BAYESIAN ADAPTIVE FILTERING: PRINCIPLES AND PRACTICAL APPROACHES

Tayeb Sadiki, Dirk T.M. Slock

Eurecom Institute
2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
Tel: +33 4 9300 2656/2606; Fax: +33 4 9300 2627
e-mail:{sadiki,slock}@eurecom.fr

ABSTRACT

While adaptive filtering is in principle intended for tracking non-stationary systems, most adaptive filtering algorithms have been designed for converging to a fixed unknown filter. When actually confronted with a non-stationary environment, they possess just one parameter (stepsize, forgetting factor) to adjust their tracking capability. Virtually the only existing optimal approach is the Kalman filter, in which the time-varying optimal filter is modeled as a vector AR(1) process. The Kalman filter is in practice never applied as an adaptive filter because of its complexity and large number of unknown parameters in its state-space (AR(1)) model. Here we consider optimal adaptive filtering for any stationary optimal filter evolution. We emphasize the various aspects of an optimal Bayesian approach, which not only include parameter variation bandwidth but also a priori parameter size and parameter dynamics. Finally we recommend some constrained versions of modest complexity and show how to estimate the parameters in the resulting Bayesian adaptive filters.

1. STATE OF THE ART

Since the introduction of the LMS algorithm by Widrow and Hopf in the 1960's, most of the further work in adaptive filtering has focused on improving the initial convergence. The Recursive Least-Squares (RLS) algorithm was also developed in the 1960's and provided an alternative algorithm for adaptive system identification. The RLS algorithm is recursive and not iterative as the LMS algorithm, solving a LS cost function exactly at each update. As a result it converges very fast since it provides an unbiased solution once the LS problem gets overdetermined. This deterministic aspect adds up to the observation that the RLS convergence is insensitive to the input signal correlation structure (approximately, since there is some dependence on the initialization). The RLS algorithm, though providing computational savings w.r.t. the plain solving of LS problems at each sampling period, is quite a bit more expensive than the LMS algorithm. This motivated on the one hand the development of fast RLS algorithms, and on the other hand the development of an intermediate category of algorithms, all less sensitive than LMS to the input correlation structure, including frequency or other transform domain LMS algorithms, prewhitened LMS versions, Fast Newton Transversal Filters and (Fast) Affine Projection Algorithms.

Eurécom's research is partially supported by its industrial partners: Hasler Foundation, Swisscom, Thales Communications, ST Microelectronics, CEGETEL, France Télécom, Bouygues Telecom, Hitachi Europe Ltd. and Texas Instruments. The work reported herein was also partially supported by a PACA regional scholarship.

At the outset, all these algorithms are developed to converge to an unknown optimal filter. When this optimal filter is actual time-varying, these algorithms need to be made adaptive. The RLS algorithms are made adaptive by the introduction of a weighting function/window. The weighted LS cost function can be viewed as the output of a filter with the instantaneous squared filtering error sequence as input. The filter should be such that its input-output relationship is simple and recursive. The LS cost function uses a discrete-time integrator as filter, which can be easily modified into a first-order recursive filter for the exponentially weighted RLS algorithm. The sliding window RLS algorithm uses a moving average filter that can also be expressed recursively. All other adaptive filtering algorithms are made adaptive by the introduction of a scalar stepsize. In fact, the time-varying stepsize sequence of stochastic gradient algorithms [1] is made time-invariant/constant to avoid convergence and permit tracking of time-varying optimal filter settings. The tracking characteristics of the LMS and RLS algorithms got analyzed only in the 1970's and 1980's, 10 to 20 years after the introduction of the algorithms, in [2] for LMS and [3] for RLS. A further inspection of these tracking characteristics revealed the surprising result that in certain cases the LMS algorithm may provide better tracking than the RLS algorithm (each with optimized stepsize or forgetting factor), see [4] for deterministic and e.g. [1] for random parameter variations. With hindsight, this is not at all surprising since LMS and RLS are just two suboptimal approaches to tracking time-varying parameters. Whereas initial convergence is about the fast reduction of the mean parameter error vector, tracking is about the optimal compromise between MSE due to estimation noise and tracking/lag noise.

