Performance Analysis of l_0 Norm Constrained Recursive Least Squares Algorithm

Samrat Mukhopadhyay, *Student Member, IEEE*, Bijit Kumar Das, *Student Member, IEEE*, and Mrityunjoy Chakraborty, *Senior Member, IEEE*

Abstract

Performance analysis of l_0 norm constrained Recursive least Squares (RLS) algorithm is attempted in this paper. Though the performance pretty attractive compared to its various alternatives, no thorough study of theoretical analysis has been performed. Like the popular l_0 Least Mean Squares (LMS) algorithm, in l_0 RLS, a l_0 norm penalty is added to provide zero tap attractions on the instantaneous filter taps. A thorough theoretical performance analysis has been conducted in this paper with white Gaussian input data under assumptions suitable for many practical scenarios. An expression for steady state MSD is derived and analyzed for variations of different sets of predefined variables. Also a Taylor series expansion based approximate linear evolution of the instantaneous MSD has been performed. Finally numerical simulations are carried out to corroborate the theoretical analysis and are shown to match well for a wide range of parameters.

Index Terms

Adaptive filters, sparsity, l_0 norm, Recursive Least Squares (RLS) algorithm, mean square deviation, performance analysis.

I. Introduction

Sparse systems are frequently encountered in many applications, such as echo paths [1], wireless communication channels, HDTV [2] etc. A system vector is called sparse if it has a very small number of nonzero entries compared to its dimension. It becomes necessary then to find identification algorithms suitable for such sparse systems. Adaptive algorithms are frequently used to identify systems whose parameters are changing with time. Due to its simplicity and ease of implementation, the least mean squares (LMS) algorithm [3] has enjoyed much success for a long time. Another frequently used adaptive algorithm is the recursive least squares (RLS) [4] which recursively tries to minimize the error between estimated and unknown system vectors using the information conveyed by the data from the beginning of reception. But such algorithms are sparsity agnostic and generally do not perform well when the unknown system is sparse. Inspired by the introduction of sparse signal processing and the nascent field of Compressive sensing (CS) [5]–[7], the last decade saw a flurry of activities on sparse adaptive filters, that has produced a number of several new algorithms that exploit the knowledge of sparsity [8]. Many of these algorithms use the knowledge of sparsity of the unknown system vector to add an l_p norm penalty to the cost function. ZA-LMS [9] uses l_1 norm penalty and l_0 LMS [10] uses l_0 norm penalty to exert zero attraction on the filter taps. l_1 norm regularized RLS algorithms have also been proposed by researchers. The SPARLS [11] algorithm suggests the use of Expectation-Maximization(EM) algorithm to minimize the l_1 norm penalized RLS cost function. The authors of [12] propose an algorithm that uses an online coordinate descent algorithm together with the l_1 regularized RLS cost function. The l_1 RLS algorithm [13] has been proposed where the cost function of conventional RLS algorithm has been modified by adding a l_1 penalty term which results in a zero point attracted RLS algorithm. In [14] a general convex penalty term is added to the RLS cost function to result in a sparsity aware convex regularized RLS algorithm.

Among the different penalty terms that can be used as a regularizer of the cost function of RLS in [14], of particular interests are the convex functions that can be used to approximate l_0 penalty term, as it was introduced in [10]. Since the l_0 norm penalty can introduce strong zero point attraction to the small taps of the estimated parameter at each step of the algorithm, for a sparse system the algorithm is expected to converge faster to a lower steady state mean square deviation. Though the author of [14] has numerically shown that mean square deviation performance of l_0 norm penalized RLS is superior to the conventional RLS, neither he or anyone else, to the best of our knowledge, has been found to make an attempt to establish that claim through a theoretical analysis of the algorithm. A detailed theoretical analysis of such an algorithm could not only just corroborate the superior performance promised by the numerical simulations of l_0 RLS but also can find out the spectrum, of the different set of predefined variables, over which the algorithm may even become worse than the conventional algorithm. A detailed theoretical analysis of l_0 LMS was carried out in [15] which inspired the present work. The present work is aimed at providing a thorough analysis of the l_0 RLS algorithm along with presenting the salient features and limitations of the performance of this algorithm.

Samrat Mukhopadhyay is with Dept. of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur 721302, India (Email: samratphysics@gmail.com)

Bijit Kumar Das is with Dept. of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur 721302, India (Email: bijitbijit@gmail.com)

Mrityunjoy Chakraborty is with Dept. of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur 721302, India (Email: mrityun@ece.iitkgp.ernet.in)

II. PRELIMINARIES

Let the system has the unknown parameter vector $\mathbf{s} = \begin{bmatrix} s_0, s_1, \cdots s_{N-1} \end{bmatrix}^T \in \mathbb{R}^N$ and let the input vector at time n be denoted by $\mathbf{x}_n = \begin{bmatrix} x(n), x(n-1), \cdots x(n-N+1) \end{bmatrix}^T \in \mathbb{R}^N$. The system produces output sequence $\{y(n)\}$ where

$$y_n = \mathbf{s}^T \mathbf{x}_n + \nu_n$$

where $\{\nu_n\}$ is an additive noise sequence. Let, the adaptive filter produces an estimate $\mathbf{w}_n = \begin{bmatrix} w_{0,n}, w_{1,n}, \cdots w_{N-1,n} \end{bmatrix}$ for the system tap vector, at time n. The instantaneous estimation error between the output of the unknown system and the output of the adaptive filter is

$$e_n = y_n - \mathbf{w}_n^T \mathbf{x}_n = (\mathbf{s} - \mathbf{w}_n)^T \mathbf{x}_n + \nu_n$$

The cost function of the conventional RLS adaptive filter with forgetting factor λ is defined as

$$\mathcal{E}_n = \sum_{m=0}^{n} \lambda^{n-m} (e_m)^2$$

In order take into account the sparsity of the unknown system vector $\mathbf{s}$, l_0-RLS modifies the cost function at each iteration by adding to it a penalty function that gives a measure of the sparsity of the system. l_0-RLS chooses the l_0 ''norm'' as the penalty function. As a result, the cost becomes,

$$\mathcal{E}_n = \sum_{m=0}^{n} \lambda^{n-m} (e_m)^2 + \gamma \|\mathbf{w}_n\|_0 \tag{1}$$

where the l_0 norm is defined as the number of non-zero entries of a vector and the parameter γ is a penalty factor that controls the balance between estimation error and penalty. In general, the l_0 norm optimization problem is known to be NP hard [7] and because of that it is often approximated by continuous(often convex) functions. A popular approximation was introduced in [10] which results, after some manipulations, in the following evolution equation of the l_0 RLS adaptive filter [14]

$$\mathbf{w}_n = w_{n-1} + \mathbf{k}_n \xi_n + \kappa \mathbf{P}_n \mathbf{g}(\mathbf{w}_{n-1}) \tag{2}$$

where

$$\kappa = \gamma(1 - \lambda) \tag{3}$$

$$\xi_n = y_n - \mathbf{w}_{n-1}^T \mathbf{x}_n \tag{4}$$

$$\mathbf{P}_n = (\mathbf{\Phi}_n)^{-1} = \left(\sum_{m=0}^{n} \lambda^{n-m} \mathbf{x}_m \mathbf{x}_m^T \right)^{-1} \tag{5}$$

$$\mathbf{k}_n = \frac{\mathbf{P}_{n-1} \mathbf{x}_n}{\lambda + \mathbf{x}_n^T \mathbf{P}_{n-1} \mathbf{x}_n} \tag{6}$$

and $\mathbf{g}(\mathbf{w}_{n-1}) = \begin{bmatrix} g(w_{0,n-1}), g(w_{1,n-1}), \cdots g(w_{N-1,n-1}) \end{bmatrix}^T$ where the function $g(\cdot)$ is defined as below

$$g(t) = \begin{cases} \beta^2 t - \beta \mathrm{sgn}(t), & |t| \leq 1/\beta \\ 0, & \text{elsewhere} \end{cases} \tag{7}$$

The third term in Eq. (2) is the *zero-point attraction* term [10] and the range $(-1/\beta, 1/\beta)$ is called the *attraction* range [15].

