BAYESIAN METHODS FOR SPARSE RLS ADAPTIVE FILTERS

H. Koeppl, G. Kubin

Christian Doppler Laboratory
for Nonlinear Signal Processing
Graz University of Technology, Austria
{heinz.koeppl,gernot.kubin}@tugraz.at

G. Paoli

System engineering group,
Infineon Technologies,
Design Center Villach, Austria
gerhard.paoli@infineon.com

ABSTRACT

This work deals with an extension of the standard recursive least squares (RLS) algorithm. It allows to prune irrelevant coefficients of a linear adaptive filter with sparse impulse response and it provides a regularization method with automatic adjustment of the regularization parameter. New update equations for the inverse auto-correlation matrix estimate are derived that account for the continuing shrinkage of the matrix size. In case of densely populated impulse responses of length M, the computational complexity of the algorithm stays $\mathcal{O}(M^2)$ as for standard RLS while for sparse impulse responses the new algorithm becomes much more efficient through the adaptive shrinkage of the dimension of the coefficient space. The algorithm has been successfully applied to the identification of sparse channel models (as in mobile radio or echo cancellation).

1. INTRODUCTION

Linear-in-parameters models

$$z[n] = \mathbf{w}^T[n]\mathbf{x}[n] + \epsilon[n] \tag{1}$$

with the observed noisy output $z[n]$, the weight vector $\mathbf{w}[n] \equiv [w_1[n], \ldots, w_M[n]]^T$, the input data vector $\mathbf{x}[n] \equiv [x[n], \ldots, x[n-M+1]]^T$ and the additive perturbation $\epsilon[n]$, are considered. Many applications of these models share the features that the excitation signal $x[n]$ for the adaptive system is not always persistently exciting and that the structure of the model does not match the structure of the reference system. One mismatch example would be a too high order of the adaptive filter. In the first case the covariance matrix estimate blows up such that the adaptive algorithm gets unstable. A common stabilization method for such situations is the regularization of the auto-correlation matrix estimate [1]. The second feature of model mismatch is due to the incomplete insight into the structure of the reference system. To guarantee some predefined error power after convergence, one has to select a conservative, i.e. overestimated, model structure which takes into account our incomplete knowledge about the reference system. In the case of an echo-canceler, where the echo-impulse response varies significantly over different environments, one has to initialize a conservative model which can handle the longest impulse-response expected to occur in practice. The inclusion of parameters in the model that are irrelevant from the viewpoint of a decrease in the error still causes an increase in the variance of the parameter estimates $\hat{\mathbf{w}}[n]$ of the adaptive system compared to the variance of the estimates for an exactly matching model structure. In addition, the tracking performance of the adaptive filter gets reduced due to the inclusion of irrelevant parameters.

In the statistics and machine learning literature this problem gets addressed by subset selection algorithms. In this work the algorithm proposed in [2], which simultaneously performs subset selection and adaptive regularization, is incorporated in a recursive least squares adaptive algorithm.

The Bayesian treatment of regularization using the evidence procedure [3] offers a simple way to estimate the regularization parameter and even allows an extension to estimate a regularization matrix [2]. In the adaptive filter literature, regularization methods can be found in, e.g. [1] and [4], on which the following presentation is based. Opposed to our contribution, these two works share the fact that no adaptive computation of the regularization term is considered.

2. DERIVATION OF THE ALGORITHM

2.1. Bayesian estimation

A Bayesian formulation of the estimation problem for the linear regression model (1) starts with the definition of the likelihood function $p(\mathbf{z}[n]|\mathbf{w}[n])$ and the prior distribution $p(\mathbf{w}[n]|\mathbf{A}[n])$ for the weights $\mathbf{w}[n]$ given the prior distribution parameter $\mathbf{A}[n]$ at a given sampling instant n, with $\mathbf{z}[n] \equiv [z[n], \ldots, z[1]]^T$. In the following, for the sake of conciseness, the sampling time index n is omitted. It will be reintroduced in section 2.3 where recursive relations are obtained. For an additive white Gaussian noise model in (1)