The RLS algorithm got introduced after the Kalman filter (KF) was invented, though the RLS algorithm is a special case of the KF for the following state-space model [5]

$$H_k^o = H_{k-1}^o \quad (1)$$

$$x_k = Y_k^T H_k^o + v_k \ . \quad (2)$$

The KF formulation requires immediately a parametric form of the optimal filter, usually a FIR filter is assumed with impulse response of N coefficients contained in the vector H_k^o. The *measurement equation* (2) expresses that the *desired-response signal* x_k is the sum of the output $\sum_{i=0}^{N-1} H_{k,i}^o y_{k-i}$ of the optimal FIR filter H_k^o with *input* y_k plus an independent *measurement noise* v_k. In KF terminology, x_k would be the *measurement* and H_k^0 the *state*.

Wiener filtering (WF'ing) is about estimating one random signal from another, let's say estimating the signal x_k on

the basis of the signal y_k. In the *system identification* (*sysid*) set-up of adaptive filtering, which is reflected in (1)-(2), the relation between these two signals is that x_k is assumed to be output of an unknown system/plant with y_k as input plus independent measurement noise v_k. In this case, the optimal Wiener (LMSSE) filter is clearly $R_{xY}R_{YY}^{-1} = H^{oT}$, which is FIR if the system to be identified is FIR. The WF is based on the joint statistical description of the random signals x_k and y_k. and is a deterministic quantity. The WF solution is not influenced by the color of the noise v_k.

KF'ing is in principle a special case of the signal-in-noise case of WF'ing. In the signal-in-noise case, the measurement signal is the sum of the signal to be estimated plus noise. For the KF, the signal to be estimated satisfies furthermore a state-space model. The adaptive filtering/RLS application of KF'ing though deviates significantly from this spirit. In RLS, the quantity (state) estimated is the set of WF coefficients H_k^o instead of its output, the filter input y_k is considered deterministic (the estimation is given y_k) and hence the filter estimate would be random if y_k would be considered random. Indeed, the KF provides an estimate H_k of the WF H_k^o. Since this KF application is now an instance of parameter estimation, the parameter estimation quality depends on e.g. the color of the noise v_k.

The KF'ing framework can be straightforwardly extended to incorporate time-varying optimal parameters. The simplest way is probably through the following stationary AR(1) model state equation for the optimal filter variation [5]

$$H_k^o \;=\; F\, H_{k-1}^o + W_k \tag{3}$$

replacing (1), where $\mathrm{E}\, W_k W_i^H = Q\,\delta_{ik}$, $\mathrm{E}\, W_k v_i^H = 0$ (noises assumed circular in complex case). This formulation lead to the widely accepted point of view that the KF would be the optimal adaptive filter. This is indeed true for the *sysid* configuration with (3)-(2) as *assumed correct model* and F, Q and r in $\mathrm{E}\, v_k v_i^H = r\,\delta_{ik}$ *assumed known*. We may note that in this model, WF'ing provides the time-varying optimal filter $H_k^{oT} = R_{x_k Y_k} R_{Y_k Y_k}^{-1}$ and the KF estimates it in a Bayesian (LMMSE) sense.

The problem with the KF viewpoint is that the model parameters, if at all the model is correct, are unknown and need to be estimated also from the same data. Those parameters can be inferred from the joint signal statistics, just like the WF itself. However, in the KF, the input signal y_k is considered deterministic which makes the state space model (3)-(2) linear but time-varying. These complications lead to approximate approaches such as exponentially weighted RLS, which can be shown [6] to correspond to the KF for certain artificial choices of F and Q_k in ([**?**]). The main issue in most applications is the so-called *generalization property* of statistical learning: what counts is the adaptive filter performance not for the given input signal realization, but when applied to other signal data, hence for the given signal statistics. The generalization capacity may be hampered by sticking too closely to one model's details when the model is approximate. Another issue is that the KF approach for tracking time-varying optimal filters only applies in the *sysid* configuration in which the filter's non-stationarity arises in the crosscorrelations between input and desired-response signals, regardless of the statistics ((non)stationarity) of the input. Communications applications of the *sysid* configuration are channel estimation and echo cancellation. In all other configurations of adaptive filtering: prediction, deconvolution/equalization and interference cancellation, the statistics of the optimal filter may be strongly intertwined with the statistics of the the input signal. In linear prediction for instance, the desired-response and input signals are the same. One rarely sees the linear prediction problem addressed as a ML estimation of or KF'ing on the parameters of an AR model, because any AR model order is likely to lead to an approximation error. Adaptive prediction is in fact a joint operation of approximation (e.g. through model order selection) and estimation. In equalization, even if the channel variation could be modeled as an AR(1) model as in (3), the optimal equalizer setting is a nonlinear function of the channel. Given all these considerations, the best practical approach is probably to specify a motivated solution structure of acceptable complexity and optimize the parameters within that structure (as is done in linear prediction) (approximation/estimation compromise). The problem considered here has of course been addressed previously and we now discuss some of this existing work.

