Regularized recursive least squares for anomaly detection in sparse channel tracking applications

Regularized Recursive Least Squares for Anomaly Detection in Sparse Channel Tracking Applications *

Behtash Babadi
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
behtash@seas.harvard.edu

Vahid Tarokh
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
vahid@seas.harvard.edu

ABSTRACT

In this paper, we study the problem of anomaly detection in sparse channel tracking applications via the ℓ_1-regularized least squares adaptive filter (SPARLS). Anomalies arise due to unexpected adversarial changes in the channel and quick detection of these anomalies is desired. We first prove analytically that the prediction error of the SPARLS algorithm can be substantially lower than that of the widely-used Recursive Least Squares (RLS) algorithm. Furthermore, we present Receiver Operating Characteristic (ROC) curves for the detection/false alarm trade-off of anomaly detection in a sparse multi-path fading channel tracking scenario. These curves reveal the considerable advantage of the SPARLS algorithm over the RLS algorithm.

Keywords

System identification, Sparsity-based signal processing, anomaly detection

1. INTRODUCTION

Adaptive filtering is an important tool in statistical signal processing, most appealing in system identification tasks based on streaming data in environments with unknown statistics [6]. For instance, it is widely used for echo cancellation in speech processing systems and for identification or equalization of wireless communication channels.

A wide range of input-output systems are described by sparse models. For example, the multi-path wireless channel has only a few significant components [2]. Other examples include echo components of sound in indoor environments and natural images. Recently, the SPARLS algorithm has been proposed for adaptive identification of such sparse systems [1]. In particular, it has been shown that the SPARLS algorithm significantly outperforms the widely used Recursive Least Squares (RLS) algorithm for system identification in terms of mean square error (MSE). Moreover, the SPARLS algorithm has a much lower computational complexity in practice.

An important problem in system identification applications is anomaly detection. Suppose that one is interested in tracking the characteristics of an input-output system, where these characteristics are expected to lie in a set of "normal" system realizations. Any observable deviation from the set of normal realizations is characterized as an anomaly. For instance, consider the scenario of under-water communications. Suppose that a number of sensors are transmitting their observations to a fusion center via the under-water acoustic channel. Then, the sudden presence of an under-water mobile object such as a submarine, will change the underlying communication channels. Often times, the detection of such events is desired, and in that case the sensors must be equipped with adaptive filtering mechanisms allowing them to detect and localize such events. The MSE advantages of the SPARLS algorithm makes it very appealing to be incorporated as the adaptive tracking unit for anomaly detection in such scenarios.

In this paper, we first study the prediction error of the SPARLS algorithm, and show that it can be substantially lower than that of the RLS algorithm. Inspired by this appealing feature of the SPARLS algorithm, we present a simple anomaly detection mechanism based on thresholding the instantaneous prediction error. We then present Receiver Operating Characteristic (ROC) curves for the detection/false alarm trade-off of the anomaly detection procedure applied to both the SPARLS and RLS algorithms. These curves reveal the considerable operational improvement of the SPARLS algorithm over the RLS algorithm.

The outline of the paper is as follows: we will give an overview of the sparse system identification setting and the SPARLS algorithm in Section 2. We will then study the prediction error performance of the RLS and SPARLS algorithms in this setting in Section 3. Simulation studies are presented in Section 4, followed by conclusion in Section 5.

2. ADAPTIVE SPARSE SYSTEM IDENTIFICATION

2.1 Adaptive Filtering Setup

*This research is supported in part by ARO MURI grant number W911NF-07-1-0376. The views expressed in this paper are those of the authors alone and not of the sponsor.

