Improved l_0-RLS adaptive filter

B.K. Das and M. Chakraborty✉

In this Letter, the authors present an improved sparse recursive least squares (RLS) algorithm, which employs a novel approximation of the l_0 norm of the filter coefficient vector for regularising the RLS cost function. The proposed algorithm achieves improved performance over existing algorithms as demonstrated via numerical simulations.

Introduction: It is well known that the most ideal way of deriving sparsity aware adaptive filters is to add the l_0 norm penalty of the filter coefficient vector $\boldsymbol{w}(n)$ to the cost function of standard adaptive filters (e.g. least mean square (LMS), recursive least squares (RLS) etc.), and minimise the sum by differentiating it w.r.t. $\boldsymbol{w}(n)$ and setting it to zero. However, the l_0 norm, being just a count of non-zero entries of a vector, is not differentiable (neither it is a true norm) and thus, the above optimisation requires combinatorial search that is NP hard. A popular approach to overcome this difficulty has been to approximate the l_0 norm by other norms like the l_1 norm $\|\boldsymbol{w}(n)\|_1$ which are differentiable almost everywhere (a.e.). Recently, it has also been shown [1, 2] that the sparsity promoting function $\sum_{i=0}^{N-1}[1-\exp(-\alpha|w_i(n)|)]$ is a better approximation of the l_0 norm of $\boldsymbol{w}(n)$ than the l_1 norm (where $w_i(n)$ denotes the ith coefficient of $w(n)_{N\times 1}$ and α is a suitable constant). An RLS-based adaptive filter has been derived in [1] that adds the above function as a sparsity promoting penalty to the RLS cost function. Unfortunately, first-order differentiation of this function w.r.t. $\boldsymbol{w}(n)$ results in expressions given in terms of sgn$[\boldsymbol{w}(n)]$ (where the operator sgn$[\cdot]$ denotes element wise 'sign' or 'signum' function), meaning minimisation of the cost function leads to equations that are non-linear. Due to this, a closed form least squares (LS) solution of $\boldsymbol{w}(n)$ and its exact RLS update become very hard to determine. To circumvent this, the term sgn$[\boldsymbol{w}(n)]$ is replaced by sgn$[\boldsymbol{w}(n-1)]$ in [1], which, however, results in the introduction of substantial gradient noise especially for 'inactive' taps (i.e. the taps having zero or close to zero optimum values), as, in the steady state, estimates of such taps fluctuate around zero. In order to counter this, we follow the philosophy of [3] and propose a new approximation of the l_0 norm that results in exact LS solution of $\boldsymbol{w}(n)$ and its RLS update without giving rise to terms like sgn$[\boldsymbol{w}(n)]$.

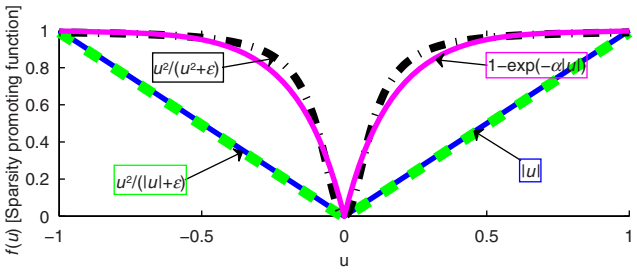

Fig. 1 *Some sparsity promoting functions*

New non-convex approximation of l_0 norm: Given a vector $\boldsymbol{a}\in\mathbb{R}^N$, we propose a new approximation of its l_0 norm by a non-convex, a.e. differentiable function $\sum_{i=0}^{N-1}(a_i^2)/(a_i^2+\epsilon)$, where ϵ is a small, positive constant. In Fig. 1, we plot the univariate version of this with ϵ taken as 1×10^{-4} and also of the aforesaid l_0 norm approximation $\sum_{i=0}^{N-1}[1-\exp(-\alpha|a_i|)]$ with $\alpha=50$, by dashed 'black' and 'magenta', respectively, as a function of a single variable u. It is seen that both curves are very close to each other, and both are much better approximations of the l_0 norm than the l_1 norm (i.e. $|u|$, as shown in 'blue'). Using the proposed approximation, the l_0 norm regularised LS cost function is given as