0-7803-8104-1/03/$17.00 ©2003 IEEE

the likelihood function reads

$$p(\mathbf{z}|\mathbf{w}) = (2\pi)^{-\frac{n}{2}}|\mathbf{\Lambda}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{X}\mathbf{w})^T\mathbf{\Lambda}(\mathbf{z}-\mathbf{X}\mathbf{w})\right\}, \quad (2)$$

with $\mathbf{X}[n] \equiv [\mathbf{x}[n],\ldots,\mathbf{x}[1]]^T$, where pre-windowing is applied. Where the diagonal exponential weighting matrix $\mathbf{\Lambda} \equiv \sigma^{-2}\,\mathrm{diag}([1,\lambda,\lambda^2,\ldots,\lambda^n])$ with the forgetting factor $0 << \lambda < 1$ was introduced, which can be interpreted as a flattening of the likelihood function (2) with covariance $\mathbf{\Lambda}^{-1}$ for samples z_k which lie further in the past. For simplicity it is assumed that the noise variance σ^2 is known. The conclusion in section 4 comments on the situation where σ^2 is not known. Similar to the method of weight decay in regularized neural networks the prior over the weights is taken to be

$$p(\mathbf{w}|\mathbf{A}) = (2\pi)^{-\frac{M}{2}}|\mathbf{A}|^{\frac{1}{2}}\exp(-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}), \quad (3)$$

where $\mathbf{A}$ is assumed to be a diagonal matrix in the sequel. Each diagonal element A_{kk} describes the a priori estimate of the inverse width of the Gaussian distribution of the values of the weight w_k. The posterior for $\mathbf{w}$ given the data $\mathbf{z}$, known prior distribution parameter $\mathbf{A}$ and known noise variance σ^2 reads

$$p(\mathbf{w}|\mathbf{z},\mathbf{A}) = \frac{p(\mathbf{z}|\mathbf{w})p(\mathbf{w}|\mathbf{A})}{\int p(\mathbf{z}|\mathbf{w})p(\mathbf{w}|\mathbf{A})\mathrm{d}\mathbf{w}}. \quad (4)$$

The maximum a posteriori (MAP) estimate of $\mathbf{w}$ gets $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\{-\log p(\mathbf{w}|\mathbf{z},\mathbf{A})\}$, which is identical to the maximum of $p(\mathbf{z}|\mathbf{w})p(\mathbf{w}|\mathbf{A})$ because the normalizing integral of (4) is not a function of $\mathbf{w}$ anymore. Thus, the objective function to be minimized for the sampling instant n is

$$L(\mathbf{w}) = \frac{1}{2}(\mathbf{z}-\mathbf{X}\mathbf{w})^T\mathbf{\Lambda}(\mathbf{z}-\mathbf{X}\mathbf{w}) + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}, \quad (5)$$

which corresponds to a regularized linear least squares problem with regularization matrix $\mathbf{A}$. For the choice $\mathbf{A} = \alpha\mathbf{I}$, this corresponds to a Tikhonov or *uniform regularization*. This special case is discussed in section 2.6 below. For now, we continue with the more general case of a diagonal matrix $\mathbf{A}$ with nonuniform entries which allows *selective regularization* for the individual elements of the weight vector. Computing the MAP estimate $\hat{\mathbf{w}}$ by taking the derivative $\partial/\partial\mathbf{w}$ of (5) and setting it to zero yields the regularized solution

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{z} \quad (6)$$

Introducing the regularized auto-correlation matrix estimate $\tilde{\mathbf{\Phi}} \equiv \mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \mathbf{A}$ and the corresponding covariance estimate $\tilde{\mathbf{P}} \equiv \tilde{\mathbf{\Phi}}^{-1}$, the posterior (4) can be rewritten as a multivariate Gaussian

$$p(\mathbf{w}|\mathbf{z},\mathbf{A}) = (2\pi)^{-\frac{M}{2}}|\tilde{\mathbf{P}}|^{-\frac{1}{2}}\cdot \exp\left\{-\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^T\tilde{\mathbf{P}}^{-1}(\mathbf{w}-\hat{\mathbf{w}})\right\}$$

with mean value $\hat{\mathbf{w}} = \tilde{\mathbf{P}}\mathbf{X}\mathbf{\Lambda}\mathbf{z}$ and covariance $\tilde{\mathbf{P}}$.

