A DYNAMIC REGULARIZED GAUSSIAN RADIAL BASIS FUNCTION NETWORK FOR NONLINEAR, NONSTATIONARY TIME SERIES PREDICTION

Paul Yee *Simon Haykin*

Communications Research Laboratory
McMaster University
Hamilton, ON L8S 4K1

ABSTRACT

A dynamic network of regularized Gaussian radial basis functions (GaRBF) is described for the one-step prediction of nonlinear, nonstationary autoregressive (NLAR) processes governed by a smooth process map and a zero-mean, independent additive disturbance process of bounded variance. For N basis functions, both full-order and reduced-order updating algorithms are introduced, having computational complexities of $\mathcal{O}(N^3)$ and $\mathcal{O}(N^2)$, respectively, per time step. Simulations on a 10,000 point, 8-bit quantized 64kbps rate speech signal show that the proposed dynamic algorithm has a prediction performance comparable and, in some cases, superior to that of AT&T's LMS-based speech predictor designed for the ITU-T G.721 standard on the 32kbps ADPCM of speech. The results indicate that the proposed dynamic regularized GaRBF predictor provides a useful tradeoff between its minimal need for prior knowledge of the speech data characteristics and its consequently heavier computational burden.

1. INTRODUCTION

Recently there has been significant interest in applying nonparametric regression techniques, such as kernel regression, to the problem of nonlinear time series prediction. In the usual context of minimum mean-square error (m.m.s.e.) prediction, such "nonlinear" time series are those precisely those that arise from non-Gaussian processes. A study of the application of kernel regression to the general nonlinear AR (NLAR) case can be found in [1]. Under the more restrictive assumption of a nonlinear MA process that can be described by a linear basis function expansion with additive noise, [2] obtains correspondingly stronger results for their algorithm's convergence and consistency. In the other direction under a similar model, [3] considers the problem of nonparametric estimation for general Hilbert-space-valued stochastic processes via empirical estimates of the covariance and cross-covariance operators. Despite the broad spectrum of these nonparametric methods, they all share the underlying assumption that the process is (almost surely) *stationary* in the strong sense, an assumption that is often violated in real-world applications. The extension of these methods to nonstationary processes would, no doubt, be a significant advance.

In this paper, we examine a particular class of nonlinear AR processes and show that the m.m.s.e. one-step prediction problem for this class can be naturally formulated as an ill-posed interpolation problem in the process state space. Tikhonov regularization is then applied to yield a Gaussian radial basis function (GaRBF) expansion that approximates the desired regression function. Using results from penalized least-squares (PLS) interpolation theory [4], we discuss briefly how the regularization principle could be justified in the context of *stationary* time series prediction. This discussion then motivates the introduction of the *dynamic regularized GaRBF* predictor for the *nonstationary* version of the NLAR process class. The proposed method is compared with the AT&T IIR filter predictor, which uses a nonlinear LMS training algorithm, specified for the ITU-T G.721 standard 32kbps ADPCM of speech data. The simulation results show that the proposed dynamic regularized GaRBF predictor can model the underlying nonstationarity and nonlinearity of the speech signal sufficiently well to match and even exceed the performance of the standard AT&T predictor. While the flexible, nonparametric nature of dynamic regularized GaRBF predictor naturally demands more data and computational resources compared with the highly optimized AT&T predictor, these results indicate that the proposed approach is nevertheless effective in problems characterized by weak prior knowledge of the underlying process structure.

2. REGULARIZED PREDICTION OF STATIONARY NLAR PROCESSES

Before we consider the nonstationary case, it is instructive to consider the one-step prediction of the NLAR(p) process $\{X_i\}_{i=1}^{\infty}$ described by

$$X_i = f(\mathbf{X}_{i-1}) + \epsilon_i, \qquad i = 1, 2, \ldots \tag{1}$$

where $\mathbf{x}_i \triangleq [x_i \; x_{i-1} \; \ldots \; x_{i-p+1}]$ can be considered the *process state vector* at time instant i and the ϵ_i are disturbance inputs with $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{E}[\epsilon_i^2] < \infty$ for every i. Note that, by construction, the process state vector in (1) forms a Markov chain in the state space $\mathcal{R}^p$. For now, we do not place any other restrictions on the nature of the disturbance

