Block-Sparsity-Induced Adaptive Filter for Multi-Clustering System Identification

Shuyang Jiang and Yuantao Gu *

Received October 18, 2014; revised March 10 and May 16; accepted June 15, 2015; to appear in *IEEE Transactions on Signal Processing*

Abstract

In order to improve the performance of least mean square (LMS)-based adaptive filtering for identifying block-sparse systems, a new adaptive algorithm called block-sparse LMS (BS-LMS) is proposed in this paper. The basis of the proposed algorithm is to insert a penalty of block-sparsity, which is a mixed $l_{2,0}$ norm of adaptive tap-weights with equal group partition sizes, into the cost function of traditional LMS algorithm. To describe a block-sparse system response, we first propose a Markov-Gaussian model, which can generate a kind of system responses of arbitrary average sparsity and arbitrary average block length using given parameters. Then we present theoretical expressions of the steady-state misadjustment and transient convergence behavior of BS-LMS with an appropriate group partition size for white Gaussian input data. Based on the above results, we theoretically demonstrate that BS-LMS has much better convergence behavior than l_0-LMS with the same small level of misadjustment. Finally, numerical experiments verify that all of the theoretical analysis agrees well with simulation results in a large range of parameters.

Keywords: adaptive filtering, block-sparse system identification, convergence behavior, performance analysis, Markov-Gaussian model.

1 Introduction

Adaptive filtering has been an important research area that attracts much interest in both theoretical and applied issues for a long time [1]. In many scenarios, the unknown systems to be identified are sparse or compressible, which means that most of the entries are zero or close to zero and only a small number of nonzero or large coefficients exist in the long impulse response (Fig. 1(a)). The typical sparse systems are digital TV transmission channels [2] and echo paths [3]. Among all kinds of sparse systems, there are families called block-sparse

*The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mail: jiangshuyang1992@126.com, gyt@tsinghua.edu.cn). The corresponding author of this work is Yuantao Gu.

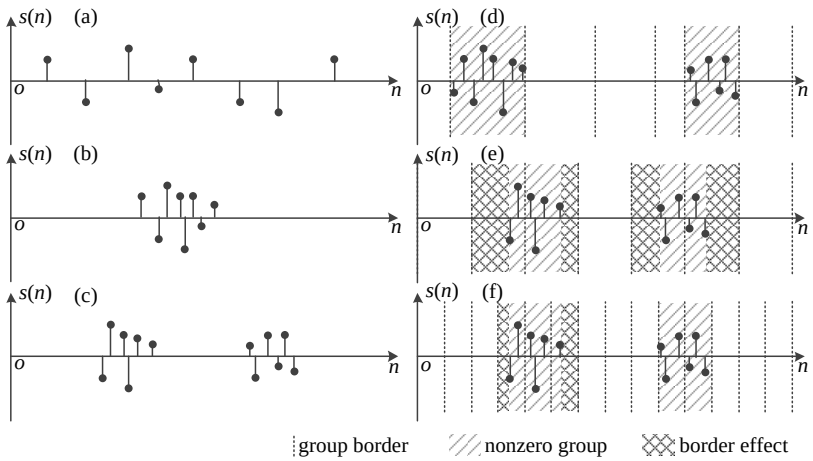

Figure 1: (a) A general sparse system. (b) A block-sparse system with one nonzero block. (c) A block-sparse system with two nonzero blocks. (d) The active regions are located randomly in known partition groups. (e) (f) The location of each cluster is arbitrary and unknown, and all of the group partition sizes are the same for practical implementation.

systems and block-compressible systems [4]. Distinguished from general sparse systems in which the nonzero coefficients may be arbitrarily located, the impulse response of a block-sparse system consists of one or more clusters, wherein a cluster is a gathering of nonzero coefficients (Fig. 1(b,c)). The acoustic echo path is a typical example of single-clustering sparse systems. In satellite-linked or in-door MIMO communications, the impulse response of the echo path consists of several long flat delay regions and disperse active regions, which is a representative of multi-clustering sparse systems [5, 6].