Consider the conventional adaptive filtering setup, consisting of a transversal filter followed by an adaptation block. The tap-input vector at time i is defined by

$$\mathbf{x}(i) := [x(i), x(i-1), \cdots, x(i-M+1)]^T \quad (1)$$

where $x(k)$ is the input at time k, $k = 1, \cdots, n$. The tap-weight vector at time n is defined by

$$\hat{\mathbf{w}}(n) := [\hat{w}_0(n), \hat{w}_1(n), \cdots, \hat{w}_{M-1}(n)]^T. \quad (2)$$

The output of the filter at time i is given by

$$y(i) := \hat{\mathbf{w}}^*(n)\mathbf{x}(i). \quad (3)$$

where $(\cdot)^*$ denotes the conjugate transpose operator. Let $d(i)$ be the desired output of the filter at time i. We can define the instantaneous (error of the filter as

$$e(i) := d(i) - y(i) = d(i) - \hat{\mathbf{w}}^*(n)\mathbf{x}(i). \quad (4)$$

The adaptation block at time n solves the following optimization problem:

$$\min_{\hat{\mathbf{w}}(n)} \quad f\big(e(1), e(2), \cdots, e(n)\big), \quad (5)$$

where $f \geq 0$ is a certain cost function. In particular, suppose that $d(i)$ is generated by an unknown tap-weight $\mathbf{w}(n)$, *i.e.*, $d(i) = \mathbf{w}^*(n)\mathbf{x}(i) + \eta(i)$, where $\eta(i)$ is the observation noise. With an appropriate choice of f, one can possibly obtain a good approximation to $\mathbf{w}(n)$ by solving the optimization problem given in (5). Note that $\mathbf{w}(n)$ reflects the true parameters which may or may not vary with time. The noise will be assumed to be i.i.d. Gaussian, i.e., $\eta(i) \sim \mathcal{N}(0, \sigma^2)$. The adaptation block has only access to input, output and observation triplet $(x(i), y(i), d(i))$.

A suitable cost function for tracking time-varying systems is defined as follows:

$$f_{RLS}\big(e(1), e(2), \cdots, e(n)\big) := \sum_{i=1}^{n} \lambda^{n-i}|e(i)|^2, \quad (6)$$

with λ a non-negative constant denoted by the *forgetting factor*. The solution to the optimization problem in Eq. (5) with f_{RLS} gives rise to the well-known Recursive Least Squares (RLS) algorithm (See, for example, [6]). Let

$$\mathbf{D}(n) := \mathrm{diag}(\lambda^{n-1}, \lambda^{n-2}, \cdots, 1), \quad (7)$$

$$\mathbf{d}(n) := [d^*(1), d^*(2), \cdots, d^*(n)]^T \quad (8)$$

and $\mathbf{X}(n)$ be an $n \times M$ matrix whose ith row is $\mathbf{x}^*(i)$, *i.e.*,

$$\mathbf{X}(n) := \begin{pmatrix} \mathbf{x}^*(1) \\ \vdots \\ \mathbf{x}^*(n-1) \\ \mathbf{x}^*(n) \end{pmatrix}. \quad (9)$$

The RLS cost function can be written in the following form:

$$f_{RLS}\big(e(1), e(2), \cdots, e(n)\big) = \big\|\mathbf{D}^{1/2}(n)\mathbf{d}(n) - \mathbf{D}^{1/2}(n)\mathbf{X}(n)\hat{\mathbf{w}}(n)\big\|_2^2, \quad (10)$$

where $\mathbf{D}^{1/2}(n)$ is a diagonal matrix with entries $D_{ii}^{1/2}(n) := \sqrt{D_{ii}(n)}$.

2.2 The SPARLS algorithm

The SPARLS algorithm, introduced in [1], iteratively minimizes the cost function

$$\frac{1}{2\sigma^2}\big\|\mathbf{D}^{1/2}(n)\mathbf{d}(n) - \mathbf{D}^{1/2}(n)\mathbf{X}(n)\hat{\mathbf{w}}(n)\big\|_2^2 + \gamma\|\hat{\mathbf{w}}(n)\|_1 \quad (11)$$

by updating the estimate $\hat{\mathbf{w}}(n)$ upon the arrival of the input and observation pair. The parameter γ represents a trade off between estimation error and sparsity of the parameter coefficients. The SPARLS algorithm can be summarized as follows:

Algorithm 1 SPARLS

Inputs: $\mathbf{B}(1) = \mathbf{I} - \frac{\alpha^2}{\sigma^2}\mathbf{x}(1)\mathbf{x}^*(1)$, $\mathbf{u}(1) = \frac{\alpha^2}{\sigma^2}\mathbf{x}(1)d^*(1)$, $\tau := \gamma\alpha^2$ and K.
Output: $\hat{\mathbf{w}}(n)$.