This work was supported in part by the Ontario Provincial Government in the form of an Ontario Graduate Scholarship

0-7803-2431-5/95 $4.00 © 1995 IEEE

input process other than that it be independent of $\{X_i\}$ and assume that the initial state $\mathbf{x}_0 = [x_0\, x_{-1} \ldots x_{1-p}]$ is given. Given sample values of the process $\{x_i\}_{i=1}^{N+1}$, we may recast the prediction problem into one of interpolation in the process state space by considering the set of ordered pairs $\{(\mathbf{x}_i, x_{i+1})\}_{i=1}^{N}$, $i = 1, 2, \ldots, N$, that describes the desired input-output relation of $f : \mathcal{R}^p \to \mathcal{R}$. Assuming that f is a sufficiently smooth function, i.e., continuously differentiable to some order as well as Lipschitz, a regularized estimate of f is given by the GaRBF expansion

$$\widetilde{f}_N(\mathbf{x}) \triangleq \sum_{i=1}^{N} c_i G(\mathbf{x} - \mathbf{x}_i) = \mathbf{c}^{\mathsf{T}} \mathbf{G}(\mathbf{x}) \qquad (2)$$

where

$$G(\mathbf{x}) = \exp\left(-\|\mathbf{x}\|_{\mathbf{W}}^2/2\right), \qquad \|\mathbf{x}\|_{\mathbf{W}} \triangleq \|\mathbf{W}\mathbf{x}\|$$

and the coefficients in $\mathbf{c}$ are computed from the one-step interpolation condition

$$\begin{aligned} (\mathbf{G} + \lambda \mathbf{I})\,\mathbf{c} &= \mathbf{y} \\ \mathbf{G} &\triangleq [G(\mathbf{x}_i - \mathbf{x}_j)]_{i,j=1}^{N} \\ \mathbf{y} &\triangleq [x_{i+1}]_{i=1}^{N} \end{aligned} \qquad (3)$$

The particular choice of the set of Gaussian basis functions $\mathcal{G} \triangleq \{G(\bullet - \mathbf{x}_i)\}_{i=1}^{N}$ can be explained with reference to *a priori* assumptions on smoothness properties of f. What is relevant is that if the $\mathbf{x}_i$ are distinct and the norm-weighting matrix $\mathbf{W}$ is non-singular, then so is the interpolation matrix $\mathbf{G}$, thus ensuring that $\mathbf{c}$ in (3) is well-defined. For further details on these and other technical aspects of RBF interpolation theory, the interested reader is referred to [5, 6].

The estimate $\widetilde{f}_N$ constructed according to (2) and (3) is optimal in the sense that it minimizes the regularized cost functional

$$H_c\mathbf{c} = \sum_{i=1}^{N} \left| x_{i+1} - \mathbf{c}^{\mathsf{T}} \mathbf{G}(\mathbf{x}_i) \right|^2 + \lambda \mathbf{c}^{\mathsf{T}} \mathbf{G} \mathbf{c} \triangleq H_0\mathbf{c} + \lambda H_s\mathbf{c}$$

over $\mathbf{c} \in \mathcal{R}^N$ given the set of basis functions $\mathcal{G}$. It is clear that the *regularization parameter* $\lambda \in \mathcal{R}^+$ balances the usual sum-of-squared-prediction error over the sample data against a G-weighted norm on $\mathbf{c}$. Under suitable conditions, this G-weighted norm is equivalent to a smoothness measure over *span*$(\mathcal{G})$ [4]. Hence λ controls the degree of smoothing introduced into the solution by the regularization procedure; $\lambda = 0$ corresponds to a standard least-squares (LS) approximation with no smoothing of the sample data while $\lambda \to \infty$ corresponds to an oversmoothed approximation.