2.2. Evidence procedure

The derived MAP estimator (6) is based on the assumption that the parameter $\mathbf{A}$ is fixed and known beforehand. If a hierarchical Bayesian model is considered, which treats $\mathbf{A}$ as a random variable with some prior distribution $p(\mathbf{A})$ the posterior distribution $p(\mathbf{w}|\mathbf{z})$ would be obtained by integrating out, i.e., marginalization of the distribution parameter $\mathbf{A}$. This normally results in nongaussian posterior distributions which can even be multimodal. Thus, the MAP estimate is hard to compute and is not representative in general. To overcome these problems the evidence procedure tries to estimate $\mathbf{A}$ from the data and then treats $\mathbf{A}$ as if it were a fixed distribution parameter and the posterior of (4) applies. This procedure choses $\mathbf{A}$ in order to maximize $p(\mathbf{A}|\mathbf{z})$, which is sometimes called the *evidence* for the parameter $\mathbf{A}$ given the data $\mathbf{z}$. For flat prior distributions over a logarithmic scale for the parameter $\mathbf{A}$, i.e. $p(\log(A_{kk})) = const$ with $k = 1,\ldots,M$ the maximum of $p(\mathbf{A}|\mathbf{z})$ coincides with the maximum of $p(\mathbf{z}|\mathbf{A})$, which is just the normalizing integral of (4), i.e.,

$$p(\mathbf{z}|\mathbf{A}) = \frac{|\mathbf{\Lambda}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}|\tilde{\mathbf{\Phi}}|^{\frac{1}{2}}}\cdot \exp\left\{-\frac{1}{2}(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z} - \hat{\mathbf{w}}^T\tilde{\mathbf{\Phi}}\hat{\mathbf{w}})\right\}. \quad (7)$$

To maximize the evidence (7) with respect to the regularization matrix $\mathbf{A}$ we minimize $L_e(\mathbf{A}) = -\log p(\mathbf{z}|\mathbf{A})$, thus

$$L_e(\mathbf{A}) = -\frac{1}{2}\log|\mathbf{\Lambda}| + \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{A}| + \frac{1}{2}\log|\tilde{\mathbf{\Phi}}| + \frac{1}{2}(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z} - \hat{\mathbf{w}}^T\tilde{\mathbf{\Phi}}\hat{\mathbf{w}}). \quad (8)$$

Using the identity $\frac{\partial}{\partial x}\log|\mathbf{A}(x)| = \mathrm{Tr}(\mathbf{A}^{-1}\frac{\partial\mathbf{A}(x)}{\partial x})$ the condition $\partial L_e(\mathbf{A})/\partial A_{kk} = 0$ gets

$$-\frac{1}{2}\mathrm{Tr}(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial A_{kk}}) + \frac{1}{2}\mathrm{Tr}(\tilde{\mathbf{\Phi}}^{-1}\frac{\partial\tilde{\mathbf{\Phi}}}{\partial A_{kk}}) + \frac{1}{2}\hat{w}_k^2 = 0$$

with $k = 1,\ldots,M$. Solving for A_{kk} gives

$$A_{kk} = \frac{1 - A_{kk}\tilde{\Phi}^{-1}_{kk}}{\hat{w}_k^2} = \frac{\gamma_k}{\hat{w}_k^2},$$

with the obvious definition for γ. This implicit equation for A_{kk} can be used as a reestimation formula with

$$A_{kk}^{i+1} = \frac{\gamma_k^i}{(\hat{w}_k^i)^2}, \tag{9}$$

where i denotes the reestimation index. One reestimation loop includes the following computations:

$$\begin{aligned} \hat{\mathbf{w}}^i &= (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \mathbf{A}^i)^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{z} \\ \gamma_k^i &= 1 - \mathbf{A}_{kk}^i(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \mathbf{A}^i)_{kk}^{-1} \\ A_{kk}^{i+1} &= \frac{\gamma_k^i}{(w_k^i)^2} \end{aligned} \tag{10}$$

Thus, at the current sampling instant n the reestimation (9) has to be iterated until $L_e(\mathbf{A})$ of (8) has reached a local minimum.