1: **for all** Input $x(n)$ **do**
2: $\quad\mathbf{B}(n) = \lambda\mathbf{B}(n-1) - \frac{\alpha^2}{\sigma^2}\mathbf{x}(n)\mathbf{x}^*(n) + (1-\lambda)\mathbf{I}$.
3: $\quad\mathbf{u}(n) = \lambda\mathbf{u}(n-1) + \frac{\alpha^2}{\sigma^2}d^*(n)\mathbf{x}(n)$.
4: $\quad$Set $\hat{\mathbf{w}}^{(0)} = \hat{\mathbf{w}}(n-1)$.
5: $\quad$**for all** $\ell = 1, 2, \cdots, K-1$ **do**
6: $\quad\quad$EM iteration: $\hat{\mathbf{w}}^{(\ell+1)} := \mathsf{ST}_\tau\big(\mathbf{B}(n)\hat{\mathbf{w}}^{(\ell)} + \mathbf{u}(n)\big)$.
7: $\quad$**end for**
8: $\quad$Update $\hat{\mathbf{w}}(n) = \hat{\mathbf{w}}^{(K)}$.
9: **end for**

The operator $\mathsf{ST}_\tau(\cdot) : \mathbb{C}^M \mapsto \mathbb{C}^M$ in line 6 of the SPARLS algorithm is denoted by *elementwise soft-thresholding* and is given by:

$$\begin{aligned}\big(\mathsf{ST}_\tau(\mathbf{w})\big)_i :=& \operatorname{sgn}\big(\Re\{w_i\}\big)\big(|\Re\{w_i\}| - \tau\big)_+ \\ &+ i\operatorname{sgn}\big(\Im\{w_i\}\big)\big(|\Im\{w_i\}| - \tau\big)_+ \end{aligned} \quad (12)$$

where $\mathrm{sgn}(\cdot)$ is the standard signum operator, ad $(x)_+ := \max(x, 0)$. The performance of the SPARLS algorithm is studied comprehensively in [1]. In particular, it has been shown that the SPARLS algorithm is capable of achieving significant MSE gains in sparse system identification, compared to the widely used RLS algorithm.

3. PREDICTION ERROR ANALYSIS

3.1 Main Results

Let $\hat{\mathbf{w}}(n-1)$ be the estimate of $\mathbf{w}(n-1)$ at time n. We consider the commonly used random-walk regeneration model for the time evolution of the channel $\mathbf{w}(n)$:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \boldsymbol{\delta}(n) \quad (13)$$

where $\boldsymbol{\delta}(n) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta})$. We assume that $\boldsymbol{\delta}(n)$, $\mathbf{x}(n)$ and $\eta(n)$ are statistically independent at all times. The prediction error at time n is given by:

$$\xi(n) := d(n) - \hat{\mathbf{w}}^*(n-1)\mathbf{x}(n). \quad (14)$$

This prediction error is often denoted by *a priori* estimation error. Let

$$\mathbf{R} := \mathbb{E}_{\mathbf{x}}\{\mathbf{x}(n)\mathbf{x}^*(n)\} \quad (15)$$

be the input covariance matrix. We are interested in studying the average steady state behavior of $\xi(n)$ for both the RLS and SPARLS adaptive filters. We have:

$$\xi(n) = (\mathbf{w}^*(n) - \hat{\mathbf{w}}^*(n-1))\mathbf{x}(n) + \eta(n)$$

Therefore,

$$\mathbb{E}_{\eta,\mathbf{x}(n),\boldsymbol{\delta}(n)}\{|\xi(n)|^2\} := \mathbb{E}_{\boldsymbol{\delta}(n)}\{\boldsymbol{\epsilon}^*(n)\mathbf{R}\boldsymbol{\epsilon}(n)\} + \sigma^2, \quad (16)$$

where

$$\boldsymbol{\epsilon}(n) := \mathbf{w}(n) - \hat{\mathbf{w}}(n-1). \quad (17)$$