1.1 Tracking Bandwidth Adjustment

Most of the work on adapting tracking capability has focused on adapting one tracking parameter. In RLS, it doesn't cost any computational complexity to make the forgetting factor (FF) time-varying. Modifications to fast RLS algorithms to allow a time-varying FF, as well as algorithms to adjust this FF on the basis of correlation matching have been pursued in [7]. The equivalent development for LMS algorithms concerns Variable StepSize (VSS) algorithms. Important developments were presented in [8],[9]. Most of the VSS algorithms use the steepest-descent strategy and the instantaneous squared error cost function of the LMS algorithm to adjust the additional parameter, which is the stepsize. A related but different approach consists in running various adaptive filters with different time constants and selecting or combining their outputs, similarly to what is done in model order selection, see [10],[11],[12],[13].

A further refinement is to allow different tracking bandwidths for differents filter components as is done in [14] with a VSS per filter coefficient and in [15] where the tracking capacity increases with frequency for the various frequency domain components of the filter. The work in [14] essentially shows that a "diagonal" state-space model (3) may allow a simplification of the KF to a LMS algorithm with a VSS per tap, but no attempt is made to automatically adjust the resulting stepsizes.

1.2 Power Delay Profile

Besides the statistical modeling of the parameter variation, another important ingredient in Bayesian adaptive filtering is the incorporation of prior knowledge on the coefficient sizes. Indeed, when tracking time-varying filters, it becomes possible to learn the variances of the filter coefficients. This aspect has been exploited for a while in a rudimentary, binary form for sparse filters: filter coefficients are either adapted or deemed to small and kept zero (for each filter coefficient, the stepsize is either 0 or a constant). More recently, a smoother evolution of the stepsize has been introduced, leading to the Proportionate LMS (PLMS) algorithm, motivated e.g. by acoustic echo cancellation in which the adaptive filter has many coefficients, but their value tapers off, see [16],[17].

Similar prior information is starting to be taken into account for (LMMSE) channel estimation in wireless communications [18], where the evolution of the channel coefficient variances along the impulse response is called the power delay profile.

1.3 Full Bayesian Approach

In a full Bayesian approach, the whole matricial spectrum $S_{H^o}(z) = S_{H^oH^o}(z)$ of H_k^o counts: not only the parameter variation speed/bandwidth but the whole spectral shape counts, not only the spectral shape but also the power delay profile counts, and in principle also the cross spectra between coefficients need to be accounted for.

The KF [5] allows to do all this in the *sysid*set-up, but ignores the estimation of $S_{H^o}(z)$. In [19], a point of view close to the one of this paper is developed. However, they require the knowledge of the (multivariate IIR) matricial spectrum of the (standard) adaptive filter gradient (this could be estimated from the observations of the gradient) and knowledge of the (multivariate IIR) matricial spectrum of the stationary filter parameter vector. This last requirement is quite unrealistic. Furthermore, the design steps suggested may be quite sensitive to estimation errors to some quantities that get estimated.