$$J(\boldsymbol{w}(n)) = \sum_{k=1}^{n}\lambda^{n-k}[d(k)-\boldsymbol{x}^{\mathrm{T}}(k)\boldsymbol{w}(n)]^2+\frac{\rho}{2}\sum_{i=0}^{N-1}\frac{w_i^2(n)}{w_i^2(n)+\epsilon}$$
$$\simeq \sum_{k=1}^{n}\lambda^{n-k}[d(k)-\boldsymbol{x}^{\mathrm{T}}(k)\boldsymbol{w}(n)]^2+\frac{\rho}{2}\sum_{i=0}^{N-1}\frac{w_i^2(n)}{w_i^2(n-1)+\epsilon}, \quad (1)$$

where $d(n)=\boldsymbol{x}^{\mathrm{T}}(n)\boldsymbol{w}_0+v(n)$ is the observable system output, $\boldsymbol{x}(n)=[x(n)x(n-1)\cdots x(n-N+1)]^{\mathrm{T}}$, $x(n)$ is the system input, $v(n)$ is an additive observation noise, and $\boldsymbol{w}_0$ is the system impulse response vector. Note that in the denominator of the second term in (1), $w_i(n)$ is approximated by $w_i(n-1)$ which is treated as a known constant. It is easy to check that as long as $w_i(n)$ and $w_i(n-1)$ remain close to each other, $\sum_{i=0}^{N-1}w_i^2(n)/(w_i^2(n-1)+\epsilon)$ still remains a good approximation of $||\boldsymbol{w}(n)||_0$. The above approximation, however, ensures that on derivation by $w_i(n)$, one obtains a first-order expression in $w_i(n)$, similar to the first term on the RHS of (1), which is the standard LS cost function and, whose derivative w.r.t. $\boldsymbol{w}(n)$ is given by

$$\big(\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{X}(n)\big)\boldsymbol{w}(n)-\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{d}(n),$$

where $\boldsymbol{X}(n)=\big[\boldsymbol{X}^{\mathrm{T}}(n-1)\quad \boldsymbol{x}(n)\big]^{\mathrm{T}}$, $\boldsymbol{X}(1)=\boldsymbol{x}^{\mathrm{T}}(1)=[x(1),0,0,\ldots,0]$, $\boldsymbol{d}(n)=[d(1),d(2),\ldots,d(n)]^{\mathrm{T}}$ and $\boldsymbol{\Lambda}(n)=\mathrm{diag}\{\lambda^{n-1},\cdots,\lambda^2,\lambda,1\}$ with λ $(0<\lambda<1)$ being a forgetting factor.

The second term on the RHS of (1), which is the sparsity promoting term, on the other hand, produces $\rho D_0(n)\boldsymbol{w}(n)$ when differentiated w.r.t. $\boldsymbol{w}(n)$, where $\boldsymbol{D}_0(n)=\mathrm{diag}\{d_{0,0}(n),d_{0,1}(n),\ldots,d_{0,N-1}(n)\}$, and $d_{0,i}(n)=1/(w_i^2(n-1)+\epsilon)$. Combining, the optimal LS solution $\boldsymbol{w}(n)$ for (1) is given by

$$\boldsymbol{w}(n)=\arg\min_{\boldsymbol{w}(n)}J(\boldsymbol{w}(n))=\boldsymbol{P}(n)\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{d}(n), \quad (2)$$

where $\boldsymbol{P}(n)=\big(\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{X}(n)+\rho\boldsymbol{D}_0(n)\big)^{-1}$.