For the sake of simplicity, suppose that the input covariance matrix is diagonal, *i.e.*, $\mathbf{R} = \mathbb{I}_{M\times M}$. Hence

$$\mathbb{E}_{\eta,\mathbf{x}(n),\boldsymbol{\delta}(n)}\{|\xi(n)|^2\} := \mathbb{E}_{\boldsymbol{\delta}(n)}\{\|\boldsymbol{\epsilon}(n)\|_2^2\} + \sigma^2. \quad (18)$$

The error term $\boldsymbol{\epsilon}(n)$ can be written in the following form:

$$\boldsymbol{\epsilon}(n) = (\mathbf{w}(n-1) - \hat{\mathbf{w}}(n-1)) + \boldsymbol{\delta}(n)$$

Hence,

$$\mathbb{E}_{\boldsymbol{\delta}(n)}\{\|\boldsymbol{\epsilon}(n)\|_2^2\} = e^2(n) + \mathrm{Tr}(\boldsymbol{\Delta}). \quad (19)$$

where

$$e(n) := \big\|\mathbf{w}(n-1) - \hat{\mathbf{w}}(n-1)\big\|_2 \quad (20)$$

is the *a posteriori* estimation error at time $n-1$, which contributes directly to the prediction error. The average steady state behavior of the *a posteriori* error is well-known in the literature. We thus state the following proposition regarding the steady state prediction error of the RLS algorithm:

PROPOSITION 3.1. *Let $\xi^{\mathsf{RLS}}(n)$ be the prediction error of the RLS algorithm in a time-varying environment where $\lambda < 1$. Suppose that $\mathbf{R}$, the input covariance matrix, is the identity matrix. Then, we have:*

$$\mathbb{E}\{|\xi^{\mathsf{RLS}}(n)|^2\} \geq \Big(\frac{1-\lambda}{1+\lambda}M + 1\Big)\sigma^2 + \mathrm{Tr}(\boldsymbol{\Delta}).$$

in the steady state, where the expectation is over $\{\mathbf{x}(i)\}_{i=1}^n$ and $\{\eta(i)\}_{i=1}^n$.

PROOF. By the result of Section III of [4] we have:

$$\mathbb{E}\Big\{e^{\mathsf{RLS}}(n)\Big\} \geq \frac{1-\lambda}{1+\lambda}M\sigma^2 \quad (21)$$

By combining Eqs. (18)–(20) and Eq. (21), the claim of the proposition follows. □

Next, we study the steady state prediction error of the SPARLS algorithm. We prove the following proposition regarding the prediction error of the SPARLS algorithm:

PROPOSITION 3.2. *Let $\xi^{\mathsf{SPARLS}}(n)$ be the* a posteriori *estimation error of the SPARLS algorithm. Let $L \leq \frac{1}{3\mu_0}$ for some $\mu_0 < 1$ and $\mathbf{R}$ be the identity matrix. Then, we have:*

$$\mathbb{E}\{|\xi^{\mathsf{SPARLS}}(n)|^2\} \leq \bigg(\frac{1}{(1-\rho^K)^2}\Big(\sqrt{3} + \frac{3\gamma}{2\sigma}\Big)^2 L + 1\bigg)\sigma^2 + \mathrm{Tr}(\boldsymbol{\Delta})$$

with probability exceeding

$$1-(M-L)\exp\Big(-\tfrac{\gamma^2}{8\sigma^2}\Big)-\exp(-L/7)-2M^2\exp\Big(-\tfrac{\mu_0^2}{6(1-\lambda)}\Big),$$

where the expectation is with respect to $\eta(n)$, $\mathbf{x}(n)$ and $\boldsymbol{\delta}(n)$, and $\rho \leq 1 - \frac{\alpha^2}{\sigma^2}(1-\tau)$ with probability exceeding

$$1 - 3M^2\exp\Big(-\frac{\tau^2}{54M^2(1-\lambda)}\Big).$$

PROOF. The proof is given in the Appendix. □

3.2 Discussion

Proposition 3.1 gives a lower bound on the steady state prediction error of the RLS algorithm. The estimation error is proportional to M, which is the length of the channel. Therefore, the lower bound is independent of L, the number of nonzero elements of $\mathbf{w}(n)$.