2.3. Recursive relations

In this section, we reintroduce the discrete-time sampling index n. The MAP estimate $\hat{\mathbf{w}}$ in (6) is subsequently derived in a recursive form. Assume for the beginning that for each sampling instant n, the regularized auto-correlation estimate $\tilde{\mathbf{\Phi}}[n]$ should have the same regularization term $\mathbf{A}[n]$, i.e. $\mathbf{A}[n] = \mathbf{A}$. Thus, in each recursion a regularization term has to be added such that in the steady state of the recursion the strength of the regularization reaches $\mathbf{A}$. For convenience, the following computations are performed on the auto-correlation estimate $\mathbf{\Phi} \equiv \sigma^2\tilde{\mathbf{\Phi}}$. The recursive expression

$$\mathbf{\Phi}[n] = \lambda\mathbf{\Phi}[n-1] + \mathbf{x}[n]\mathbf{x}[n]^T + \sigma^2\mathbf{\Delta}, \tag{11}$$

with the regularization pump term $\mathbf{\Delta} = \mathbf{A}(1-\lambda)$, in the steady state coincides with

$$\mathbf{\Phi}[n] = \sigma^2\mathbf{X}^T[n]\mathbf{\Lambda}[n]\mathbf{X}[n] + \sigma^2\mathbf{A}. \tag{12}$$

Adding the regularization term $\sigma^2\mathbf{\Delta}$ in (11) corresponds to a full-rank update and thus no fast recursive computation of $\mathbf{P}[n] \equiv \mathbf{\Phi}^{-1}[n]$ using the matrix inversion lemma can be performed. For updating only one entry of $\sigma^2\mathbf{\Delta}$ per sampling instant n, a rank-one update is sufficient [4]. Thus, by introducing the M-periodic sequence of M-dimensional pivot vectors $\mathbf{v}[n] \equiv [0,0,\ldots,1,0,\ldots,0]^T$ with the nonzero entry at position $j = 1 + n \bmod N$ one can replace (11) with

$$\mathbf{\Phi}[n] = \breve{\mathbf{\Phi}}[n] + \mathbf{x}[n]\mathbf{x}[n]^T, \tag{13}$$

and

$$\breve{\mathbf{\Phi}}[n] \equiv \lambda\mathbf{\Phi}[n-1] + \kappa[n]\mathbf{v}[n]\mathbf{v}[n]^T. \tag{14}$$

The steady state behavior of (13) and (14) is periodic with M. By averaging over one full period, this behavior can be matched with the steady state according to (12) if we select

$$\kappa[n] = \sigma^2 M A_{jj}(1-\lambda) \quad \text{with} \quad j = 1 + n \bmod M. \tag{15}$$

Introducing $\breve{\mathbf{P}}[n] \equiv \breve{\mathbf{\Phi}}^{-1}[n]$ and applying the matrix inversion lemma to (13) and (14) yields

$$\breve{\mathbf{P}}[n] = \left[\mathbf{I} - \frac{\kappa[n]\mathbf{P}[n-1]\mathbf{v}[n]\mathbf{v}[n]^T}{\lambda + \kappa[n]\mathbf{v}[n]^T\mathbf{P}[n-1]\mathbf{v}[n]}\right]\lambda^{-1}\mathbf{P}[n-1], \tag{16}$$

$$\mathbf{P}[n] = \left[\mathbf{I} - \frac{\breve{\mathbf{P}}[n]\mathbf{x}[n]\mathbf{x}[n]^T}{1 + \mathbf{x}[n]^T\breve{\mathbf{P}}[n-1]\mathbf{x}[n]}\right]\breve{\mathbf{P}}[n], \tag{17}$$