To obtain the RLS update relation from (2), we follow the approach of [3]. In [3], an l_1 norm regularised RLS algorithm (termed as ZA-RLS-II) was derived which approximates the l_1 norm of a given $\boldsymbol{a}\in\mathbb{R}^N$ by $\sum_{i=0}^{N-1}(a_i^2)/(|a_i|+\epsilon)$ (shown by the dashed 'green' curve in Fig. 1 for the univariate case, for $\epsilon=1\times10^{-4}$). For this, we first note that

$$\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{d}(n)=\lambda\boldsymbol{X}^{\mathrm{T}}(n-1)\boldsymbol{\Lambda}(n-1)\boldsymbol{d}(n-1)+\boldsymbol{x}(n)d(n). \quad (3)$$

Next, we write

$$\begin{aligned}\boldsymbol{P}^{-1}(n)&=\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{X}(n)+\rho\boldsymbol{D}_0(n)\\&=\lambda\boldsymbol{X}^{\mathrm{T}}(n-1)\boldsymbol{\Lambda}(n-1)\boldsymbol{X}(n-1)+\boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n)+\rho\boldsymbol{D}_0(n)\\&=\lambda\boldsymbol{P}^{-1}(n-1)+\boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n)+\rho(\boldsymbol{D}_0(n)-\lambda\boldsymbol{D}_0(n-1)).\end{aligned}$$

The above implies

$$\begin{aligned}\lambda\boldsymbol{P}(n)\boldsymbol{P}^{-1}(n-1)=\boldsymbol{I}&-\boldsymbol{P}(n)\boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n)\\&-\rho\boldsymbol{P}(n)(\boldsymbol{D}_0(n)-\lambda\boldsymbol{D}_0(n-1))\end{aligned} \quad (4)$$

From (3) and (4), the time-recursive version of (2) is then obtained as

$$\begin{aligned}\boldsymbol{w}(n)&=\boldsymbol{P}(n)\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{d}(n)\\&=\boldsymbol{P}(n)[\lambda\boldsymbol{X}^{\mathrm{T}}(n-1)\boldsymbol{\Lambda}(n-1)\boldsymbol{d}(n-1)+\boldsymbol{x}(n)d(n)]\\&=\boldsymbol{P}(n)[\lambda\boldsymbol{P}^{-1}(n-1)\boldsymbol{w}(n-1)+\boldsymbol{x}(n)d(n)]\\&=[\boldsymbol{I}-\boldsymbol{P}(n)\boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n)]\boldsymbol{w}(n-1)\\&\quad-\rho\boldsymbol{P}(n)(\boldsymbol{D}_0(n)-\lambda\boldsymbol{D}_0(n-1))\boldsymbol{w}(n-1)+\boldsymbol{P}(n)\boldsymbol{x}(n)d(n)\\&=\big[\boldsymbol{I}-\rho\boldsymbol{P}(n)\big(\boldsymbol{D}_0(n)-\lambda\boldsymbol{D}_0(n-1)\big)\big]\boldsymbol{w}(n-1)\\&\quad+\boldsymbol{P}(n)\boldsymbol{x}(n)e_p(n),\end{aligned} \quad (5)$$

where $e_p(n)=d(n)-\boldsymbol{w}^{\mathrm{T}}(n-1)\boldsymbol{x}(n)$ is the so-called a priori error.

Next we try to evaluate $\boldsymbol{P}(n)$, which we write as $[Q^{-1}(n)+\rho\boldsymbol{D}_0(n)]^{-1}$, where $\boldsymbol{Q}(n)=\big(\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{\Lambda}(n)\boldsymbol{X}(n)\big)^{-1}$. To compute the inverse, we invoke the Woodbury matrix inversion lemma [4] which states that given $\boldsymbol{A}$ and $\boldsymbol{C}$ to be two $l\times l$ and $r\times r$ invertible matrices, and $\boldsymbol{U}$ and $\boldsymbol{V}$ be two $l\times r$ and $r\times l$ matrices, respectively, the following holds:

$$(\boldsymbol{A}+\boldsymbol{UCV})^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{U}\big(\boldsymbol{C}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U}\big)^{-1}\boldsymbol{VA}^{-1}. \quad (6)$$