At this point, one may question the utility of the additional regularizing term $\lambda H_s\mathbf{c}$ in $H_c\mathbf{c}$. It turns out, however, that if we define the empirical total squared sample fitting error as

$$T(\lambda) \triangleq \sum_{i=1}^{N} \left| f(\mathbf{x}_i) - \widetilde{f}_N(\mathbf{x}_i) \right|^2 \qquad (4)$$

and seek to $\min\left\{\mathrm{E}[T(\lambda)] : \lambda \in \mathcal{R}^+\right\}$, where the expectation is taken over all possible realizations of the process sample $\{X_i\}_{i=1}^{N+1}$, under the condition that the ϵ_i are i.i.d. then in all but degenerate cases the optimum values for λ are nonzero [4]. In this sense, regularized fitting is canonically better than standard least-squares fitting where $\lambda = 0$. This result formally verifies the intuition that given that the model (1) specifies observations corrupted by additive noise, the best policy is not to fit the given time series sample data $\{(\mathbf{x}_i, x_{i+1})\}_{i=1}^{N}$ exactly if we wish to recover a good estimate for f.

So far the discussion has centred on the properties of the regularization parameter λ as applied to data fitting. The natural question then arises as to the relevance of minimizing $\mathrm{E}[T(\lambda)]$ when one-step prediction is set as an interpolation problem. The problem is that minimizing $\mathrm{E}[T(\lambda)]$ over the process sample $\{X_i\}_{i=1}^{N+1}$ does not necessarily minimize the m.m.s.e. prediction cost function

$$\mathrm{E}\left[\left| X_{N+2} - \widetilde{f}_N(\mathbf{X}_{N+1}) \right|^2\right]$$

for the as yet unobserved process datum x_{N+2}, where the expectation is now taken with respect to $\{X_i\}_{i=1}^{N+2}$. One immediate answer to this question is intimately tied to the existence of a stationary distribution for the process at hand. Note that this condition is stronger than the usual notion of (strict) stationarity. Recall that when the disturbance process $\{\epsilon_i\}$ is i.i.d., as is often assumed, and the mapping f is both Lipschitz and exponentially asymptotically stable in the large about 0, the Markov chain described in (1) is *geometrically ergodic* and hence has a stationary distribution [7]. For this stationary case, if we consider the probability of a large deviation in the prediction error for the next process datum x_{N+2}, the probabilistic triangle inequality leads to

$$\begin{aligned} &\Pr\left\{\left| X_{N+2} - \widetilde{X}_{N+2} \right| > \delta\right\} \\ &\quad = \Pr\left\{\left| f(\mathbf{X}_{N+1}) + \epsilon_{N+2} - \widetilde{f}_N(\mathbf{X}_{N+1}) \right| > \delta\right\} \\ &\quad \leq \Pr\left\{\left| f(\mathbf{X}_{N+1}) - \widetilde{f}_N(\mathbf{X}_{N+1}) \right| > \frac{\delta}{2}\right\} + \Pr\left\{\left|\epsilon_{N+2}\right| > \frac{\delta}{2}\right\} \end{aligned}$$

Since only the first term depends on λ in the regularization procedure, it is sufficient to consider the Chebyshev bound

$$\begin{aligned} &\Pr\left\{\left| f(\mathbf{X}_{N+1}) - \widetilde{f}_N(\mathbf{X}_{N+1}) \right| > \delta\right\} \\ &\quad \leq \mathrm{E}\left[\left| f(\mathbf{X}_{N+1}) - \widetilde{f}_N(\mathbf{X}_{N+1}) \right|^2\right] / \delta^2 \end{aligned} \qquad (5)$$

which requires only the joint distribution of the process up to time $N+1$ to compute. Roughly speaking, when f and the disturbance process $\{\epsilon_i\}$ are such that $\{\mathbf{X}_i\}$ has a stationary distribution, the expectation on the r.h.s. of (5) is closely approximated by $\mathrm{E}[T(\lambda)]/N$ as $N \to \infty$. Therefore selecting λ to minimize $\mathrm{E}[T(\lambda)]$ asymptotically minimizes the probability of a large deviation in the prediction error on the next time step, i.e., the next time series datum outside of the current process sample. The authors are currently working on extending this interpretation to more general cases in which the existence of a stationary distribution cannot be assumed.