respectively. Their corresponding Kalman gain vectors are

$$\breve{\mathbf{k}}[n] = \frac{\kappa[n]\mathbf{P}[n-1]\mathbf{v}[n]}{\lambda + \kappa[n]\mathbf{v}[n]^T\mathbf{P}[n-1]\mathbf{v}[n]},$$

and

$$\mathbf{k}[n] = \frac{\breve{\mathbf{P}}[n]\mathbf{x}[n]}{1 + \mathbf{x}[n]^T\breve{\mathbf{P}}[n]\mathbf{x}[n]}.$$

The update equations become

$$\breve{\mathbf{P}}[n] = \lambda^{-1}\mathbf{P}[n-1] - \lambda^{-1}\breve{\mathbf{k}}[n]\mathbf{v}[n]^T\mathbf{P}[n-1], \tag{18}$$

$$\mathbf{P}[n] = \breve{\mathbf{P}}[n] - \mathbf{k}[n]\mathbf{x}[n]^T\breve{\mathbf{P}}[n], \tag{19}$$

and

$$\hat{\mathbf{w}}[n] = (\mathbf{I} - \kappa[n]\mathbf{v}[n]\mathbf{v}^T[n]\mathbf{P}[n])\hat{\mathbf{w}}[n-1] + \mathbf{k}[n]\xi[n], \tag{20}$$

with the a priori error $\xi[n] = z[n] - \mathbf{w}^T[n-1]\mathbf{x}[n]$. The weight update equation (20) differs from the standard RLS update equation in the term $-\kappa[n]\mathbf{v}[n]\mathbf{v}^T[n]\mathbf{P}[n]\hat{\mathbf{w}}[n-1]$, which is a leakage term due to the regularization [5]. Considering (15) and taking into account the positiveness of the elements of $\mathbf{A}$ (cf. to (3)) and the positive-definiteness of $\mathbf{P}[n]$ allows to show that all eigenvalues of $(\mathbf{I} - \kappa[n]\mathbf{v}[n]\mathbf{v}^T[n]\mathbf{P}[n])$ are bounded by 1 in magnitude and, therefore, to guarantee the stability of the difference equation (20). The updating of the regularization matrix $\mathbf{A}$ using $\kappa[n]\mathbf{v}[n]\mathbf{v}^T[n]$ together with the auto-correlation update $\mathbf{x}[n]\mathbf{x}^T[n]$ results in two rank-one updates. This concludes the treatment of the RLS algorithm with constant regularization term.

In non-recursive estimation the application of the reestimation formula (10) showed convergence in a few steps. Therefore, the following heuristic procedure is proposed to allow simultaneous reestimation of the regularization parameters A_{kk} while performing the MAP estimator recursions according to (11) with (6) in the full-rank update and (16–20) in the rank-one case, respectively. Thus, for the full-rank update the reestimation index i of (10) coincides with the sampling index n, i.e.

$$A_{kk}[n+1] = \frac{\gamma_k[n]}{(w_k[n])^2} \quad \text{for} \quad k = 1,\ldots,M.$$

For the case of the rank-one update (14) of the regularization term using the pivot vector $\mathbf{v}[n]$, each component

of the regularization term gets reestimated after M_A sampling instants. To guarantee that the algorithm settles subject to the new regularization matrix, it is reasonable to chose $M_A = mM$ with $m \in \mathbb{N}$. The adaptive regularization pump term $\kappa[n]$ becomes

$$\kappa[n] = \sigma^2 A_{jj}[n] M(1-\lambda) \quad \text{with} \quad j = 1 + n \bmod M.$$