Taking $\boldsymbol{A}=\rho\boldsymbol{D}_0(n)$, $\boldsymbol{C}=\boldsymbol{Q}^{-1}(n)$, and $\boldsymbol{U}$ and $\boldsymbol{V}$ both to be identity matrices, we have

$$\begin{aligned}\boldsymbol{P}(n)&=\big(\boldsymbol{Q}^{-1}(n)+\rho\boldsymbol{D}_0(n)\big)^{-1}\\&=\frac{1}{\rho}\boldsymbol{D}_0^{-1}(n)-\frac{1}{\rho}\boldsymbol{D}_0^{-1}(n)[\boldsymbol{Q}(n)+\frac{1}{\rho}\boldsymbol{D}_0^{-1}(n)]^{-1}\frac{1}{\rho}\boldsymbol{D}_0^{-1}(n).\end{aligned}$$

Simplifying

$$\boldsymbol{P}(n)=\Big[\boldsymbol{I}-\boldsymbol{D}_{0,\mathrm{inv}}(n)\big\{\boldsymbol{Q}(n)+\boldsymbol{D}_{0,\mathrm{inv}}(n)\big\}^{-1}\Big]\boldsymbol{D}_{0,\mathrm{inv}}(n), \quad (7)$$

where $\boldsymbol{D}_{0,\mathrm{inv}}(n)=(\rho\boldsymbol{D}_0(n))^{-1}=(1/\rho)\boldsymbol{D}_0^{-1}(n)$.

It is seen from (7) that we still need a matrix inversion, i.e. $\{\boldsymbol{Q}(n)+\boldsymbol{D}_{0,\text{inv}}(n)\}^{-1}$ to evaluate $\boldsymbol{P}(n)$. Similar to [3], in the sequel, we show that this matrix inversion can be performed alternatively using an iterative update rule exploiting the structures of the individual matrices. First recall that $\boldsymbol{D}_{0,\text{inv}}(n)$ is a diagonal matrix with the ith diagonal element given by $(1/\rho)[w_i^2(n-1)+\epsilon]$. For large n, it is reasonable to expect that $\boldsymbol{D}_{0,inv}(n)$ will have approximately $r=\|\boldsymbol{w}_0\|_0$ number of significant diagonal entries. Using a threshold δ (which is a very small positive constant), we retain only those diagonal elements $[\boldsymbol{D}_{0,\text{inv}}(n)]_{i,i}$ for which $|w_i(n)|>\delta$. Define the index set $S=\{0,1,\ldots,N-1\}$ and the subset $NZ\subset S$ such that for every $i\in NZ$, $|w_i(n)|>\delta$ and for every $j\in S\backslash NZ$, $|w_i(n)|\leq\delta$. Let $|NZ|=l$, which, for large n, will be close to $r=\|w_0\|_0$. We arrange the indices belonging to NZ in a $l\times 1$ vector $\boldsymbol{\theta}_n$, i.e. for every $i\in\{1,2,\ldots,l\}$, $\theta_n(i)\in NZ$ and if $i\neq j$, $\theta_n(i)\neq\theta_n(j)$. Then $\boldsymbol{D}_{0,\text{inv}}(n)$ can be approximated as $\boldsymbol{D}_{0,\text{inv}}(n)\simeq\sum_{i=1}^{l}\boldsymbol{D}^i_{0,\text{inv}}(n)$, where $\boldsymbol{D}^i_{0,\text{inv}}(n)$ is a rank one diagonal matrix with a non-zero entry at the $\theta_n(i)$th diagonal position. Defining $\boldsymbol{Q}_j(n)=\boldsymbol{Q}(n)+\sum_{i=1}^{j}\boldsymbol{D}^i_{0,\text{inv}}(n)$ for $j=1,\ldots,l$, and also, $\boldsymbol{Q}_0(n)=\boldsymbol{Q}(n)$, the task then boils down to computing the inverse of $\boldsymbol{Q}_l(n)$. We do this job recursively. Suppose, the inverse of $\boldsymbol{Q}_j(n)$ has been computed for some $j\in\{0,1,\ldots,l-1\}$. Then, to compute the inverse of $\boldsymbol{Q}_{j+1}(n)=\boldsymbol{Q}_j(n)+D^{j+1}_{0,\text{inv}}(n)$, we again use the aforesaid matrix inversion lemma [4] for which we choose $\boldsymbol{A}=\boldsymbol{Q}_j(n)$, $C=1$, $\boldsymbol{u}$: an $N\times 1$ vector with all elements zero except having $+\sqrt{(1/\rho)(w^2_{\theta_n(j+1)}(n-1)+\epsilon)}\simeq\sqrt{(1/\rho)}w_{\theta_n(j+1)}(n-1)$ at the $\theta_n(j+1)$th position and $\boldsymbol{v}=\boldsymbol{u}^{\text{T}}$. (Here, the scalar C and the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are equivalent to the matrices $\boldsymbol{C}$, $\boldsymbol{U}$, and $\boldsymbol{V}$, respectively, of (6).)