2. MODELING OF STANDARD ADAPTIVE FILTER BEHAVIOR

The adaptive filter is H_k and the a prior error signal is $e_k = x_k - Y_k^T H_{k-1}$. Consider the (complex) LMS algorithm first

$$H_k^{lms} = H_{k-1}^{lms} + \mu\, Y_k^*\, e_k \tag{4}$$

whereas the RLS filter update is of the form

$$H_k^{rls} = H_{k-1}^{rls} + \widehat{R}_k^{-1}\, Y_k^*\, e_k \tag{5}$$

where $\widehat{R}_k = \lambda\, \widehat{R}_{k-1} + Y_k^* Y_k^T$. Let $R = \mathrm{E}\, Y_K^* Y_k^T$. Then, assuming the adaptation speed is not too fast, we get approximately

$$\begin{aligned} H_k^{lms} &= \left[I - (I-\mu R)\, q^{-1}\right]^{-1} \mu R\, (H_k^o + R^{-1}\, Y_k^*\, v_k) \\ H_k^{rls} &= \frac{1-\lambda}{1-\lambda\, q^{-1}}\, (H_k^o + R^{-1}\, Y_k^*\, v_k) \end{aligned} \tag{6}$$

where $q^{-1} H_k = H_{k-1}$. Using averaging analysis at low adaptation speed, these results for the *sysid*set-up hold approximately also for the other adaptive filtering applications. Note that $H_k^o + R^{-1}\, Y_k^*\, v_k$ is closely related to $G_k = R^{-1}\, Y_k^*\, x_k$, which is a mixed quantity in that it is averaged in the input covariance but instantaneous in the input-desired-response correlation.

One may remark that in the context of tracking (slowly) time-varying parameters, the exact least-squares property of RLS becomes quite unimportant. In fact, the main property that continues to count is the decorrelation property leading to insensitivity of the tracking dynamics to the variation in the input signal spectrum.

3. GENERAL BAYESIAN ADAPTIVE FILTER SOLUTION

We shall introduce, mostly for the purpose of analysis, a somewhat idealized Bayesian solution which is based on the assumption that R can be estimated well. This solution will be based on LMMSE estimation (Wiener filtering) of H_k from

$$G_k = R^{-1}\, Y_k^*\, x_k = H_k^o + R^{-1}\, Y_k^*\, v_k + (R^{-1}\, Y_k^* Y_k^T - I)\, H_k^o \tag{7}$$

where for slow parameter variations, the last term can be neglected since it is the product of low-pass noise H_k^o with high-pass noise $R^{-1}\, Y_k^* Y_k^T - I$. The optimal Bayesian adaptive filter would be to apply the KF to (7), $G_k = H_k + \widetilde{G}_k$, which can be considered as a measurement equation for the state H_k^o. In steady-state, the KF converges to the WF

$$H_k = F(q)\, G_k \tag{8}$$

where in the non-causal case

$$F(q) = S_{H^oG}(q)\, S_{GG}^{-1}(q) = I - S_{\widetilde{G}\widetilde{G}}(q)\, S_{GG}^{-1}(q) \tag{9}$$

Neglecting the last term in (7) and assuming that v_k is white noise (hence $\widetilde{G}_k$ is white), we have $S_{\widetilde{G}\widetilde{G}}(q) = R_{\widetilde{G}\widetilde{G}} = \sigma_v^2\, R^{-1}$. Hence the non-causal WF is fairly straightforward to find since $S_{GG}(q)$ can be estimated simply from the observations of G_k, though σ_v^2 is somewhat trickier to derive from the observed MSE (the details are omitted here). For the causal case, consider $\overline{G}_k = P(q)\, G_k$ where $P(q)$ is the (∞ length) (monic) multivariate prediction error filter for the vector signal G_k and $\overline{G}_k$ is the resulting white prediction error with covariance matrix $R_{\overline{G}\overline{G}}$. Then the causal WF is

$$F(q) = \left\{S_{H^oG}(q)\, P^\dagger(q)\right\}_+ R_{\overline{G}\overline{G}}^{-1} P(q) = I - R_{\widetilde{G}\widetilde{G}} R_{\overline{G}\overline{G}}^{-1} P(q) \tag{10}$$