III. MODELLING AND ASSUMPTIONS

Following the approach adopted by Su et.al [15], based on the magnitudes of the entries of the unknown system vector $\mathbf{s}$, we partition the set of indices $\{1, 2, \cdots, N\}$ into three sets:

$$\mathcal{C}_0 := \{k : s_k = 0\} \tag{8}$$

$$\mathcal{C}_L := \{k : |s_k| > 1/\beta\} \tag{9}$$

$$\mathcal{C}_S := \{k : 0 < |s_k| \leq 1/\beta\} \tag{10}$$

Thus, if $\mathbf{s}$ is K-sparse, $|\mathcal{C}_L \cup \mathcal{C}_S| = K$, $|\mathcal{C}_0| = N - K$.

We adopt the following assumptions:

A.1 The data sequence $\{x(n)\}$ is a white sequence with zero mean and variance P_x and is independent of the additive noise sequence $\{\nu_n\}$ which is also assumed to be a zero mean sequence.

A.2 (*Independence assumption*) The incoming sequence of vectors $\{\mathbf{x}_n\}$ are independent.

A.3 λ is chosen *sufficiently* close to 1 such that $\frac{N}{N+1} << \lambda < 1$, so that for large n, $\mathbf{P}_n \approx \mathbb{E}(\mathbb{P}_n) = \frac{1-\lambda}{1-\lambda^{n+1}} \mathbf{R}^{-1}$ where $\mathbf{R}$ is the autocorrelation matrix of the incoming data sequence.

A.4 The parameters κ and β are chosen such that $\beta^2\kappa(1-\lambda) << P_x$.

A.5 The tap weights $w_{k,n}$, $\forall k \in \mathcal{C}_0$ are gaussian distributed.

A.6 $w_{k,n}$ are assumed to be of the same sign as that of s_k, $k \in \mathcal{C}_L \cup \mathcal{C}_S$.

A.7 $w_{k,n}$ are assumed to be out of the attraction range for $k \in \mathcal{C}_L$ and inside attraction range elsewhere.

The following points attempt to justify the use of these assumptions:

1) The assumption **A.1** is generally adopted to leverage the simple properties of a gaussian data sequence. This assumption can be slightly generalized by dropping the assumption that the sequence is independent, which forces one to work with a coloured gaussian sequence. However, a coloured sequence be easily *pre-whitened* by pre-multiplying any vector of interest with the unitary matrix that diagonalizes the covariance matrix of the gaussian sequence [4], which is why assumption **A.1** can be considered without loss of generality.

2) Assumption **A.2** is the *independence* assumption and is widely used in the literature for simplified analysis of adaptive algorithms [4], [16].

3) Assumption **A.3** has generally been used in the literature for simplified analysis of RLS [17]. One justification for this assumption can be provided by the following lemma which assumes assumptions **A.1** and **A.2**.

   Lemma 3.1. *If the sequence $\{x(n)\}$ is assumed to follow assumption A.1, then with $0 < \lambda \le 1$*

$$\lim_{n\to\infty} \mathbb{E}\left(\left\|\frac{\boldsymbol{\Phi}_n}{\frac{1-\lambda^{n+1}}{1-\lambda}} - \mathbf{R}\right\|^2\right) = \left(\frac{1-\lambda}{1+\lambda}\right)(N+1)P_x^2 \tag{11}$$

   where $\|\cdot\|$ denotes the 2-matrix norm.

   Proof: A short proof is provided in Appendix A. ∎

   Lemma 3.1 encourages the use of assumption **A.3**. Furthermore, as it will be seen in the performance analysis of l_0 RLS, this assumption simplifies the analysis significantly because without this assumption, the nonlinear contribution of past data vector $\mathbf{x}_{n-1}$, in matrix $\mathbf{P}_n$ makes carrying out the analysis difficult.

4) Assumption **A.4** is a result of experimental observation. It basically implies that for the l_0 RLS to be stable, $\kappa, \beta, 1-\lambda$ have to be small compared to the signal power.

5) The use of assumptions **A.5**, **A.6**, and **A.7** are found suitable for this analysis. These exactly same assumptions are taken in [15] for the analysis of l_0 LMS. They justifications of the assumptions there are based on intuitive discussion and logical assumptions which also were probably justified by experimental observations. In the same spirit we also performed extensive simulations to verify these assumptions. Also, since the structure of the l_0 RLS algorithm is similar to that of the l_0 LSM algorithm, save the time varying gain matrix, it is expected that the logical discussions similar to those justifying the use of these assumptions in the work of Su etal [15] can also justify the use of these assumptions in our work.

IV. Performance analysis

The convergence analysis of RLS itself is not easy because of the presence of the time dependent gain matrix $\mathbf{P}_n$. However the use of assumption **A.3** significantly simplifies the analysis [17]. We then use the assumptions taken in Section III to carry out the analysis in a simplified manner.

A. Mean convergence analysis:

Define $\mathbf{h}_n = \mathbf{w}_n - \mathbf{s}$ as the weight deviation vector. Recalling the equation of evolution for the adaptive filter from Eq. (2), the recursive update equation for $\mathbf{h}(n)$ can be written as

$$\mathbf{h}_n = (\mathbf{I} - \mathbf{k}_n\mathbf{x}_n^T)\mathbf{h}_{n-1} + \mathbf{k}_n\nu_n + \kappa\mathbf{P}_n\mathbf{g}(\mathbf{w}_{n-1})$$

where the definition of ξ_n from Eq. (4) and the equation for y_n have been evoked. The sequence of inverse matrices $\{\mathbf{P}_n\}$ evolve according to the following well known *Riccati* equation [17]

$$\mathbf{P}_n = \lambda^{-1}(\mathbf{I} - \mathbf{k}_n\mathbf{x}_n^T)\mathbf{P}_{n-1} \tag{12}$$

Using this update quation of $\mathbf{P}_n$, the filter evolution equation takes the form

$$\mathbf{h}_n = \lambda\mathbf{P}_n\boldsymbol{\Phi}_{n-1}\mathbf{h}_{n-1} + \mathbf{k}_n\nu_n + \kappa\mathbf{P}_n\mathbf{g}(\mathbf{w}_{n-1}) \tag{13}$$

Utilizing assumptions **A.3** and **A.1**, we can further simplify the evolution equation to get (for large n)

$$\mathbf{h}_n = \eta_n\mathbf{h}_{n-1} + \mathbf{k}_n\nu_n + \rho_n\mathbf{g}(\mathbf{w}_{n-1}) \tag{14}$$

where the following symbols are be used to compactly represent the expressions that will be derived in the paper:

$$\eta_n = \lambda \left(\frac{1 - \lambda^n}{1 - \lambda^{n+1}} \right) \tag{15}$$

$$\rho_n = \frac{\kappa}{P_x} \left(\frac{1 - \lambda}{1 - \lambda^{n+1}} \right) \tag{16}$$

$$c_n = \lambda^n \left(\frac{1 - \lambda}{1 - \lambda^{n+1}} \right) \tag{17}$$

$$d_n = \frac{\kappa}{P_x} \left(\frac{1 - \lambda^n}{1 - \lambda^{n+1}} \right) \tag{18}$$

$$\theta = \frac{\beta\kappa(1 - \lambda)}{P_x} \tag{19}$$

Then the following theorem describes the evolution and convergence of the mean of deviation vector $\mathbf{h}_n$.