2.4. Incorporating weight pruning

The parameter $A_{kk}[n]$ of the prior (3) governs the inverse variance of the zero-mean prior distribution for the corresponding weight $w_k[n]$. This inverse variance is estimated via the evidence procedure (9) from the observed data $\mathbf{z}[n]$. If the data $\mathbf{z}[n]$ does not show any contribution of the weight $w_k[n]$ its corresponding variance parameter $A_{kk}^{-1}[n]$ will tend to zero. Thus, the prior distribution $p(\mathbf{w}[n]|\mathbf{A}[n])$ for weight $w_k[n]$ gets highly peaked at zero. For infinite $A_{kk}[n]$ the weight $w_k[n]$ can be removed without changing the error signal. In practice, weights with $A_{kk}[n] > B$, with $B \approx 10^8$, will be removed from the model. Thus, it is necessary to reduce the dimensions of the involved matrices and vectors. To prune $w_k[n]$, for the auto-correlation estimate $\mathbf{\Phi}[n]$ this would mean to prune the k-th row and k-th column. The question turns up, if one could directly prune the covariance estimate $\mathbf{P}[n]$, such that the efficient recursive computation (17) can still be applied. For the sake of conciseness the sampling time index n is omitted again. In the following, the case where the last row and the last column of $\mathbf{\Phi}$ has to be removed, is considered. This situation was chosen because of the possibility of a more compact notation of the following matrix algebra compared to the case where another row-column pair is removed. That this does not cause a loss in generality is seen from the relation $(\mathbf{M}^T\mathbf{\Phi}\mathbf{M})^{-1} = \mathbf{M}^T\mathbf{\Phi}^{-1}\mathbf{M}$, where $\mathbf{M}$ is a permutation matrix permuting the k-th column with the M-th column. Thus, $\mathbf{\Phi}$ can be thought of being partitioned as

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}_1 & \mathbf{b} \\ \mathbf{b}^T & c \end{pmatrix} \quad \text{and} \quad \mathbf{\Phi}^{-1} = \begin{pmatrix} \mathbf{C} & \mathbf{d} \\ \mathbf{d}^T & f \end{pmatrix}, \qquad (21)$$

where in addition a partition of $\mathbf{\Phi}^{-1}$ was introduced for reasons which become obvious later. Using Gaussian block-elimination with $\mathbf{\Phi}_1$ as pivot element for inverting the matrix $\mathbf{\Phi}$ gives

$$\mathbf{\Phi}^{-1} = \begin{pmatrix} \mathbf{\Phi}_1^{-1} + \beta\mathbf{\Phi}_1^{-1}\mathbf{b}\mathbf{b}^T\mathbf{\Phi}_1^{-1} & -\beta\mathbf{\Phi}_1^{-1}\mathbf{b} \\ -\beta\mathbf{b}^T\mathbf{\Phi}_1^{-1} & \beta \end{pmatrix}, \qquad (22)$$

with $\beta \equiv (c - \mathbf{b}^T\mathbf{\Phi}_1^{-1}\mathbf{b})^{-1}$. Making the obvious identification of the matrix entries of the partition of $\mathbf{\Phi}^{-1}$ in (21) with the entries of (22), the shrunken inverse matrix $\mathbf{\Phi}_1^{-1}$ is obtained as

$$\mathbf{\Phi}_1^{-1} = \mathbf{C} - f^{-1}\mathbf{d}\mathbf{d}^T. \qquad (23)$$

The elements $\mathbf{C}$, $\mathbf{d}$ and f can be read off from $\mathbf{\Phi}^{-1}$. The numerical effect of the subtraction in (23) on the positive-definiteness of $\mathbf{\Phi}_1^{-1}$ is subject to future research.

2.5. Pseudo-code of the proposed algorithm

Table 1 shows the pseudo-code of the proposed algorithm. Function `Prune()` performs the shrinkage of the weight vector $\hat{\mathbf{w}}[n]$ and the covariance matrix estimate $\mathbf{P}[n]$ according to (23) if one or more diagonal elements of $\mathbf{A}[n]$ exceeds the threshold B.

Algorithm 1: Recursive least squares algorithm with adaptive, selective regularization and weight pruning.