which is based on the same quantities as the non-causal WF. It can be shown that in the case of FIR Wiener filtering (causal or not), an expression similar to the one in (10) holds in which $P(q)$ would then denote the LMMSE estimation error filter for estimating G_k on the basis of the other G's involved, and $R_{\overline{G}\overline{G}}$ denotes the corresponding estimation error covariance matrix. The use of a general filter $F(q)$ will lead to an estimation error $\widetilde{H}_k = H_k^o - H_k = (I - F(q))H_k^o - F(q)\widetilde{G}_k$ with covariance matrix

$$R_{\widetilde{H}\widetilde{H}} = \oint \frac{dz}{2\pi j z} (I-F) S_{H^oH^o} (I-F)^\dagger + \oint \frac{dz}{2\pi j z} F R_{\widetilde{G}\widetilde{G}} F^\dagger \tag{11}$$

where $F = F(z)$ and $S_{H^oH^o} = S_{H^oH^o}(z)$, and results in Excess MSE (EMSE) $\mathrm{tr}\left\{R_{\widetilde{H}\widetilde{H}} R\right\}$.

4. STRUCTURED/REDUCED-COMPLEXITY CASES OF INTEREST

The analysis is easiest when the input is white, $R = I$. It is of interest to analyze the following structured models for the optimal filter Doppler spectrum:

(i) subspace model: $H_k^o = A\, W_k$ where A is tall and $S_{WW}(z)$ is diagonal

(ii) decoupled coefficient dynamics: $S_{H^oH^o}$ diagonal

(iii) uniform dynamics plus power delay profile: $S_{H^oH^o}(z) = S_{hh}(z)\, D$ where S_{hh} is scalar and D is a constant diagonal

A very low complexity solution for (ii) or (iii) would be a LMS algorithm with individual stepsizes for the coefficients.

It is of interest to compare the resulting EMSE with optimized individual stepsizes to the classical LMS with an optimized global stepsize. For case (iii), if the power delay profile (D) is binary then in the classical solution the EMSE will be proportional to the total number of adaptive filter coefficients whereas in the optimized individual stepsize case, only the coefficients with non-zero variance will contribute. Also, in case (iii), with $D = I$, one can come up with a spectrum $S_{hh}(z)$ for which RLS is optimum.

5. ADAPTIVE BAYESIAN ADAPTIVE FILTERING

5.1 Predictive Bayesian Adaptive Filter

In the case of white input (such as in communications channel estimation or electrical echo cancellation, in which cases the input is the transmitted symbol sequence), the ideal Bayesian adaptive filter introduced above is immediately applicable. Constrained structure solutions such as LMS-variants (individualized stepsize) can be considered. Classical VSS solutions can also be applied to individualized stepsizes.

In the case of colored inputs, one can go to transform domain LMS: frequency domain, subbands, wavelets. Or use FNTF, prewhitened LMS or IV-LMS with Instrumental Variable (IV) $z_k = S_{yy}^{-1}(q)\, y_k$.

5.2 Two-Stage Solutions

In this case we consider a first stage with a fast standard adaptive filter, e.g. NLMS with stepsize equal to 1. The second stage is a Wiener filter on the filter estimates provided by the first stage. Even in fast time-varying filter cases, the filter variation bandwidth will typically be only a fraction of the signal bandwidth. This means that the filter estimates from the first stage can be (first lowpass filtered (by e.g. simple averages) to reduce the estimation noise and) subsampled and WF (or KF) can be applied in the second stage, to a slightly modified form of (7) (with quite generally white measurement noise due to the subsampling) or perhaps even to a modified form of (6) if one doesn't want to neglect the adaptation dynamics in the first stage. This second stage then becomes similar to the filtering approaches suggested for brute periodic channel estimates in [20],[21]. Since the 2nd stage works at reduced rate, the added complexity of working in 2 stages becomes acceptable.

REFERENCES

[1] D.T.M. Slock. "On the Convergence Behavior of the LMS and the Normalized LMS Algorithms". *IEEE Trans. on Signal Processing*, 41(9):2811–2825, Sept. 1993.

[2] B. Widrow *et al.* "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter". *Proc. IEEE*, 64(8):1151–1162, Aug. 1976.

