$\psi_n(\cdot) = \sum_{j=1}^{m} \alpha_j \, \kappa(\cdot, \boldsymbol{u}_{\omega_j}), \tag{4}$$

where the ω_j's form an m-element subset $\mathcal{J}_n$ of $\{1, \ldots, n\}$. We call $\{\kappa(\cdot, \boldsymbol{u}_{\omega_j})\}_{j=1}^{m}$ the dictionary. These approaches rely on a two-stage process at each iteration: a model order selection step, and a parameter update step. In the model order selection step, at time instant n, the kernel function $\kappa(\cdot, \boldsymbol{u}_n)$ is inserted into the dictionary if it cannot be reasonably well represented by a combination of the other kernel functions of the dictionary. This condition usually takes the form

$$\min_{\boldsymbol{\gamma}} \|\kappa(\cdot, \boldsymbol{u}_n) \; - \sum_{\omega_j \in \mathcal{J}_{n-1}} \gamma_j \, \kappa(\cdot, \boldsymbol{u}_{\omega_j})\|_{\mathcal{H}}^2 > \epsilon_0, \tag{5}$$

where κ is a unit-norm kernel, that is, $\kappa(\boldsymbol{u}_k, \boldsymbol{u}_k) = 1$; otherwise replace $\kappa(\cdot, \boldsymbol{u}_k)$ by $\kappa(\cdot, \boldsymbol{u}_k)/\sqrt{\kappa(\boldsymbol{u}_k, \boldsymbol{u}_k)}$ in the above expression. The threshold ϵ_0 determines the level of sparsity of the model. These approaches, while accurate, are computationally prohibitive.

The coherence is a fundamental parameter to characterize a dictionary in sparse approximation problems [6, 19]. In our kernel-based context, we define the coherence parameter as

$$\mu = \max_{i \neq j} |\langle \kappa(\cdot, \boldsymbol{u}_{\omega_i}), \kappa(\cdot, \boldsymbol{u}_{\omega_j}) \rangle_{\mathcal{H}}| = \max_{i \neq j} |\kappa(\boldsymbol{u}_{\omega_i}, \boldsymbol{u}_{\omega_j})| \tag{6}$$

where κ is a unit-norm kernel. It reflects the most extreme correlations in the dictionary and, consequently, it is equal to zero for every orthonormal basis. Rather than solving a problem of the form (5), we suggest to insert $\kappa(\cdot, \boldsymbol{u}_n)$ into the dictionary provided that its coherence remains below a given threshold μ_0, namely,

$$\max_{\omega_j \in \mathcal{J}_{n-1}} |\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_j})| \leq \mu_0, \tag{7}$$

where μ_0 is a parameter in $[0, 1[$ determining both the level of sparsity and the coherence of the dictionary. The motivation for using this test is two-fold. First, it is easy to calculate and its computational complexity is only linear in the dictionary size. Second, it offers several attractive properties that can be exploited to assess novelty of input kernel functions, in particular:

- If $\mathcal{U}$ is compact, the dictionary of kernel functions determined under the rule (7) is finite. This implies that the order of the asymptotic model $\psi_\infty(\cdot)$ is finite and depends on μ_0.
- If $(m-1) < 1/\mu_0$, the elements of the dictionary $\{\kappa(\cdot, \boldsymbol{u}_{\omega_j})\}_{j=1}^m$ are linearly independent.
- If $(m-1) < 1/2\mu_0$, the left side of equation (5) is lower bounded by $1 - \frac{(m-1)\mu_0^2}{(1-(m-1)\mu_0)}$, which establishes a connection between ϵ_0 and μ_0.