Init $\mathbf{A}[0] = \alpha_0\mathbf{I}$, $\mathbf{P}[0] = \mathbf{A}^{-1}[0]$
for $n \leftarrow 1$ *to* N **do**
- $j = 1 + n \bmod M$
- Generate $\mathbf{v}[n]$ with "1" at j
- $\kappa[n] = \sigma^2 M A_{jj}[n](1-\lambda)$
- $\xi[n] = z[n] - \hat{\mathbf{w}}^T[n-1]\mathbf{x}[n]$
- $\breve{\mathbf{k}}[n] = \frac{\kappa[n]\mathbf{P}[n-1]\mathbf{v}[n]}{\lambda + \kappa[n]\mathbf{v}[n]^T\mathbf{P}[n-1]\mathbf{v}[n]}$
- $\mathbf{k}[n] = \frac{\breve{\mathbf{P}}[n]\mathbf{x}[n]}{1+\mathbf{x}[n]^T\breve{\mathbf{P}}[n]\mathbf{x}[n]}$
- $\breve{\mathbf{P}}[n] = \lambda^{-1}\mathbf{P}[n-1] - \lambda^{-1}\breve{\mathbf{k}}[n]\mathbf{v}[n]^T\mathbf{P}[n-1]$
- $\mathbf{P}[n] = \breve{\mathbf{P}}[n] - \mathbf{k}[n]\mathbf{x}[n]^T\breve{\mathbf{P}}[n]$
- $\hat{\mathbf{w}}[n] =$
- $(\mathbf{I} - \kappa[n]\mathbf{v}[n]\mathbf{v}^T[n]\mathbf{P}[n])\hat{\mathbf{w}}[n-1] + \mathbf{k}[n]\xi[n]$
- **if** $n \bmod M_A = 0$ & $n > N_t$ **then**
  - $A_{kk}[n+1] =$
  - $(1 - \sigma^2 A_{kk}[n]P_{kk}[n])(\hat{w}_k[n])^{-2} \quad \forall k$
  - $(\hat{\mathbf{w}}[n], \mathbf{P}[n])$ = `Prune` $(\hat{\mathbf{w}}[n], \mathbf{P}[n], \mathbf{A}[n])$

2.6. Adaptive uniform regularization

For the choice $\mathbf{A} = \alpha\mathbf{I}$ no selective regularization for each weight or tap input is possible. If the derivative $\partial/\partial\alpha$ of the log-evidence (8) is set to zero, the reestimation formula for the regularization parameter α simplifies to

$$\alpha^{i+1} = \frac{M - \alpha^i \operatorname{Tr}(\tilde{\mathbf{P}}^i)}{\hat{\mathbf{w}}^{iT}\hat{\mathbf{w}}^i},$$

with

$$\tilde{\mathbf{P}}^i = (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \alpha^i\mathbf{I})^{-1}.$$

The remaining equations of this uniform regularization RLS algorithm are identical to those of the selective regularization algorithm presented in section 2.3.

3. SIMULATION RESULTS

For illustration of the regularization and pruning performance of the proposed scheme, a $M = 64$ taps sparse impulse response $\mathbf{w}_r$ of a mobile radio channel, shown in Fig. 1, is taken as a reference. The input signal is chosen to be a discrete multi-tone (DMT) signal with $N_c = 20$

carriers. As performance index for the adaptation quality the normalized squared norm of the misalign vector is used

$$Q[n] = 10\log_{10}\left\{\frac{(\mathbf{w}_r - \hat{\mathbf{w}}[n])^T(\mathbf{w}_r - \hat{\mathbf{w}}[n])}{\mathbf{w}_r^T\mathbf{w}_r}\right\}. \quad (24)$$