Theorem 4.1. *The mean deviation coordinates $\mathbb{E}h_{k,n}$ evolve according to the following recursive equation*

$$\mathbb{E}h_{k,n} = \begin{cases} c_n \mathbb{E}h_{k,0} + d_n g(s_k), & k \in \mathcal{C}_S \\ c_n \mathbb{E}h_{k,0}, & k \in \mathcal{C}_0 \cup \mathcal{C}_L \end{cases} \tag{20}$$

As a result,

$$\mathbb{E}h_{k,\infty} = \begin{cases} \frac{\kappa}{P_x} g(s_k), & k \in \mathcal{C}_S \\ 0, & k \in \mathcal{C}_0 \cup \mathcal{C}_L \end{cases} \tag{21}$$

Proof: The proof is postponed to Appendix B. ∎

B. Mean Square convergence analysis:

We begin by investigating the evolution of the correlation matrix of the mean deviation vector, i.e. $\mathbb{E}\mathbf{h}_n\mathbf{h}_n^T$. From Eq. (14) we get

$$\mathbb{E}\mathbf{h}_n\mathbf{h}_n^T = \mathbf{M}_1 + (\mathbf{M}_2 + \mathbf{M}_2^T) + (\mathbf{M}_3 + \mathbf{M}_3^T) \tag{22}$$
$$+ (\mathbf{M}_4 + \mathbf{M}_4^T) + \mathbf{M}_5 + \mathbf{M}_6 \tag{23}$$

where

$$\mathbf{M}_1 = \lambda^2 \mathbb{E}\left(\mathbf{P}_n \mathbf{\Phi}_{n-1} \mathbf{h}_{n-1} \mathbf{h}_{n-1}^T \mathbf{\Phi}_{n-1} \mathbf{P}_n \right) \tag{24}$$

$$\mathbf{M}_2 = \lambda \mathbb{E}\left(\mathbf{P}_n \mathbf{\Phi}_{n-1} \mathbf{h}_{n-1} \mathbf{k}_n^T \nu_n \right) \tag{25}$$

$$\mathbf{M}_3 = \lambda\kappa \mathbb{E}\left(\mathbf{P}_n \mathbf{\Phi}_{n-1} \mathbf{h}_{n-1} \mathbf{g}^T(\mathbf{w}_{n-1}) \mathbf{P}_n \right) \tag{26}$$

$$\mathbf{M}_4 = \kappa \mathbb{E}\left(\mathbf{P}_n \mathbf{g}(\mathbf{w}_{n-1}) \mathbf{k}_n^T \nu_n \right) \tag{27}$$

$$\mathbf{M}_5 = \kappa^2 \mathbb{E}\left(\mathbf{P}_n \mathbf{g}(\mathbf{w}_{n-1}) \mathbf{g}(\mathbf{w}_{n-1})^T \mathbf{P}_n \right) \tag{28}$$

$$\mathbf{M}_6 = \mathbb{E}\left(\nu_n^2 \mathbf{k}_n \mathbf{k}_n^T \right) \tag{29}$$

By using assumptions **A.1**, **A.2**, and **A.3**, we get the following simplified equations for the terms in the right hand side of (22):

$$\mathbf{M}_1 = \eta_n^2 \mathbb{E}\mathbf{h}_{n-1}\mathbf{h}_{n-1}^T \tag{30}$$

$$\mathbf{M}_2 = \mathbf{0} \tag{31}$$

$$\mathbf{M}_3 = \eta_n \rho_n \left(\frac{1 - \lambda^n}{1 - \lambda^{n+1}} \right) \mathbb{E}\left(\mathbf{h}_{n-1} \mathbf{g}^T(\mathbf{w}_{n-1}) \right) \tag{32}$$

$$\mathbf{M}_4 = \mathbf{0} \tag{33}$$

$$\mathbf{M}_5 = \rho_n^2 \mathbb{E}\left(\mathbf{g}(\mathbf{w}_{n-1}) \mathbf{g}(\mathbf{w}_{n-1})^T \right) \tag{34}$$

$$\mathbf{M}_6 = P_\nu \mathbb{E}\mathbf{k}_n \mathbf{k}_n^T \tag{35}$$

Thus, the evolution equation for the correlation matrix of $\mathbf{h}_n$ can be expressed as

$$\mathbb{E}\mathbf{h}_n\mathbf{h}_n^T = \eta_n^2 \mathbb{E}\mathbf{h}_{n-1}\mathbf{h}_{n-1}^T + \eta_n \rho_n \mathbb{E}\left(\mathbf{h}_{n-1}\mathbf{g}^T(\mathbf{w}_{n-1}) + \mathbf{g}(\mathbf{w}_{n-1})\mathbf{h}_{n-1}^T \right) + \rho_n^2 \mathbb{E}\left(\mathbf{g}(\mathbf{w}_{n-1})\mathbf{g}(\mathbf{w}_{n-1})^T \right) + P_\nu \mathbb{E}\mathbf{k}_n\mathbf{k}_n^T \tag{36}$$

Taking the kth diagonal element of the error covariance matrix we get the corresponding evolution equation:

$$\mathbb{E}h_{k,n}^2 = \eta_n^2 \mathbb{E}h_{k,n-1}^2 + 2\eta_n \rho_n \mathbb{E}(h_{k,n-1}g(w_{k,n-1})) + \rho_n^2 \mathbb{E}(g^2(w_{k,n-1})) + P_\nu \mathbb{E}(k_{k,n})^2 \tag{37}$$

To do the mean square convergence analysis, we introduce the notations that will be henceforth used to succinctly represent the results of the mean square convergence analysis.

$$D_n := \mathbb{E}\left\|\mathbf{h}_n\right\|_2^2 \tag{38}$$

$$\Omega_n := \sum_{k\in\mathcal{C}_0} \mathbb{E}h_{k,n}^2 \tag{39}$$

$$\omega_n^2 := \mathbb{E}h_{k,n}^2 \quad \forall\ k \in \mathcal{C}_0 \tag{40}$$

$$G(s) := \sum_{k\in\mathcal{C}_0} g^2(s_k) \tag{41}$$

$$G'(s) := \sum_{k\in\mathcal{C}_0} s_k g(s_k) \tag{42}$$

1) Instantaneous approximate mean square deviation analysis: In this section we provide the result of an approximate analysis for the instantaneous MSD.

Theorem 4.2. *The instantaneous power of the nonzero and zero taps of the l_0 RLS filter evolve, approximately, according to the following linear dynamical system:*

$$\begin{bmatrix} D_n \\ \Omega_n \end{bmatrix} = \mathbf{A}_n \begin{bmatrix} D_{n-1} \\ \Omega_{n-1} \end{bmatrix} + \mathbf{b}_n \tag{43}$$

where

$$\mathbf{A}_n = \begin{bmatrix} \eta_n^2 & -\frac{2\beta\rho_n\eta_n}{\sqrt{2\pi\omega_\infty^2}} \\ 0 & \eta_n^2 - \frac{2\beta\rho_n\eta_n}{\sqrt{2\pi\omega_\infty^2}} \end{bmatrix} \tag{44}$$

and

$$\mathbf{b}_n = \begin{bmatrix} b_n(1) \\ b_n(2) \end{bmatrix} \tag{45}$$

where

$$b_n(1) = NP_\nu p_n^2 + (N-K)\beta^2\rho_n^2 - 2(N-K)\beta\rho_n\eta_n\omega_\infty^2/\sqrt{2\pi\omega_\infty^2}$$
$$- 2\rho_n c_n \eta_n G'(s) + (2\rho_n d_n \eta_n + \rho_n^2)G(s)$$
$$b_n(2) = (N-K)(P_\nu p_n^2 + \beta^2\rho_n^2) - 2(N-K)\beta\rho_n\eta_n\omega_\infty^2/\sqrt{2\pi\omega_\infty^2}$$

and,

$$\omega_\infty = \frac{-2\lambda\theta/\sqrt{2\pi} + \sqrt{2\lambda^2\theta^2/\pi + (1-\lambda^2)(\theta^2 + P_\nu p_\infty^2)}}{1 - \lambda^2} \tag{46}$$

where θ is defined as in Equation 19.