For lack of space, proofs are omitted here but will be presented in the full-length paper. We describe now the parameter update step, whose purpose is to solve problem (3) recursively. Given the least-squares estimate $\boldsymbol{\alpha}_n = \boldsymbol{P}_n \boldsymbol{K}_n^t \boldsymbol{d}_n$, one of the following two alternatives holds with the arrival of $\boldsymbol{u}_{n+1}$. Either $\kappa(\cdot, \boldsymbol{u}_{n+1})$ does not satisfy the rule (7). It is not inserted into the dictionary and an iteration of the RLS algorithm is performed to get $\boldsymbol{\alpha}_{n+1}$ and $\boldsymbol{P}_{n+1}$. The computational cost is $\mathcal{O}(m^2)$. Or $\kappa(\cdot, \boldsymbol{u}_{n+1})$ is added to the dictionary. The order of the model is increased by one, and $\boldsymbol{\alpha}_{n+1}$ and $\boldsymbol{P}_{n+1}$ are updated accordingly using order-update relations that will be presented in the full-length paper. Each time it is run, the order-update iteration also requires $\mathcal{O}(m^2)$ operations.

The purpose of this part is to illustrate the performance of the proposed approach. Consider the discrete-time nonlinear dynamical system

$$d(n) = 0.5\, d(n-1)^3 + 0.3\, d(n-1)\, u(n-1) + 0.2\, u(n-1) + 0.05\, d(n-1)^2 + 0.6\, u(n-1)^3, \qquad (8)$$

where d_n is the desired output, and u_n is the input sampled from a zero-mean Gaussian distribution with variance 0.1. The data were generated from $d_0 = 0.1$. The system output d_n was corrupted by an additive zero-mean white Gaussian noise with variance 0.05, resulting in a signal-to-noise ratio of 1.15 dB. Our approach was used to identify a nonlinear model of the form $d_n = \psi(d_{n-1}, u_{n-1})$. Preliminary experiments were conducted to select the kernel and determine the best settings for the algorithm. The Gaussian kernel defined as $\kappa(\boldsymbol{u}_i, \boldsymbol{u}_j) = \exp\big(-\|\boldsymbol{u}_i - \boldsymbol{u}_j\|^2/2\,\beta_0^2\big)$ with $\beta_0 = 0.25$ was shown to be very accurate, and the coherence threshold μ_0 was set to $5 \cdot 10^{-2}$. The characteristics of the model were examined over one hundred 10000-sample test sequences. The final order m of the kernel expansion was 9.5 on average. Condition $(m-1) < 1/\mu_0$ mentioned above is then satisfied, in average, which indicates that the kernel functions of the dictionary were most frequently, if not always, chosen linearly independent. The normalized prediction mean-square prediction error was $3.32 \cdot 10^{-3}$. Some illustrative examples are shown in Figure 1. The convergence behavior of the method is presented in Figure 2. In the full-length paper, experiments with different values of the kernel bandwidth β_0 and colored noises will be also proposed.

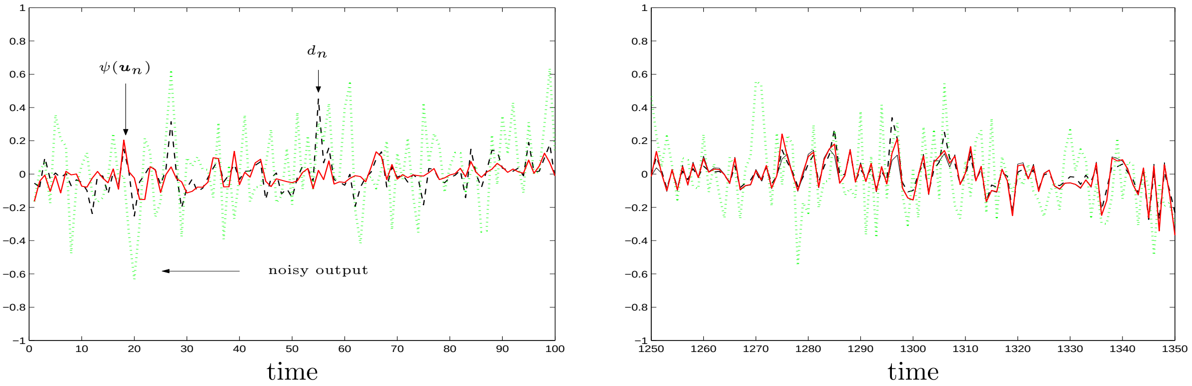

Figure 1: Desired output d_n, predicted output $\psi(\boldsymbol{u}_n)$ and system output with measurement noise during first instants (left) and later (right).