On the other hand, Proposition 3.2 presents an upper bound on the prediction error of the SPARLS algorithm, where the estimation error is indeed proportional to L. Hence, in the low sparsity regime, the prediction error of the SPARLS algorithm can be substantially lower than that of the RLS algorithm, for large values of L and M. This result will be confirmed in Section 4, in a multi-path fading channel tracking application. Note that the result of Proposition 3.2 is probabilistic. However, with appropriate choices of γ and λ for a given M and L, the success probability can be made arbitrarily close to 1.

4. SIMULATION STUDIES

4.1 Simulation setting

We consider the scenario of tracking a sparse multi-path fading channel, generated by the Jake's model [7]. In the Jake's model, each channel component is sampled from a Rayleigh random process with autocorrelation function given by

$$R(n) = \mathrm{J}_0(2\pi n f_d T_s) \quad (22)$$

where $\mathrm{J}_0(\cdot)$ is the zeroth order Bessel function, f_d is the Doppler frequency shift and T_s is the channel sampling interval. The dimensionless parameter $f_d T_s$ gives a measure of how fast each tap is changing over time. For the purpose of simulations, T_s is normalized to 1.

The channel length is $M = 100$, where as there are only $L = 5$ nonzero elements. The channel vector is normalized to have norm 1. We probe the channel by a sequence of i.i.d. Gaussian inputs.

4.2 Prediction error

Fig. 1 shows the steady state prediction error of the SPARLS and RLS algorithms, for $f_d T_s = 0.0005$, and $\sigma^2 = 0.0001$. The choice of λ for the RLS algorithm in this case is 0.98. The SPARLS algorithm is operating with the choice of parameters $\lambda = 0.98$, $\gamma = 35$, $\alpha = \sigma/2$ and $K = 1$ (See Tables I and II of [1] for details). The normalized steady state prediction error is defined as:

$$10\log_{10}\bigg(\frac{\xi(n)}{\mathbb{E}\{|y(n)|^2\}}\bigg), \quad (23)$$

where $y(n) := \mathbf{w}^*(n)\mathbf{x}(n)$ is the noiseless output of the channel, and the expectation is over $\mathbf{w}(n)$ and $\mathbf{x}(n)$. A rectangular smoothing window of length 100 is used for the prediction error for the sake of presentation.

The small-scale fluctuations of the prediction error are due to the observation noise, whereas the large-scale fluctuations are due to the temporal variations of the underlying channel. As it can be observed from Fig. 1, the SPARLS algorithm has a gain of about 5dB over the RLS algorithm in terms of the prediction error.

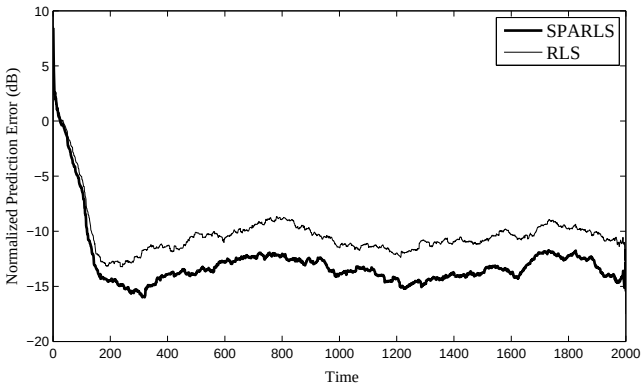

Figure 1: Normalized prediction error of SPARLS and RLS for $f_dT_s = 0.0005$, and $\sigma^2 = 0.0001$.