Proof: The proof is postponed to Appendix C. ■

2) Steady state mean square deviation analysis: Unlike the instantaneous analysis, we can get the expression for steady state MSD exactly under the assumptions taken in Sec. III. The result of that analysis is showed in the form of the following theorem.

Theorem 4.3. *The steady state MSD has the following expression:*

$$D_\infty = \frac{NP_\nu p_\infty^2}{1 - \lambda^2} + \beta_1\theta^2 - \beta_2\theta\sqrt{\theta^2 + \beta_3} \tag{47}$$

where

$$\beta_1 := \frac{N-K}{1-\lambda^2} + \frac{G(s)}{\beta^2(1-\lambda)^2} + \frac{4\lambda^2(N-K)}{\pi(1-\lambda^2)^2}$$

$$\beta_2 := \frac{4\lambda(N-K)}{\sqrt{2\pi}(1-\lambda^2)^2}\sqrt{\frac{2\lambda^2}{\pi} + 1 - \lambda^2}$$

$$\beta_3 := \frac{P_\nu p_\infty^2}{\frac{2\lambda^2}{\pi(1-\lambda^2)} + 1}$$

Proof: The proof is postponed to Appendix D. ■

The appearance of the form of the steady state MSD is identical to the one derived by the authors of [15] since our analysis actually follows the same methodology as theirs. But the terms that calculate the MSD are quite different and also the way the terms $\beta_1, \beta_2, \beta_3$ depend upon the attraction parameter β is different from the way the dependence is for l_0 LMS (See [15] for details). The first term in Eq. (47) is the steady state MSD for conventional RLS and the second and third terms comprise of the "excess" MSD produced by the l_0 attraction term. Note that this excess MSD can very well be negative, for certain range of κ, which results in improved performance of l_0 RLS. In fact, paralleling Corollary 1 of [15], we can get the following corollaries from straight forward calculations:

Corollary 4.1. *For fixed β, l_0 RLS outperforms conventional RLS if the parameter κ is chosen such that the following holds*

$$0 < \theta < \sqrt{\frac{\beta_2^2 \beta_3}{\beta_1^2 - \beta_2^2}} \tag{48}$$

Proof: The proof follows by noticing that l_0 RLS outperforms conventional RLS in steady state MSD if $D_\infty < \frac{NP_\nu p_\infty^2}{(1-\lambda^2)} \implies \beta_1 \theta^2 - \beta_2 \theta \sqrt{\beta_3 + \theta^2} < 0$ and recalling that $\theta > 0$. ∎

Corollary 4.2. *In terms of minimum obtainable MSD from l_0 RLS, the best choice of κ is found from*

$$\theta_{\mathrm{opt}} = \frac{\sqrt{\beta_3}}{2} \left(\sqrt[4]{\frac{\beta_1 + \beta_2}{\beta_1 - \beta_2}} - \sqrt[4]{\frac{\beta_1 - \beta_2}{\beta_1 + \beta_2}} \right) \tag{49}$$

and the minimum MSD is

$$D_\infty^{\mathrm{min}} = \frac{NP_\nu p_\infty^2}{1 - \lambda^2} + \frac{\beta_3}{2} \left(\sqrt{\beta_1^2 - \beta_2^2} - \beta_1 \right) \tag{50}$$

Proof: The proof is the same as the proof of Corollary 1 in [15]. The readers are referred to Appendix A of [15] for details. ∎

From the definitions of $\beta_1, \beta_2, \beta_3$ in Theorem 4.3, it is evident from Corollary 4.2 that the minimum MSD given by l_0 RLS is a function of the attraction parameter β. The following corollary shows that this minimum MSD is, in fact, constant if β is large.

Corollary 4.3. *The minimum steady state MSD D_∞^{min} is a decreasing function of β and as $\beta \to \infty$, the ratio of minimum MSD of l_0 RLS, as found in Corollary 4.2 and the steady state MSD of conventional RLS converges to*

$$\lim_{\beta \to \infty} \frac{D_\infty^{\mathrm{min}}}{D_{\mathrm{RLS}}} = \frac{\pi(1 - \lambda^2) + 2\frac{K}{N}\lambda^2}{\pi(1 - \lambda^2) + 2\lambda^2} \tag{51}$$

which is $\approx K/N$ if λ is close to 1.

Proof: First, observe that β_2, β_3 are independent of β and the only dependence of the steady state MSD on β is through the term β_1. From Equation 47 it is clear that the steady state MSD is an increasing function of β_1 and from the expression of β_1 it is clear that β_1 is a decreasing function of β, which proves the first part of the corollary.

To see how the second part of the corollary comes up, observe that the the expression for β_1 can be rewritten from Theorem 4.3 as

$$\beta_1 = \frac{G(s)}{\beta^2(1-\lambda)^2} + \frac{N-K}{(1-\lambda^2)^2}\left(\frac{4\lambda^2}{\pi} + 1 - \lambda^2\right)$$

$$\implies \lim_{\beta \to \infty} \beta_1 = \frac{N-K}{(1-\lambda^2)^2}\left(\frac{4\lambda^2}{\pi} + 1 - \lambda^2\right)$$

Now, to make the expressions look less formidable, let

$$f_0 = P_\nu p_\infty^2, \ f_1 = \frac{(N-K)}{(1-\lambda^2)^2}, \ f_2^2 = \frac{2\lambda^2}{\pi} + 1 - \lambda^2, \ f_3^2 = \frac{2\lambda^2}{\pi}$$

then,

$$\lim_{\beta \to \infty} \beta_1 = f_1(f_2^2 + f_3^2), \ \beta_2 = 2f_1 f_2 f_3, \ \beta_3 = \frac{f_0(1-\lambda^2)}{f_2^2}$$

so that

$$\lim_{\beta \to \infty} D_\infty^{\min}$$

$$= \frac{(1-\lambda^2)Nf_0f_1}{N-K} + \frac{\beta_3}{2}\left(\sqrt{\beta_1^2 - \beta_2^2} - \beta_1\right)$$

$$= \frac{(1-\lambda^2)Nf_0f_1}{N-K} + \frac{f_0f_1(1-\lambda^2)}{2f_2^2}\left(\sqrt{(f_2^2+f_3^2)^2 - 4f_2^2f_3^2} - (f_2^2+f_3^2)\right)$$

$$= \frac{(1-\lambda^2)Nf_0f_1}{N-K} - \frac{(1-\lambda^2)f_0f_1f_3^2}{f_2^2}$$

$$= D_{RLS}\left(1 - \frac{(N-K)f_3^2}{Nf_2^2}\right)$$

from where the result follows after plugging in the expressions for f_2^2 and f_3^2. ■

Another important observation is that the expression of minimum steady state MSD in Equation 50 is dependent upon the unknown system parameters in the set $\mathcal{C}_S$. This dependence is via $G(s)$ which appears in the expression of β_1. Interestingly, the extent of this dependence is controlled by the attraction parameter β, and as seen from the Corollary 4.3, this dependence vanishes when β becomes large and then the MSD is only a function of λ and the system sparsity to length ratio K/N. In this regard, the following simple corollary connects the behaviour of the minimum steady state MSD with the sparsity of the system and the attraction of the small unknown parameters $G(s)$.