Then, we obtain from the aforementioned matrix inversion lemma

$$\begin{aligned}\boldsymbol{Q}^{-1}_{j+1}(n)&=[\boldsymbol{I}-\boldsymbol{Q}^{-1}_j(n)\boldsymbol{u}[1+\boldsymbol{u}^{\text{T}}\boldsymbol{Q}^{-1}_j(n)\boldsymbol{u}]^{-1}\boldsymbol{u}^{\text{T}}]\boldsymbol{Q}^{-1}_j(n)\\&=[\boldsymbol{I}-\frac{1}{1+\text{Tr}(\boldsymbol{D}^{j+1}_{0,\text{inv}}(n)\boldsymbol{Q}^{-1}_j(n))}\boldsymbol{Q}^{-1}_j(n)\boldsymbol{D}^{j+1}_{0,\text{inv}}(n)]\boldsymbol{Q}^{-1}_j(n)\\&=\boldsymbol{Q}^{-1}_j(n)-\frac{w^2_{\theta_n(j+1)}(n-1)}{\rho+w^2_{\theta_n(j+1)}(n-1)q_{j+1}(n)}\boldsymbol{q}_{j+1}(n)\boldsymbol{q}^{\text{T}}_{j+1}(n),\end{aligned}$$

for $j=0,1,2,\ldots,l-1$, where $\boldsymbol{q}_{j+1}(n)$ and $q_{j+1}(n)$ are the $\theta_n(j+1)$th column and $\theta_n(j+1)$th diagonal element of the matrix $\boldsymbol{Q}^{-1}_j(n)$, respectively.

For highly sparse $\boldsymbol{w}_0$, $l\ll N$. Thus, the number of iterations required to compute $\boldsymbol{Q}^{-1}_l(n)$ can be kept considerably small.

Finally, $\boldsymbol{Q}^{-1}(n)$ is updated time recursively as

$$\boldsymbol{Q}^{-1}(n)=\lambda\boldsymbol{Q}^{-1}(n-1)+\boldsymbol{x}(n)\boldsymbol{x}^{\text{T}}(n). \tag{8}$$