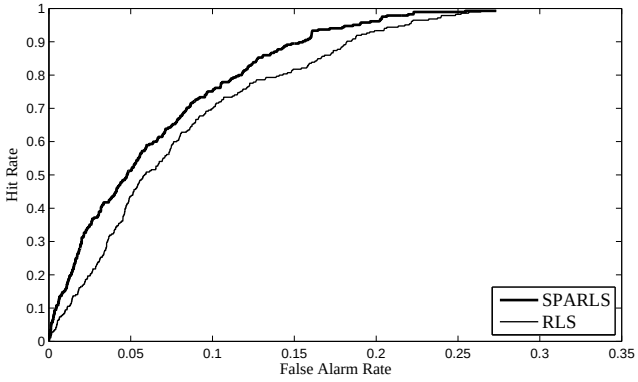

Figure 2: ROC curves for SPARLS and RLS in an anomaly detection application.

4.3 Receiver Operating Characteristic

Next, we consider an anomaly detection scenario in a channel tracking application. Suppose that the set of "normal" channels are those with L nonzero element, where each element is a Raleigh fading component. Moreover, suppose that these elements are located among the first 20 coordinates of $\mathbf{w}(n)$.

We consider the following on-off anomaly class: suppose that at certain time instances, the last 50 elements of $\mathbf{w}(n)$ take random values uniformly distributed in the interval $[0.05, 0.15]$. These values are persistent for t_0 time instances and disappear. Suppose that these anomalies occur according to a Poisson process with rate p. For instance, this anomaly class can model the sudden appearance and disappearance of a mobile underwater vessel which gives rise to high delay components in the underwater acoustic channel. The objective is to detect and localize the anomalies by observing the prediction error of the adaptive filters tracking the channel.

As for the anomaly detection procedure, we monitor $\xi(n)$, the prediction error of the filter (smoothed with a moving average window of length $t_0/2 = 25$) and compare it with a threshold value ξ_{th}. An anomaly is reported if the instantaneous prediction error exceeds the threshold ξ_{th}.

In order to compare the performance of the SPARLS and RLS algorithms in this scenario, we look at the Receiver Operating Characteristic (ROC) curves obtained for both filters. The ROC curve shows the hit rate (detection) versus the false alarm rate of each filter, by sweeping through an admissible range values for ξ_{th}. The hit rate is defined as the ratio of correctly detected anomalies to the total number of anomalies. An anomaly is detected correctly, if it is detected within a time interval of $t_0/2$ from the occurred anomalous event. The false alarm rate is defined as the ratio of falsely detected anomalies over the total number of segments of length t_0 in the observed data. In this sense, the false alarm rate denotes the probability of a false detection per time instance of the observed data. For obtaining the ROC curves, the threshold value ξ_{th} varies uniformly in the interval $[0, 0.1]$.

Fig. 2 shows the ROC curve for the SPARLS and RLS algorithms for $f_dT_s = 0.0005$ and $\sigma^2 = 0.0001$. The Poisson process has a rate of $p = 0.05$ and $t_0 = 50$, and the curves are obtained by averaging over 50000 time instances. As it can be observed from Fig. 2, the ROC curve of the SPARLS algorithm lies above that of the RLS algorithm. In other words, for a given false alarm rate, the SPARLS algorithm has a higher hit rate, and for a given hit rate, the SPARLS algorithm has a lower false alarm rate. For instance, at a false alarm rate of 10%, the SPARLS and RLS algorithms have hit rates of 75% and 70%, respectively. Similarly, at a hit rate of 85%, the SPARLS and RLS algorithms have false alarm rates of 13% and 17%, respectively.

5. COLCLUSION

In this paper, we have studied the prediction error performance of the SPARLS and RLS algorithms, revealing the significant advantage of the SPARLS algorithm over the RLS. Moreover, we have proposed a simple anomaly detection scheme and presented the corresponding ROC curves for both the SPARLS and RLS algorithms, in a multi-path fading channel tracking scenario. These curves suggest that the SPARLS algorithm outperforms the RLS algorithm in terms of detection/false alarm trade-off.

6. REFERENCES

[1] B. Babadi, N. Kalouptsidis, and V. Tarokh. SPARLS: The sparse RLS algorithm. *IEEE Transactions on Signal Processing*, 58(8):4013–4025, 2010.