Corollary 4.4. *The minimum steady state MSD in Eq. (50) is a monotonically increasing function of the small set attraction $G(s)$ and the sparsity K.*

Proof: We can write the expression for the minimum steady state MSD as

$$D_\infty^{\min} = \frac{Np_{\infty^2}}{(1-\lambda^2)} - \frac{\beta_2^2\beta_3}{\sqrt{\beta_1^2 - \beta_2^2} + \beta_1}$$

which shows that $D_\infty^{\min}$ increases with the increase of β_1. Then, as β_1 is an increasing function of $G(s)$, $D_\infty^{\min}$ is also a monotonically increasing function of $G(s)$.

To investigate the dependence of the minimum steady state MSD on the sparsity K, first note that the first term is independent of K and hence the behaviour of the second term will suffice for our purpose. Now, let us define, for the sake of simplicity of the expressions,

$$f_1 = \frac{(N-K)}{(1-\lambda^2)^2}, \ f_2^2 = \frac{2\lambda^2}{\pi} + 1 - \lambda^2, \ f_3^2 = \frac{2\lambda^2}{\pi}, f_4 = \frac{G(s)}{\beta^2(1-\lambda)^2},$$

then,

$$\beta_1 = f_1(f_2^2 + f_3^2) + f_4, \ \beta_2 = 2f_1f_2f_3$$

Then, note that we can express the second term as a function of f_1 (and hence as a function of $N-K$) in the following manner:

$$\frac{\beta_3}{2}\left(\sqrt{\beta_1^2 - \beta_2^2} - \beta_1\right)$$

$$= \frac{\beta_3}{2}\left(\sqrt{f_1^2(f_2^2 - f_3^2)^2 + 2f_1f_4(f_2^2+f_3^2) + f_4^2} - (f_1(f_2^2+f_3^2) + f_4)\right)$$

$$= \frac{-2\beta_3 f_1^2 f_2^2 f_3^2}{\sqrt{f_1^2(f_2^2f_3^2)^2 + 2f_1f_4(f_2^2+f_3^2) + f_4^2} + (f_1(f_2^2+f_3^2) + f_4)}$$

$$= \frac{E_1}{E_2}$$

It is trivial to note that E_1 is negative and decreases as f_1 increases. In the same way it is easy to verify that E_2 is positive and increases with f_1. Thus, the second term decreases as f_1 increases, which implies, that the second term increases when K increases. This proves that the minimum steady state MSD increases with the increase in sparsity K. ■

V. NUMERICAL EXPERIMENTS

Numerical experiments are carried out to verify the accuracy of our analysis. In order to perform the experiments, the unknown system vector **s**, is generated by generating its components as independent samples of a $\mathcal{N}(0, 1)$ random variable. Each simulation result is averaged over 100 iterations. Table I documents the various parameter values that are used during the experiments.

TABLE I
PARAMETER VALUES FOR DIFFERENT NUMERICAL EXPERIMENTS

Experiment	N	K	λ	β	κ	SNR
1	64	6	0.995	5	$5 \times 10^{-7} \rightarrow 10^{0.1}\kappa_{\max}$	50dB/25dB
2	64	6	0.995	$10^{-1} \rightarrow 50$	κ_{opt}	50dB
3	64	$1 \rightarrow 61$	0.995	5	κ_{opt}	50dB
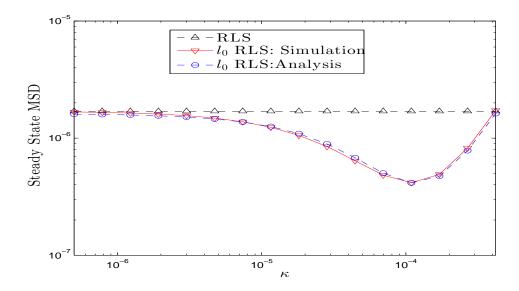
Fig. 1.   Steady state MSD vs κ for SNR 50dB

Figures 1 and 2 compare the steady state MSD of conventional RLS, MSD of l_0 RLS obtained from simulation and MSD of l_0 RLS obtained from the analysis that resulted in Eq. (47) as κ is varied. The figure clearly shows that the theory is in good agreement with the simulation. Also, the value of the optimal κ_{opt} is seen to be well matched with that found from simulation. It can be seen that the tally is better when SNR is 50 dB than when SNR is 25dB. This is expected since decrease in SNR makes the assumptions **A.5** and **A.7** weak.

Figures 3 and 4 plot the variation of steady state MSD with β. it can be seen that the result from analysis matches well with the theory. Also, it is interesting to observe that the decrease in the MSD for l_0 RLS is almost by a factor of $1/10$ compared to the steady state MSD of conventional RLS. This result matches quite closely with the result stated in Corollary 4.3, according to which, this factor should be$\approx K/N = 1/10.667$ using the values of K, N from Table. I for experiment 2.

Figure 5 plots the variation of steady state MSD with sparsity K. The figure clearly verifies the claim of Corollary 4.4.
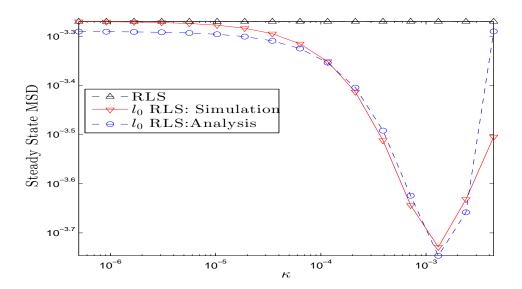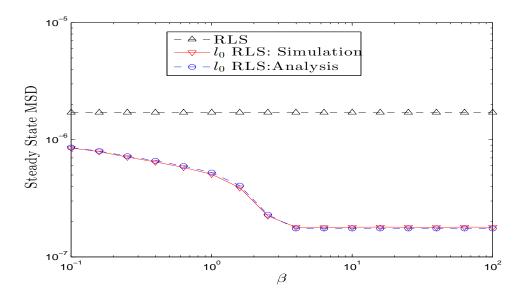
Fig. 2.   Steady state MSD vs κ for SNR 25dB

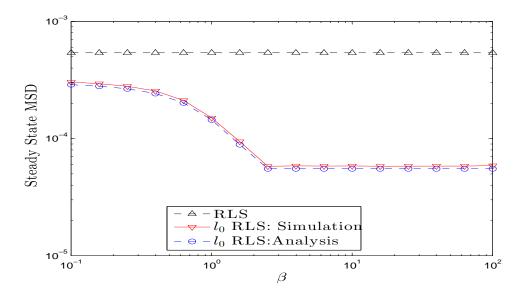Fig. 3.   Steady state MSD vs β for SNR 50dB

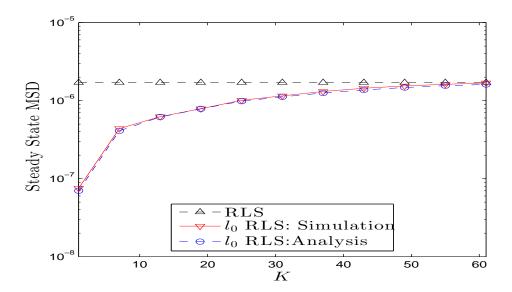Fig. 4.   Steady state MSD vs β for SNR 25dB

Fig. 5.   Steady state MSD vs sparsity K for SNR 50dB

VI. Conclusion

In this paper a theoretical analysis of l_0 RLS is carried out. Inspired by the work in [15], relevant common assumptions are taken along with some new ones and their applicability are discussed. Also the taps are divided into different sets according to their magnitudes and the effect of the set with coefficients with small magnitude is analyzed in detail. The expressions for steady state MSD as well as a linear evolution model of the instantaneous MSD are derived and analyzed for the effects of different parameter settings. Several numerical simulations are done to verify the claims made by the analysis and are seen to match well with the theoretical predictions for a range of parameter values.