3. NONSTATIONARITY AND THE DYNAMIC REGULARIZED GARBF PREDICTOR

When the process described in (1) is stationary, it is clear that the statistical characteristics of the predictor function $\widetilde{f}_N$ are the same whether it was designed using the process sample $\mathcal{S}_k \triangleq \{x_i\}_{i=k}^{k+N}$ or $\mathcal{S}_{k+n} \triangleq \{x_i\}_{i=k+n}^{k+n+N}$, $n, k = 1, 2, \ldots$ On the other hand, many situations exist for which the disturbance process $\{\epsilon_i\}$ is not i.i.d. or f is not sufficiently well-behaved to ensure the existence of a stationary distribution for the process $\{X_i\}$. In such cases, if we assume slowly varying process statistics, the idea of dynamically updating the GaRBF predictor periodically by recomputing the solution to (3) for $\mathcal{S}_n$, $\mathcal{S}_{n+1}$, $\ldots$, has some intuitive appeal. To be more precise, the suggested prediction-update algorithm is

1. given $\mathcal{S}_k = \{x_i\}_{i=k}^{k+N}$, compute the GaRBF expansion coefficients $\mathbf{c}$ according to (3).
2. predict $\widetilde{x}_{k+N+1} \triangleq \widetilde{f}_N(\mathbf{x}_{k+N})$.
3. when the actual x_{k+N+1} becomes available, update the interpolation matrix $\mathbf{G}$ according to $\mathcal{S}_{k+1}$.
4. with the updated $\mathbf{G}$, compute the new value of $\mathbf{c}$ according to (3). Together with the new set of GaRBF expansion centres $\mathcal{S}_{k+1}$, the updated function $\widetilde{f}_N$ is defined according to (2).
5. repeat from step 2 with $k \rightarrow k+1$.

A few comments are in order here. In step 4, a full matrix inversion of the updated $\mathbf{G} + \lambda\mathbf{I}$ is not necessary if the change to $\mathbf{G}$ is restricted to updating only those elements which are related to the x_k being replaced, as described in step 3. With such a small rank change in $\mathbf{G}$, the matrix inversion lemma may be applied to reduce the computational requirements for the new $\mathbf{c}$ from $\mathcal{O}\left(N^3\right)$ to $\mathcal{O}\left(N^2\right)$. Indeed, the simulation results for both full and partial updating of $\mathbf{G}$ in step 3 of the dynamic GaRBF predictor show that such an incremental approach to its evolution is a viable tradeoff between the algorithm's computational load and its performance for processes of slow statistical variation. Note also that while we have limited the resources used in the GaRBF predictor to N basis functions, we have not otherwise constrained the choice of λ and norm weighting matrix $\mathbf{W}$; in practice, we have found that $\mathbf{W}$ recursively estimated from the empirical covariance matrix of the process sample and a sufficiently small λ provide reasonable results when N is adequately large [4]. Of course, if N is very small, say less than ten, then the performance is correspondingly more sensitive to truly optimal choices of λ and $\mathbf{W}$ in the algorithm.

4. SIMULATION RESULTS

The proposed dynamic regularized GaRBF predictor is applied to a 10,000 point speech sample and compared with the G.721 standard AT&T IIR filter predictor trained according to a nonlinear LMS algorithm. The speech data, which appear to have no discernible noise, are a male voice sampled at 8kHz and 8bits per sample while speaking the fragment "When recording audio data ...". Before processing, the speech data were approximately recentred to zero-mean and normalized to unit total amplitude range. Table 1 gives the absolute and relative SNR for the two predictors over the speech sample, where "noise" in this case refers to prediction error. Only results for which the GaRBF predictor exhibited comparable or better performance than the AT&T predictor are included. In all cases, the GaRBF predictor is designed with $\lambda = 0.01$ and diagonal $\mathbf{W}^{-1}$ uniformly set to the trace of the process sample's empirical covariance matrix at each time step, as these choices appeared to yield reasonable figures in some preliminary simulations. Note that this choice of updating $\mathbf{W}$ leads to a full-rank change in $\mathbf{G}$ at every time step, implying that the results have complexity $\mathcal{O}\left(N^3\right)$.