[2] W. Bajwa, J. Haupt, G. Raz, and R. Nowak. Compressed channel sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS'08)*, March 2008.

[3] Z. Ben-Haim, Y. C. Eldar, and M. Elad. Coherence-based performance guarantees for estimating a sparse vector under random noise. *IEEE Transactions on Signal Processing*, 58(10):5030Ű–5043, Oct. 2010.

[4] E. Eleftheriou and D. Falconer. Tracking properties and steady-state performance of RLS adaptive filter algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(5):1097–1110, Oct. 1986.

[5] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications

to sparse channel estimation. *IEEE Transactions on Information Theory*, 56(11):5862–5875, Nov. 2010.

[6] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 3rd edition, 1996.

[7] W. C. Jakes, editor. *Microwave Mobile Communications*. New York: John Wiley & Sons Inc, 1975.

[8] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1303–1338, October 2000.

APPENDIX

A. PROOF OF PROPOSITION 3.2

The proof is mainly based on Section IV-B of [1] and Theorem 3 of [3]. Let $\tilde{\mathbf{w}}(n)$ be the solution to the ℓ_1-regularized quadratic cost function. Let $e^{\mathsf{SPARLS}}(n) := \|\mathbf{w}(n)-\hat{\mathbf{w}}^{\mathsf{SPARLS}}(n)\|_2$. It has been shown in [1] that we have:

$$e^{\mathsf{SPARLS}}(n) \leq \rho^K e^{\mathsf{SPARLS}}(n-1) + \|\mathbf{w}(n) - \tilde{\mathbf{w}}(n)\|_2 \quad (24)$$

where $\rho := 1 - \frac{\alpha^2}{\sigma^2} s_{\min}\big(\mathbf{X}^*(n)\mathbf{D}(n)\mathbf{X}(n)\big)$, where $s_{\min}(\cdot)$ denotes the minimum eigen-value. Also, from Theorem 3 of [3], we have:

$$\|\mathbf{w}(n) - \tilde{\mathbf{w}}(n)\|_2^2 \leq \Big(\sqrt{3} + \frac{3\gamma}{2\sigma}\Big)^2 L\sigma^2 \quad (25)$$

with probability exceeding

$$1 - (M-L)\exp\Big(-\frac{\gamma^2}{8\sigma^2}\Big) - \exp(-L/7)$$

given $L \leq 1/3\mu$, where μ is the *coherence* of the matrix $\mathbf{D}^{1/2}(n)\mathbf{X}(n)$. The coherence of an $N \times M$ matrix $\mathbf{A}$ with columns $\{\mathbf{a}_i\}_{i=1}^M$ is defined as

$$\mu := \max_{i \neq j} |\mathbf{a}_i^* \mathbf{a}_j|. \quad (26)$$

It only remains to establish bounds on the coherence and the minimum singular value of the matrix $\mathbf{D}^{1/2}(n)\mathbf{X}(n)$.

For simplicity, we assume that $x(i) \sim \mathcal{N}(0, \nu^2)$, for all $i = 1, 2, \cdots, n$. The following lemma establishes a lower bound on the minimum singular value of $\mathbf{D}^{1/2}(n)\mathbf{X}(n)$:

LEMMA A.1. *Let $\nu^2 \frac{1-\lambda^{n+1}}{1-\lambda} = 1$. Then, the eigen-values of $\mathbf{C}(n) := \mathbf{X}^*(n)\mathbf{D}(n)\mathbf{X}^*(n)$ lie in the interval $[1-\tau, 1+\tau]$ with probability exceeding*

$$1 - 3M^2 \exp\Big(-\frac{\tau^2}{54M^2(1-\lambda)}\Big).$$

PROOF. Let $n_\lambda := \frac{1-\lambda^{n+1}}{1-\lambda}$. The ith diagonal element of $\mathbf{C}(n)$ is given by

$$C_{ii}(n) = \sum_{k=1}^n \lambda^{n-k} x_i^2(k)$$

with $\mathbb{E}(C_{ii}(n)) = n_\lambda \nu^2$, where $n_\lambda := \frac{1-\lambda^{n+1}}{1-\lambda}$. Using the standard χ^2 tail bounds given in Lemma 1 in Section 4.1 of [8] with $a_i := \lambda^{n-i}$, $i = 1, 2, \cdots, n$ we get:

$$\mathbb{P}\big(|C_{ii}(n) - n_\lambda \nu^2| \geq 4\nu^2 \sqrt{n_{\lambda^2} t}\big) \leq 2\exp(-t)$$

for $0 \leq t \leq 1$. Also, a slight modification of Lemma 6 in [5] yields:

$$\mathbb{P}\big(|C_{ij}(n)| \geq t\big) \leq 2\exp\Big(-\frac{t^2}{4\nu^2(n_{\lambda^2}\nu^2 + t/2)}\Big)$$

Similar to [5], we seek conditions on λ, n and ν^2 such that the eigenvalues of $\mathbf{C}(n)$ lie in the interval $[1-\tau, 1+\tau]$, where $\tau < 1$ is a positive constant. It can be shown that if $n_\lambda \nu^2 = 1$, and n is large enough so that $n_\lambda \approx \frac{1}{1-\lambda}$, by an application of the Gersgorin's disc theorem [5] the eigen-values of $\mathbf{C}(n)$ lie in the above interval with probability exceeding

$$1 - 3M^2 \exp\Big(-\frac{\tau^2}{54M^2(1-\lambda)}\Big), \quad (27)$$

which establishes the claim of the lemma. □

Next, we present the following lemma establishing an upper bound on the coherence of the matrix $\mathbf{D}^1/2(n)\mathbf{X}(n)$:

LEMMA A.2. *Let $x(i) \sim \mathcal{N}(0, \nu^2)$, for all $i = 1, 2, \cdots, n$. Then, if $\nu^2 \frac{1-\lambda^{n+1}}{1-\lambda} = 1$, the coherence of the matrix $\mathbf{D}^{1/2}(n)\mathbf{X}(n)$ is bounded by μ_0 for some arbitrary constant $\mu_0 < 1$, with probability exceeding*

$$1 - 2M^2 \exp\Big(-\frac{\mu_0^2}{6(1-\lambda)}\Big) \quad (28)$$

for $n \gg 1$.

PROOF. The coherence of the matrix $\mathbf{D}^1/2(n)\mathbf{X}(n)$ is given by

$$\mu = \max_{i,j} \Big|\sum_{k=1}^n \lambda^{n-k} x_i^*(k) x_j(k)\Big| \quad (29)$$

Again, by a slight modification of Lemma 6 of [5] we get:

$$\mathbb{P}\Big(\big|\sum_{k=1}^n \lambda^{n-k} x_i^*(k) x_j(k)\big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{4\nu^2(n_{\lambda^2}\nu^2 + t/2)}\Big)$$

where $n_{\lambda^2} := \frac{1-\lambda^{2(n+1)}}{1-\lambda^2}$. Assuming that n is large enough so that $n_{\lambda^2} \approx \frac{1}{1-\lambda^2}$, and the choice of $n_\lambda \nu^2 = 1$, we will get

$$\mathbb{P}\Big(\mu \geq \mu_0\Big) \leq 2M^2 \exp\Big(-\frac{\mu_0^2}{6(1-\lambda)}\Big) \quad (30)$$

which establishes the statement of the lemma. □

Note that if $\lambda \geq 1 - \mathcal{O}(1/\log M)$, then the probability of $\mu \leq \mu_0$ goes to 1 at a polynomial rate in M. Similarly, if $\lambda \geq 1 - \mathcal{O}(1/M^2 \log M)$, the probability that the eigen-values lie in the interval $[1-\tau, 1+\tau]$ tends to 1 at a polynomial rate in M. One can take such a choice of λ, and therefore the coherence of the matrix $\mathbf{D}^{1/2}(n)\mathbf{X}(n)$ is upper bounded by μ_0 and the minimum singular value is lower bounded by $1-\tau$, for some constants $\mu_0 < 1$ and $\tau < 1$, with high probability. By Lemmas A.1 and A.2, and combining Eqs. (24) and (25), the result of the proposition follows.