Appendix A
Proof of Lemma 3.1

When $\lambda = 1$, the proof follows from the ergodicity of the $\{x(n)\}$ sequence.

For $\lambda \in (0, 1)$, note that from Equation (5) one can write

$$\boldsymbol{\epsilon}_n := \boldsymbol{\Phi}_n - \frac{1 - \lambda^{n+1}}{1 - \lambda}\mathbf{R} = \sum_{m=0}^{n} \lambda^{n-m} \left(\mathbf{x}_m \mathbf{x}_m^T - \mathbf{R}\right)$$

Then,

$$\mathbb{E}\left(\boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^T\right) = \sum_{l,m} \lambda^{2n-l-m} \left(\mathbb{E}\left(\mathbf{x}_l \mathbf{x}_l^T \mathbf{x}_m \mathbf{x}_m^T\right) - \mathbb{E}(\mathbf{x}_l \mathbf{x}_l^T)\mathbf{R} - \mathbf{R}\mathbb{E}(\mathbf{x}_m \mathbf{x}_m^T) + \mathbf{R}^2\right)$$

$$= \sum_{l,m} \lambda^{2n-l-m} \left(\mathbb{E}\left(\mathbf{x}_l \mathbf{x}_l^T \mathbf{x}_m \mathbf{x}_m^T\right) - \mathbf{R}^2\right)$$

Now, using *Gaussian mean factoring theorem* [18], one can find an expression for the $(i,j)^{\text{th}}$ element $(0 \le i, j \le N-1)$ of $\mathbb{E}\left(\mathbf{x}_l \mathbf{x}_l^T \mathbf{x}_m \mathbf{x}_m^T\right)$ in the following way:

$$\left[\mathbb{E}\left(\mathbf{x}_l \mathbf{x}_l^T \mathbf{x}_m \mathbf{x}_m^T\right)\right]_{(i,j)}$$

$$= \sum_{n=0}^{N-1} \mathbb{E}(x(l-n)x(m-n)x(l-i)x(m-j))$$

$$= \sum_{n=0}^{N-1} \left[\mathbb{E}(x(l-n)x(m-n))\mathbb{E}(x(l-i)x(m-j)) + \mathbb{E}(x(l-n)x(m-j))\mathbb{E}(x(m-n)x(l-j))\right.$$

$$\left. + \mathbb{E}(x(l-n)x(l-i))\mathbb{E}(x(m-n)x(m-j))\right]$$

$$= \sum_{n=0}^{N-1} P_x^2 \left(\delta(l-m)\delta(l-m-i+j) + \delta(l-m-n+j)\delta(m-n-l+j) + \delta(i-n)\delta(j-n)\right) \quad \text{(using assumption \textbf{A.1})}$$

$$= \begin{cases} (N+2)P_x^2, & l = m, \ i = j \\ P_x^2, & l \ne m, \ i = j \\ 0, & \text{otherwise} \end{cases}$$

Then, recalling that under assumption **A.1**, $\mathbf{R} = P_x \mathbf{I}$, we get

$$\boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^T = (N+1)P_x^2 \mathbf{I} \sum_{m=0}^{n} \lambda^{2n-2m}$$

$$= \frac{(1 - \lambda^{2(n+1)})(N+1)P_x^2}{1 - \lambda^2}\mathbf{I}$$

So, $\forall \mathbf{u} \in \mathbb{R}^N$ such that $\|\mathbf{u}\|_2 = 1$, we have

$$\left\| \frac{\boldsymbol{\epsilon}_n}{\frac{1-\lambda^{n+1}}{1-\lambda}}\mathbf{u} \right\|_2^2$$

$$= \left(\frac{1-\lambda}{1+\lambda}\right) \left(\frac{1+\lambda^{n+1}}{1-\lambda^{n+1}}\right) (N+1)P_x^2 \to \frac{1-\lambda}{1+\lambda}(N+1)P_x^2$$

as $n \to \infty$. This proves the claim.

APPENDIX B
PROOF OF THEOREM 4.1

Taking expectations on both sides of Eq. (14), and using Assumptions **A.1** and **A.2**, we get

$$\mathbb{E}\mathbf{h}_n = \eta_n \mathbb{E}\mathbf{h}_{n-1} + \rho_n \mathbb{E}\mathbf{g}(\mathbf{w}_{n-1}) \tag{52}$$

To solve the linear system in Eq. (52), an expression for $\mathbb{E}\mathbf{g}(\mathbf{w}_{n-1})$ is needed. Using the assumptions **A.5**, **A.6**, and **A.7**, we get,

$$g(w_{k,n-1}) = \begin{cases} 0 & \forall\ k \in \mathcal{C}_L \\ \beta^2 h_{k,n-1} + g(s_k) & \forall\ k \in \mathcal{C}_S \\ g(h_{k,n-1}) & \forall\ k \in \mathcal{C}_0 \end{cases} \tag{53}$$

Thus, from Eq. (52) it follows that

$$\mathbb{E}h_{k,n} = \begin{cases} \eta_n \mathbb{E}h_{k,n-1} & \forall\ k \in \mathcal{C}_L \\ \eta_n \mathbb{E}h_{k,n-1} + \rho_n g(s_k) & \forall\ k \in \mathcal{C}_S \\ \eta_n \mathbb{E}h_{k,n-1} & \forall\ k \in \mathcal{C}_0 \end{cases} \tag{54}$$

where in Eq. (54) the assumption **A.4** is used to simplify the expression for $\mathbb{E}g(w_{k,n-1})$ for $k \in \mathcal{C}_S$. The expression for $\mathbb{E}g(w_{k,n-1})$ for $k \in \mathcal{C}_0$ is obtained in the following way, using assumption **A.5** and the definition of function $g(\cdot)$ in Eq. (7):

$$\mathbb{E}g(w_{k,n-1}) = \frac{1}{\sqrt{2\pi\omega_{n-1}^2}} \int_{-1/\beta}^{1/\beta} (\beta^2 x - \operatorname{sgn}(x)) e^{-x^2/2\omega_{n-1}^2} dx$$
$$= 0$$

where $\omega_n^2 := \mathbb{E}h_{k,n}^2 \ \forall k \in \mathcal{C}_0$. Then, it follows that $\forall k \in \mathcal{C}_S$,

$$\mathbb{E}h_{k,n} = \begin{cases} \prod_{k=1}^n \eta_k \mathbb{E}h_{k,0} + \left(\sum_{k=1}^n \rho_k \prod_{j=k+1}^n \eta_j\right) g(s_k), & k \in \mathcal{C}_S \\ \prod_{k=1}^n \eta_k \mathbb{E}h_{k,0}, & k \in \mathcal{C}_0 \cup \mathcal{C}_L \end{cases}$$

From definitions of η_n, ρ_n, we find that $\prod_{k=1}^n \eta_k = \lambda^n \frac{1-\lambda}{1-\lambda^{n+1}} = c_n$, and

$$\sum_{k=1}^n \rho_k \prod_{j=k+1}^n \eta_j = \sum_{k=1}^n \left(\frac{\kappa}{P_x}\frac{1-\lambda}{1-\lambda^{k+1}}\right) \cdot \left(\lambda^{n-k}\frac{1-\lambda^{k+1}}{1-\lambda^{n+1}}\right)$$
$$= \rho_n \sum_{k=1}^n \lambda^{n-k}$$
$$= \frac{\kappa}{P_x}\frac{1-\lambda^n}{1-\lambda^{n+1}} = d_n$$

From this the evolution equation for $\mathbb{E}h_{k,n}$ follows. Taking, $n \to \infty$ trivially results in Eq. (21).