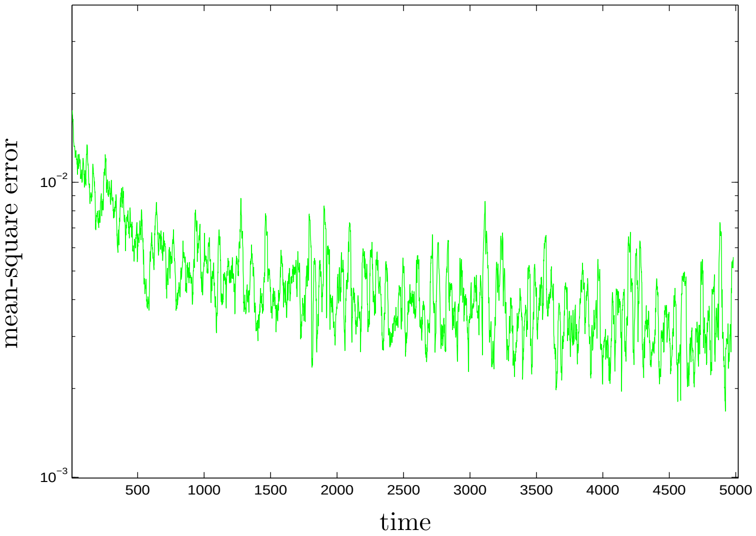

Figure 2: Convergence behavior of our approach.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] M. A. Azerman, E. M. Braverman, and L. I. Rozonoer. The method of potential functions for the problem of restoring the characteristic of a function converter from randomly observed points. *Automation and Remote Control*, 25(12):1546–1556, 1964.

[3] T. J. Dodd, V. Kadirkamanathan, and R. F. Harrison. Function estimation in Hilbert space using sequential projections. In *Proc. IFAC Conference on Intelligent Control Systems and Signal Processing*, pages 113–118, 2003.

[4] T. J. Dodd, B. Mitchinson, and R. F. Harrison. Sparse stochastic gradient descent learning in kernel models. In *Proc. Second International Conference on Computational Intelligence, Robotics and Autonomous Systems*, 2003.

[5] D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. In *Proc. National Academy of Sciences of the USA*, volume 100, pages 2197–2202, 2003.

[6] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[7] M. Elad and A.M. Bruckstein. A generalized uncertainty principle and sparse representations in pairs of bases. *IEEE Transactions on Information Theory*, 48(9):2558–2567, 2002.

[8] Y. Engel, S. Mannor, and R. Meir. Kernel recursive least squares. *IEEE Transactions on Signal Processing*, 52(8):2275–2285, 2004.

[9] G. B. Giannakis and E. Serpedin. A bibliography on nonlinear system identification. *Signal Processing*, 81:553–580, 2001.

[10] A.C. Gilbert, S. Muthukrishnan, and M.J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proc. 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2003.

[11] S. Haykin. *Adaptive filter theory.* Prentice Hall, Upper Saddle River, NJ, fourth edition, 2002.

[12] R. Herbrich. *Learning kernel classifiers. Theory and algorithms.* The MIT Press, Cambridge, MA, 2002.

[13] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

[14] S. Mallat and Z. Zhang. Matching pursuit with time-frequency disctionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[15] A. H. Sayed. *Fundamentals of adaptive filtering.* John Wiley & Sons, Hoboken, NJ, 2003.

[16] B. Schölkopf, J. C. Burges, and A. J. Smola. *Advances in kernel methods.* MIT Press, Cambridge, MA, 1999.

[17] B. Schölkopf, R. Herbrich, and R. Williamson. A generalized representer theorem. Technical Report NC2-TR-2000-81, NeuroCOLT, Royal Holloway College, University of London, UK, 2000.

[18] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. Technical Report NC-TR-98-030, NeuroCOLT, Royal Holloway College, University of London, UK, 1998.

[19] J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.