GaRBF parameters	SNR (dB)	dB/AT&T
$N = 10$, $p = 50$	9.869	-0.0958
$N = 10$, $p = 100$	9.955	-0.0102
$N = 20$, $p = 50$	10.35	0.3804
$N = 20$, $p = 100$	10.68	0.7150
$N = 50$, $p = 50$	9.887	-0.0774
$N = 50$, $p = 100$	11.48	1.5113
$N = 100$, $p = 2$	9.742	-0.2234
$N = 100$, $p = 5$	11.48	1.5115
$N = 100$, $p = 10$	13.10	3.1372
$N = 100$, $p = 50$	14.37	4.4081
$N = 100$, $p = 100$	13.73	3.7664
AT&T pred.	10.01	0

Table 1: GaRBF predictor (full update algorithm) results for 10,000 point speech waveform

Window (250 pts.)	SNR (dB)	dB/AT&T
1	12.82	-0.52
2	11.93	2.36
3	12.62	4.64
4	21.09	8.33
Over all 1000 pts.	12.64	3.00
Segmental SNR	14.61	3.70

Table 2: GaRBF predictor (partial update algorithm) results for $N = 100$, $p = 50$, $\lambda = 0.01$

For the same data, Figures 1 and 2 show, over 30ms windows, how the $N = 100$, $p = 50$, $\lambda = 0.01$ GaRBF predictor apparently captures the dynamics of both the slowly and quickly varying portions of the speech sample. The corresponding segmental SNR (SEGSNR) and total SNR for these and the subsequent two windows are compared with those of the AT&T predictor in Table 2. In this case, the updated norm weighting matrix $\mathbf{W}$ at each time step is used only in calculations involving the newest datum x_{k+N+1} in step 3 of the dynamic regularized GaRBF predictor algorithm; this induces only a partial update to the interpolation matrix $\mathbf{G}$ at each time step and so allows the matrix inversion lemma to be exploited as mentioned previously,

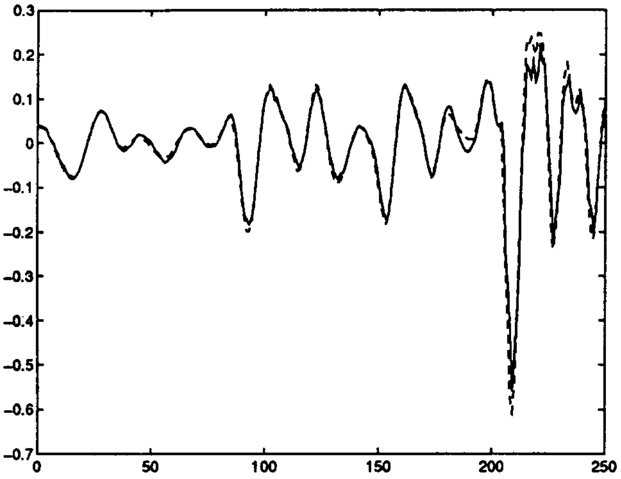

Figure 1: Proposed predictor performance over 1st window of 250 points ('–' is predicted, '- -' is actual)

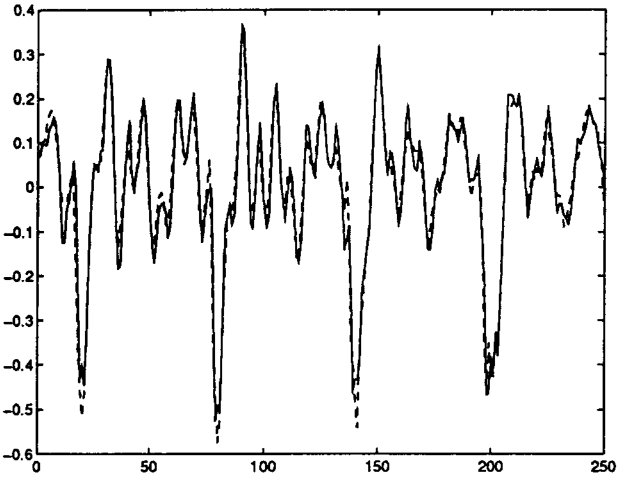

Figure 2: Proposed predictor performance over 2nd window of 250 points ('–' is predicted, '- -' is actual)