APPENDIX C
PROOF OF THEOREM 4.2

To solve the recursion in Eq.(37), we need to evaluate the terms $\mathbb{E}(h_{k,n-1}g(w_{k,n-1}))$, $\mathbb{E}(g^2(w_{k,n-1}))$, and $\mathbb{E}(k_{k,n}^2)$ for each $k \in \{1, 2, \cdots, N\}$.

A. *Evaluating* $\mathbb{E}(h_{k,n-1}g(w_{k,n-1}))$

From Eq. (53) and recalling that $h_{k,n} = w_{k,n} - s_k$, we get

$$h_{k,n-1}g(w_{k,n-1}) = \begin{cases} 0 & \forall\ k \in \mathcal{C}_L \\ \beta^2 h^2_{k,n-1} + g(s_k)h_{k,n-1} & \forall\ k \in \mathcal{C}_S \\ h_{k,n-1}g(h_{k,n-1}) & \forall\ k \in \mathcal{C}_0 \end{cases} \tag{55}$$

So, taking expectations on both sides we get

$$\mathbb{E}h_{k,n-1}g(w_{k,n-1}) = \begin{cases} 0 & \forall\ k \in \mathcal{C}_L \\ \beta^2 \mathbb{E}h^2_{k,n-1} + g(s_k)\mathbb{E}h_{k,n-1} & \forall\ k \in \mathcal{C}_S \\ \mathbb{E}h_{k,n-1}g(h_{k,n-1}) & \forall\ k \in \mathcal{C}_0 \end{cases}$$

To get the expression for $\mathbb{E}h_{k,n-1}g(w_{k,n-1})$ for $k \in \mathcal{C}_0$, we note that, for $k \in \mathcal{C}_0$, using the definition of function $g(\cdot)$ in Eq.(7), we get

$$\mathbb{E}h_{k,n-1}g(h_{k,n-1}) = \frac{1}{\sqrt{2\pi\omega_{n-1}^2}} \int_{-1/\beta}^{1/\beta} (\beta^2 x^2 - \beta|x|)e^{-x^2/2\omega_{n-1}^2}dx$$

Note that assumption **A.5** implies that $\omega_n < 1/\beta, \ \forall n \geq 1$, which permits to approximate the above integral as

$$\mathbb{E}h_{k,n-1}g(h_{k,n-1}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\beta^2\omega_{n-1}^2 x^2 - \omega_{n-1}\beta|x|)e^{-x^2/2}dx$$

$$= \beta^2\omega_{n-1}^2 - \frac{2\beta\omega_{n-1}}{\sqrt{2\pi}}$$

Thus

$$\mathbb{E}h_{k,n-1}g(w_{k,n-1}) = \begin{cases} 0 & \forall\ k \in \mathcal{C}_L \\ \beta^2\mathbb{E}h^2_{k,n-1} + g(s_k)\mathbb{E}h_{k,n-1} & \forall\ k \in \mathcal{C}_S \\ \beta^2\omega_{n-1}^2 - \frac{2\beta\omega_{n-1}}{\sqrt{2\pi}} & \forall\ k \in \mathcal{C}_0 \end{cases} \tag{56}$$

where $\omega_{n-1}^2 := \mathbb{E}h_{k,n-1}^2, \ k \in \mathcal{C}_0$.

B. Evaluating $\mathbb{E}g^2(w_{k,n-1})$

Again using the definition of function $g(\cdot)$ in Eq.(7), we get

$$\mathbb{E}g^2(w_{k,n-1}) = \begin{cases} 0 & \forall\ k \in \mathcal{C}_L \\ \beta^4\mathbb{E}h^2_{k,n-1} + g^2(s_k) + 2\beta^2 g(s_k)\mathbb{E}(h_{k,n-1}) & \forall\ k \in \mathcal{C}_S \\ \mathbb{E}g^2(h_{k,n-1}) & \forall\ k \in \mathcal{C}_0 \end{cases}$$

Again, using assumption **A.5**, we get, $\forall k \in \mathcal{C}_0$,

$$\mathbb{E}g^2(h_{k,n-1}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\beta^2\omega_{n-1}x - \beta\mathrm{sgn}(x))^2 e^{-x^2/2}dx$$

$$= \beta^2 - 2\beta^3\omega_{n-1}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x|e^{-x^2/2}dx + \beta^4\omega_{n-1}^2\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}dx$$

$$= \beta^2 - \frac{4\beta^3\omega_{n-1}}{\sqrt{2\pi}} + \beta^4\omega_{n-1}^2$$

Thus,

$$\mathbb{E}g^2(w_{k,n-1}) = \begin{cases} 0 & \forall\ k \in \mathcal{C}_L \\ \beta^4\mathbb{E}h^2_{k,n-1} + g^2(s_k) + 2\beta^2 g(s_k)\mathbb{E}(h_{k,n-1}) & \forall\ k \in \mathcal{C}_S \\ \beta^2 - \frac{4\beta^3\omega_{n-1}}{\sqrt{2\pi}} + \beta^4\omega_{n-1}^2 & \forall\ k \in \mathcal{C}_0 \end{cases} \tag{57}$$

C. Evaluating $\mathbb{E}k_{k,n}^2$

From the definition of the *gain vector* $\mathbf{k}_n$ in Eq.(6), along with the assumptions **A.1** and **A.3**, we get the following simplified expression for $\mathbf{k}_n$:

$$\mathbf{k}_n = \frac{\mathbf{x}_n}{a_n^2 + \|\mathbf{x}_n\|_2^2} \tag{58}$$

where $a_n^2 := (1 - \lambda^n)a^2$ and $a^2 := \frac{\lambda}{1-\lambda}P_x$. Then,

$$k_{k,n} = \frac{x_{k,n}}{a_n^2 + \|\mathbf{x}_n\|_2^2}$$

Let, $p_{k,n}^2 := \mathbb{E}k_{k,n}^2$. It follows from assumption **A.1** that $p_{0,n} = p_{1,n} = \cdots = p_{N-1,n} = p_n$ where

$$p_n^2 = \frac{1}{N}\mathbb{E}\left(\frac{\|\mathbf{x}_n\|_2^2}{(a_n^2 + \|\mathbf{x}_n\|_2^2)^2}\right)$$

Now, because of the choice of λ in assumption **A.3**, we have $\frac{\lambda}{1-\lambda} >> N$. Then, we can simplify the expression for p_n as an approximation

$$p_n^2 \approx \frac{1}{N}\mathbb{E}\left(\frac{\|\mathbf{x}_n\|_2^2}{(a_n^2)^2}\right)$$

$$= \frac{P_x}{a_n^4}$$

$$\implies p_n^2 = \frac{(1-\lambda)^2}{\lambda^2(1-\lambda^n)^2 P_x} \tag{59}$$

$$p_\infty^2 = \frac{(1-\lambda)^2}{\lambda^2 P_x} \tag{60}$$

D. *Putting everything together*

Thus, using the expressions found in equations 56, 57, and 58 in Eq. (37) and using the assumption **A.4**, we get

$$\mathbb{E}h_{k,n}^2 = \begin{cases} \eta_n^2 \mathbb{E}h_{k,n-1}^2 + P_\nu p_n^2, & k \in \mathcal{C}_L \\ \eta_n^2 \mathbb{E}h_{k,n-1}^2 + P_\nu p_n^2 + 2\rho_n\eta_n g(s_k)\mathbb{E}h_{k,n-1} + \rho_n^2 g^2(s_k), & k \in \mathcal{C}_S \\ \eta_n^2 \mathbb{E}h_{k,n-1}^2 + P_\nu p_n^2 - 4\beta\rho_n\eta_n\frac{\sqrt{\mathbb{E}h_{k,n-1}^2}}{\sqrt{2\pi}} + \beta^2\rho_n^2, & k \in \mathcal{C}_0 \end{cases} \tag{61}$$