[3] E. Eleftheriou and D. Falconer. "Tracking Properties and Steady-State Performance of RLS Adaptive Filter Algorithms". *IEEE Trans. ASSP*, ASSP-34(5):1097–1110, Oct. 1986.

[4] O.M. Macchi and N.J. Bershad. "Adaptive Recovery of a Chirped Sinusoid in Noise, Part I: Performance of the RLS Algorithm, Part II: Performance of the LMS Algorithm". *IEEE Trans. SP*, SP-39:583–602, March 1991.

[5] S. Haykin, A.H. Sayed, J.R. Zeidler, P. Yee, and P.C. Wei. "Adaptive tracking of linear time-variant systems by extended RLS algorithms". 45(5):1118–1128, May 1997.

[6] B.D.O. Anderson and J.B. Moore. *"Optimal Filtering"*. Prentice-Hall, Englewood Cliffs, NJ, 1979.

[7] D.T.M. Slock. "Fast Transversal Filters with Data Sequence Weighting". *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37(3):346–359, March 1989.

[8] V.J. Mathews and Z. Xie. "A Stochastic Gradient Adaptive Filter with Gradient Adaptive Step Size". *IEEE Trans. Signal Proc.*, 41(6):2075–2087, June 1993.

[9] T. Aboulnasr and K. Mayyas. "A Robust Variable Step-Size LMS-Type Algorithm: Analysis and Simulations". *IEEE Trans. on Signal Proc.*, 45(3), March 1997.

[10] W.J. Song and M.S. Park. "A Comlementary Pair LMS Algorithm for Adaptive Filtering". In *Proc. ICASSP*, Münich, Germany, Apr. 1997.

[11] W.S. Chaer, R.H. Bishop, and J. Ghosh. "A Mixture-of-Experts Framework for Adaptive Kalman Filtering". *IEEE Trans. Systems, Man and Cybernetics, Part B*, 27(3), June 1997.

[12] S.S. Kozat and A.C. Singer. "Further Results in Multistage Adaptive Filtering". In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Proc.*, March 2002.

[13] J. Arenas-Garcia, V. Gomez-Verdejo, M. Martinez-Ramon, and A.R. Figueiras-Vidal. "Separate-Variable Adaptive Combination of LMS Adaptive Filters for Plant Identification". In *Proc. IEEE NNSP Workshop*, Toulouse, France, Sept. 2003.

[14] W. Liu. "Performance of Joint Data and Channel Estimation Using Tap Variable Step-Size (TVSS) LMS for Multipath Fast Fading Channel". In *Proc. Globecom*, pages 973–978, 1994.

[15] D.T.M. Slock. "Fractionally-Spaced Subband and Multiresolution Adaptive Filters". In *Proc. ICASSP*, Toronto, Canada, May 1991.

[16] R.K. Martin and Jr. C.R. Johnson. "NSLMS: a Proportional Weight Algorithm for Sparse Adaptive Filters". In *Proc. Asilomar Conf. Signals Systems and Computers*, Pacific Grove, CA, USA, 2001.

[17] J. Benesty and S.L. Gay. "An Improved PNLMS Algorithm". In *Proc. ICASSP*, Orlando, FL,USA, 2002.

[18] D. Schafhuber, G. Matz, F. Hlawatsch, and P. Loubaton. "MMSE Estimation of Time-Varying Channels for DVB-T Systems with Strong Co-Channel Interference". In *Proc. European Signal Processing Conference (EUSIPCO)*, Toulouse, France, Sept. 2002.

[19] M. Sternad, L. Lindbom, and A. Ahlèn. "Wiener Design of Adaptation Algorithm with Time-Invariant Gains". *IEEE Trans. on Signal Proc.*, 50(8), Aug. 2002.

[20] M. Lenardi and D. Slock. "Estimation of Time-Varying Wireless Channels and Application to the UMTS W-CDMA FDD Downlink". In *Proc. European Wireless (EW)*, Florence, Italy, Feb. 2002.

[21] G. Montalbano and D.T.M. Slock. "Joint Common-Dedicated Pilots Based Estimation of Time-Varying Channels for W-CDMA Receivers". In *Proc. Vehic. Tech. Conf. (VTCfall)*, Orlando, FL, USA, Sept. 2003.