resulting in an $\mathcal{O}\left(N^2\right)$ complexity. Experience shows that the decrease in the prediction performance between the algorithm using partial versus full updates to **G** is acceptable. Nevertheless, care must be taken in implementing the partial updating scheme to avoid the accumulation of round-off errors and the possibility of ill-conditioning in $(\mathbf{G}+\lambda\mathbf{I})^{-1}$ as it is propagated from time step to time step; failure to do so can lead to seemingly discontinuous outputs from the predictor. The authors conjecture that these numerical precision difficulties can be alleviated by the judicious development of equivalents to square-root filters and their related counterparts in the linear case.

From these results, one sees that without any voice-specific prior knowledge or optimizations, the proposed dynamic regularized GaRBF predictor can attain or surpass the performance offered by the AT&T predictor. As may be expected, the results demonstrate the existence of a tradeoff between the number of basis functions N and the predictor memory p required to obtain a given level of performance. With sufficiently high memory, e.g., $p = 100$, the GaRBF predictor with as few as ten basis functions can match the performance of the AT&T speech predictor. On the other hand, with a sufficient number of basis functions, e.g., $N = 100$, the proposed predictor need only have a memory of two to provide performance comparable to that of the AT&T predictor. The set of results with one hundred basis functions and memory greater than five, despite being computationally burdensome, is included to show that the proposed GaRBF predictor's parameterization is flexible enough to model the underlying dynamics of the test speech signal to a degree not possible with the simpler AT&T speech predictor. The results also suggest that there is room for non-trivial improvement in the performance of the AT&T predictor.

5. CONCLUSIONS

Applying the principles of regularized fitting, in this paper we have described a nonparametric predictor based on a Gaussian basis function expansion for a general class of nonlinear AR processes. The relevance of regularization to the one-step prediction of such NLAR processes is examined for the stationary case. In the nonstationary case, an algorithm for dynamically updating the GaRBF predictor is described and shown to meet or exceed the performance of the AT&T standard speech predictor when tested on a 10,000 sample speech waveform. Although there is, of course, a complexity tradeoff between the amount of prior knowledge embedded in the AT&T predictor and the nonparametric nature of the dynamic regularized GaRBF predictor, the results indicate that the proposed predictor can be a viable base upon which further speech-specific optimizations may be applied to achieve significant gains in predictor performance.

6. REFERENCES

[1] Y. Kang. Lee and Don H. Johnson, "Nonparametric prediction of non-gaussian time series", *in Proc. ICASSP vol.4*, pp. 480–483, 1993.

[2] M. Pawlak and W. Greblicki, *Nonparametric estimation of a class of nonlinear time series models*, vol. 335 of *NATO ASI C*, pp. 541–552, Kluwer, 1991.

[3] D. Bosq, *Modelization, nonparametric estimation and prediction for continuous time processes*, vol. 335 of *NATO ASI C*, pp. 509–529, Kluwer, 1991.

[4] M. Von Golitschek and L. L. Schumaker, "Data fitting by penalized least squares", in J. C. Mason and M. G. Cox, editors, *Algorithms for approximation II*, pp. 210–227, London, Great Britain, July 1990. Cranfield Institute of Technology, Chapman and Hall.

[5] T. Poggio and F. Girosi, "Networks for approximation and learning", *Proc. IEEE*, vol. 78, pp. 1484–1487, 1990.

[6] P. Yee and S. Haykin, "Pattern classification as an ill-posed, inverse problem: a regularization approach", *internal manuscript*, Feb. 1994.

[7] H. Tong, *Non-linear time series: a dynamical systems approach*, Oxford Science, 1990.