This along with Eq. (20) produces the following linear recursion:

$$\mathbb{E}h_{k,n}^2 = \begin{cases} \eta_n^2 \mathbb{E}h_{k,n-1}^2 + P_\nu p_n^2, & k \in \mathcal{C}_L \\ \eta_n^2 \mathbb{E}h_{k,n-1}^2 + P_\nu p_n^2 - 2\rho_n c_n \eta_n s_k g(s_k) + 2\rho_n d_n \eta_n g^2(s_k) + \rho_n^2 g^2(s_k), & k \in \mathcal{C}_S \\ \eta_n^2 \mathbb{E}h_{k,n-1}^2 + P_\nu p_n^2 - 4\beta\rho_n\eta_n\frac{\sqrt{\mathbb{E}h_{k,n-1}^2}}{\sqrt{2\pi}} + \beta^2\rho_n^2, & k \in \mathcal{C}_0 \end{cases} \tag{62}$$

where we have assumed that $\mathbb{E}w_{0,k} = 0, \ \forall k$. Then, it follows from Eq. (62)

$$D_n - \Omega_n = \eta_n^2(D_{n-1} - \Omega_{n-1}) + KP_\nu p_n^2 - 2\rho_n c_n \eta_n G'(s) + (2\rho_n d_n \eta_n + \rho_n^2)G(s) \tag{63}$$

Also, it follows from Eq. (62)

$$\Omega_n = \eta_n^2 \Omega_{n-1} + (N-K)(\beta^2\rho_n^2 + P_\nu p_n^2) - (N-K)\frac{4\beta\rho_n\eta_n}{\sqrt{2\pi}}\omega_{n-1}$$

Observing that $\Omega_n = (N-K)\omega_n^2$, it follows that

$$\Omega_n = \eta_n^2 \Omega_{n-1} + (N-K)(\beta^2\rho_n^2 + P_\nu p_n^2) - \sqrt{N-K}\frac{4\beta\rho_n\eta_n}{\sqrt{2\pi}}\sqrt{\Omega_{n-1}} \tag{64}$$

Thus, using $\Omega_n = (N-K)\omega_n^2, \ \forall n$, it follows from Eq. (64), as $n \to \infty$,

$$\omega_\infty^2 = \eta_\infty^2 \omega_\infty^2 + \beta^2\rho_\infty^2 + P_\nu p_\infty^2 - \frac{4\beta\rho_\infty\eta_\infty}{\sqrt{2\pi}}\omega_\infty$$

$$\implies \omega_\infty = \frac{-2\beta\rho_\infty\eta_\infty/\sqrt{2\pi} + \sqrt{2\beta^2\rho_\infty^2\eta_\infty^2/\pi + (1-\eta_\infty^2)(\beta^2\rho_\infty^2 + P_\nu p_\infty^2)}}{(1-\eta_\infty^2)} \quad (\because \omega_\infty \geq 0)$$

Now, $\eta_\infty = \lambda$, $\rho_\infty = \frac{\kappa(1-\lambda)}{P_x}$. recalling Hence, we have the desired parametric expression for ω_∞ in terms of θ as promised in Theorem 4.2.

For large n, however, an approximate linear evolution for Ω_n can be obtained by a first order Taylor series approximation of $\sqrt{\mathbb{E}h_{k,n-1}^2}$ to get

$$\mathbb{E}h_{k,n-1}^2 \approx \sqrt{\mathbb{E}h_{k,\infty}^2} + \frac{\mathbb{E}h_{k,n-1}^2 - \mathbb{E}h_{k,\infty}^2}{2\sqrt{\mathbb{E}h_{k,\infty}^2}}$$

$$= \frac{\mathbb{E}h_{k,n-1}^2 + \mathbb{E}h_{k,\infty}^2}{2\sqrt{\mathbb{E}h_{k,\infty}^2}}$$

then Eq. (64) will become

$$\Omega_n = \eta_n^2 \Omega_{n-1} + (N-K)(\beta^2\rho_n^2 + P_\nu p_n^2) - \frac{2\beta\rho_n\eta_n}{\sqrt{2\pi\omega_\infty^2}}(\Omega_{n-1} + (N-K)\omega_\infty^2) \tag{65}$$

For large n, thus, Eq. (63) and Eq. (65) together produce the results in Equations (43) and (44).

APPENDIX D
PROOF OF THEOREM 4.3

From Eq. (63), taking $n \to \infty$ and using Eq. (46), we get

$$D_\infty = \frac{(N-K)\beta^2\rho_\infty^2 + NP_\nu p_\infty^2 - 2\rho_\infty c_\infty \eta_\infty G'(s) + (2\rho_\infty d_\infty \eta_\infty + \rho_\infty^2)G(s) - (N-K)\frac{4\beta\rho_\infty \eta_\infty}{\sqrt{2\pi}}\omega_\infty}{1 - \eta_\infty^2}$$

Observing that

$$\rho_\infty = \frac{\kappa(1-\lambda)}{P_x}$$
$$\eta_\infty = \lambda$$
$$c_\infty = 0$$
$$d_\infty = \frac{\kappa}{P_x}$$

we get

$$D_\infty = \frac{(N-K)\theta^2 + NP_\nu p_\infty^2 + \left(\frac{2\lambda\theta^2}{\beta^2(1-\lambda)} + \frac{\theta^2}{\beta^2}\right)G(s) - \frac{4\lambda(N-K)\theta}{\sqrt{2\pi}}\omega_\infty}{1 - \lambda^2} \tag{66}$$

which, together with Eq. (46), yields the desired result.

REFERENCES

[1] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 5, pp. 508–518, 2000.

[2] W. F. Schreiber, "Advanced television systems for terrestrial broadcasting: Some problems and some proposed solutions," *Proceedings of the IEEE*, vol. 83, no. 6, pp. 958–981, 1995.

[3] B. Widrow and S. D. Stearns, "Adaptive signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1985, 491 p.*, vol. 1, 1985.

[4] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.

[5] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, 2006.

[6] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[7] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5406–5425, 2006.

[8] R. L. Das and M. Chakraborty, "Sparse adaptive filters-an overview and some new results," in *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2745–2748.

[9] Y. Chen, Y. Gu, and A. O. Hero III, "Sparse lms for system identification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3125–3128.

[10] Y. Gu, J. Jin, and S. Mei, "norm constraint lms algorithm for sparse system identification," *Signal Processing Letters, IEEE*, vol. 16, no. 9, pp. 774–777, 2009.

[11] B. Babadi, N. Kalouptsidis, and V. Tarokh, "Sparls: The sparse rls algorithm," *Signal Processing, IEEE Transactions on*, vol. 58, no. 8, pp. 4013–4025, 2010.

[12] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, "Online adaptive estimation of sparse signals: Where rls meets the-norm," *Signal Processing, IEEE Transactions on*, vol. 58, no. 7, pp. 3436–3447, 2010.

[13] E. M. Ekşioğlu, "Rls adaptive filtering with sparsity regularization," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 550–553.

[14] E. M. Eksioglu *et al.*, "Rls algorithm with convex regularization," *Signal Processing Letters, IEEE*, vol. 18, no. 8, pp. 470–473, 2011.

[15] G. Su, J. Jin, Y. Gu, and J. Wang, "Performance analysis of norm constraint least mean square algorithm," *Signal Processing, IEEE Transactions on*, vol. 60, no. 5, pp. 2223–2235, 2012.

[16] H. J. Kushner and G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

[17] A. Uncini, *Fundamentals of Adaptive Signal Processing*. Springer, 2015.

[18] M. B. Priestley, "Spectral analysis and time series," 1981.