It is clear that a DMT signal comprising only $N_c = 20$ carrier is not capable to persistently excite a linear $M = 64$ taps filter. Thus, without pruning or regularization the covariance matrix estimate $\mathbf{P}[n]$ will eventually blow up. In Fig. 2 the performance index $Q[n]$ for the standard RLS and for the proposed RLS algorithm is depicted. Due to the pruning of the irrelevant weights a lower misalignment error can be reached and the auto-correlation matrix estimate does not get singular. The condition number of the auto-correlation matrix estimate for the proposed algorithm and the standard RLS are shown Fig. 3, which illustrates the ill-posedness of the standard RLS estimation. The results in Fig. 2 and Fig. 3 are averaged over 100 different realization of the additive noise $\epsilon[n]$, where a SNR of 45 dB is chosen. The algorithm setting $\lambda = 1 - 1/3M$, α_0 =1E-5 and $m = 4$ is used. In Fig. 1, in addition to the estimated and reference impulse responses, the pruned filter weights are indicated as well (the two responses actually coincide in the graphics).

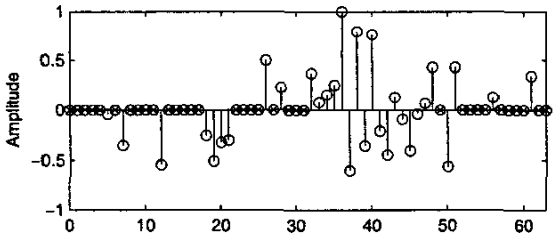

Fig. 1. Sparse impulse response of a mobile radio channel, reference $\mathbf{w}_r$ (circle) and estimation $\hat{\mathbf{w}}$ (circle), pruned filter weights (cross).

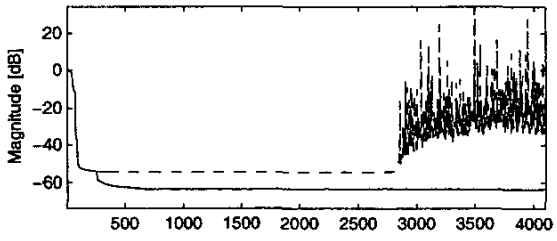

Fig. 2. Squared norm of the misalignment vector $Q[n]$ (cf. to (24)) for each sampling instant n; Standard RLS algorithm (dashed) and proposed algorithm (solid).

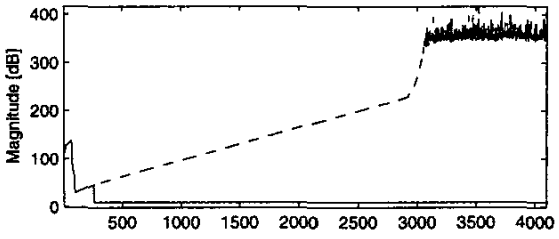

Fig. 3. Condition number of the auto-correlation matrix estimate for standard RLS (dashed) and for the proposed algorithm (solid).

4. CONCLUSION

The evidence procedure from Bayesian estimation is applied to the regularization of the RLS algorithm. Due to the use of a regularization matrix it is possible to distinguish the relevant model weights from the irrelevant ones. The regularization matrix is successively updated using a rank-one update. Thus the computational complexity of the proposed algorithm stays at $\mathcal{O}(M^2)$ when no model parameters can be pruned. In the case of k irrelevant weights the complexity decreases to $\mathcal{O}((M-k)^2)$, which can be much lower than for the standard $\mathcal{O}(M^2)$ RLS. It is also possible to estimate the noise variance σ^2 via the evidence maximization, in analogy to the parameter $\mathbf{A}$. As a negative point, in several situations the proposed algorithm exhibits stability problems if the regularization matrix update was performed too fast. To stabilize this updating is subject to future research.

5. REFERENCES

[1] L. Ljung and T. Söderstrom, *Theory and Practice of Recursive Identification*, The MIT Press, Cambridge, MA, 1983.

[2] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, June 2001.

[3] D. J. C. Mackay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[4] S. L. Gay, "Dynamically regularized fast RLS with application to echo cancellation," in *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing.*, Atlanta, GA, USA, May 1996, vol. 2, pp. 957–960.

[5] L. Ljung and J. Sjöberg, "A comment on leakage in adaptive algorithms," in *4th IFAC International Symposium on Adaptive Systems in Control and Signal Processing*, Grenoble, France, Jul 1992, pp. 377–382.