Numerical simulations: We conduct an adaptive system identification experiment using zero mean, unit variance white Gaussian random process as input and a sparse $\boldsymbol{w}_0$ as the system to be identified. The system $\boldsymbol{w}_0$ has 128 coefficients, out of which four coefficients with random locations have value 1 each, while rest of the coefficients have zero value. The additive measurement noise is taken to be a zero mean white Gaussian noise with variance 0.01, giving rise to an SNR of 20 dB. Since ZA-RLS-II [3] is known to outperform the other contemporary sparse RLS algorithms (e.g. OCCD-TNWL [5], SPARLS [6] etc.), and behaves identical to ZA-RLS-I [3], we compare the performance of the proposed algorithm with that of ZA-RLS-II [3]. Apart from ZA-RLS-II, the l_0 norm regularised RLS [1], the standard sparsity unaware RLS, and the l_0-normalised least mean square (NLMS) [2] are also deployed for identifying $\boldsymbol{w}_0$, and their convergence behaviours are compared with that of the proposed one. The forgetting factor for all the RLS-based filters is chosen as $\lambda=0.95$. The other parameters for the ZA-RLS-II and the proposed algorithm are chosen as follows: $\rho=0.0005$, $\epsilon=0.0001$. The parameters for l_0-RLS [1] are taken to be $\beta=50$, $\kappa=2$. In the case of l_0-NLMS, parameters are tuned to obtain minimum steady-state mean-square deviation (MSD). We carry out the experiment for 375 iterations and the results are averaged over 100 ensembles. Fig. 2 shows the learning curves (i.e. MSD vs. iteration index) for (i) the proposed algorithm, (ii) ZA-RLS-II [3], (iii) the sparsity unaware RLS, (iv) the l_0-RLS [1], and (v) the l_0-NLMS [2] (plotted by dashed 'black', dashed 'green', 'blue', 'magenta', and 'red', respectively). It is seen that the proposed algorithm outperforms all the existing algorithms including the recently proposed ZA-RLS-II [3] (dashed 'green') in terms of steady-state MSD. The convergence speed of the proposed algorithm is also substantially higher than that of the l_0-RLS [1], the l_0-NLMS [2], and the RLS, and is at par with that of the ZA-RLS-II [3].

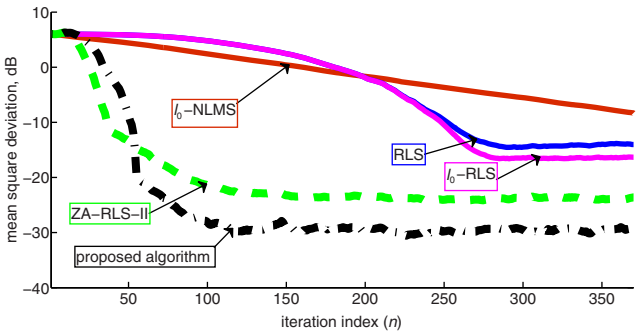

Fig. 2 *MSD versus iteration index n for different sparse adaptive filters*

Conclusion: In this Letter, we have proposed a new non-convex and a. e. differentiable approximation of the sparsity promoting l_0 norm which can be expressed as a weighted l_2 norm, and have employed it to derive a new RLS-type algorithm which is significantly superior to the existing sparse RLS algorithms including the newly proposed ZA-RLS-II [3] which was derived using a similar weighted l_2 norm-based approximation of the convex l_1 norm. Numerical simulations demonstrated that our proposed algorithm outperforms the existing algorithms including that of [3].

Submitted: *11 September 2017* E-first: *7 November 2017*
doi: 10.1049/el.2017.3441
One or more of the Figures in this Letter are available in colour online.

B.K. Das (*Department of E&ECE, IIT Kharagpur, West Bengal 721302, India*)

References

1 Eksioglu, E.M., and Tanc, A.K.: 'RLS algorithm with convex regularization', *Signal Process. Lett.*, 2011, **18**, (8), pp. 470–473
2 Su, G., Jin, J., Gu, Y., *et al.*: 'Performance analysis of l_0 norm constraint least mean square algorithm', *Trans. Signal Proc.*, 2012, **60**, (5), pp. 2223–2235
3 Hong, X., Gao, J., and Chen, S.: 'Zero attracting recursive least squares algorithms', *Trans. Veh. Technol.*, 2017, **66**, (1), pp. 213–221
4 Horn, R., and Johnson, C.R.: 'Matrix analysis' (Cambridge University, Cambridge, 1985)
5 Angelosante, D., Bazerque, J. A., and Giannakis, G. B.: 'Online adaptive estimation of sparse signals: where RLS meets the l_1-norm', *Trans. Signal Proc.*, 2010, **58**, (7), pp. 3436–3447
6 Babadi, B., Kalouptsidis, N., and Tarokh, V.: 'SPARLS: the sparse RLS algorithm', *Trans. Signal Proc.*, 2010, **58**, (8), pp. 4013–4025