The least mean square (LMS) algorithm [7] is widely used in various applications due to its low computational cost, easy implementation, and high robustness. However, the traditional LMS has no particular improvement on block-sparse system identification. Many algorithms have been proposed to take advantage of the prior knowledge of block-sparsity. In some algorithms, an auxiliary filter is needed to estimate the positions of the disperse regions. Based on the location information, a number of short adaptive filters are centered at these clusters. The auxiliary filter may be realized as an adaptive delay filter (ADF) [8, 9] or a full-tap adaptive filter which is operated at a reduced sampling rate [10]. In other algorithms, the dispersive regions are detected through the process of convergence. Stochastic Taps NLMS (STNLMS) [11] and its two variants [12, 13] locate the active region in a stochastic manner. Select and queue with a constraint (SELQUE) algorithm [14] categorizes all taps into two groups: active taps and inactive taps, the latter of which are kept in two queues. The active tap with the minimum absolute coefficient value is replaced by a tap in the queue that is exclusively used for inactive tap indexes residing in the constrained region. An improved M-SELQUE algorithm [15] is applicable to identify an

unknown number of multiple dispersive regions. Furthermore, region-based wavelet-packet adaptive algorithm (RBWP) [16] detects the active taps in transform domain and has been shown to be specially effective. In the above mentioned algorithms, the active taps are first located then estimated. The explicitly separated two steps may decelerate the convergence rate and reduce the robustness.

This work focuses on the fast identification of block-sparse systems via the framework of adaptive filter. Inspired by a sparsity constraint adaptive algorithm named l_0-LMS [17], we propose block-sparse LMS (BS-LMS) to improve the performance of block-sparse system identification. In l_0-LMS, the gradient descent of filter tap-weights are adjusted by approximated l_0 norm constraint to learn a general sparse system response. However, it does not utilize the prior knowledge of block-sparsity and has no particular gain when it is applied to identify a clustering sparse system. Motivated by this, we improve l_0 norm in the cost function to mixed $l_{2,0}$ norm with equal group partition sizes and exert the sparsity constraint in partitions. We then propose a Markov-Gaussian (M-G) model to generate and describe block-sparse systems. Based on this model, theoretical analysis on the proposed algorithm is conducted. It is proved that BS-LMS outperforms l_0-LMS, when the partition size is appropriately chosen.

This paper is organized as follows. Related works are briefly reviewed in Section 2. BS-LMS is proposed in Section 3. The theoretical results on steady-state performance and convergence behavior of BS-LMS are presented in Section 4. The Markov-Gaussian model for generating block-sparse system is proposed and studied in Section 5. The key part of this work goes in Section 6, where the optimal group partition size is studied and superior performance of BS-LMS compared to l_0-LMS is theoretically explained based on the proposed M-G model. Numerical experiments are implemented to verify the above theoretical results in Section 7. The conclusion is drawn in Section 8.

2 Related Work

In this section, we briefly review the available (block)-sparsity-constraint-based adaptive algorithms, which are highly relevant to the proposed BS-LMS, from various approaches.

2.1 Sparsity-Constraint LMS

The identification of an unknown system with sparse impulse response can be accelerated and enhanced by introducing a sparsity constraint into the cost function of LMS, where the sparsity constraint can be approximated l_0 norm [17], l_1 norm [18], reweighted l_1 norm [18, 19], smoothed l_0 norm [20, 21], l_p norm [22, 23], or a convex sparsity penalty [24]. However, literature on adaptive filtering algorithms benefiting from block-sparsity is scarce. Thus, it is important to further improve the performance by utilizing block structure. Among the above algorithms, l_0-LMS [17] demonstrates rather good performance in experiments and

has comprehensive theoretical guarantee [25]. Therefore, in this work we generalize l_0-LMS to BS-LMS by utilizing block-sparsity. Part of our derivations (mainly in Section 4) are based on the approach in [25]. However, the main contributions of this paper, including BS-LMS algorithm (Section 3), the Markov-Gaussian block-sparse model (Section 5), and part of the performance analysis (Section 6.2), are brand-new compared to the above references.

2.2 Block-Sparse Signal Recovery

The idea of using mixed norm, such as $l_{2,1}$ norm [26–28], approximated $l_{2,0}$ norm [29], $l_{q,1}$ norm [30], to handle block-sparsity has been adopted in sparse signal recovery. By exploiting block structure, recovery may be possible under more general conditions, which demonstrates superior performance brought about by mixed norm. Furthermore, after mixed norm is introduced, the reconstruction error in the presence of noise becomes smaller compared with the conventional algorithms. Besides mixed norm, there are some other approaches in block-sparse signal recovery, including greedy algorithms [31–34], Bayesian CS framework-based algorithms [35, 36], the dynamic programming-based algorithm [37] and the decoding-based algorithm [38].

2.3 Group Sparsity Cognizant RLS

Recursive least squares (RLS) is another important branch in adaptive filtering. Its faster convergence rate compared to LMS makes RLS an intriguing adaptive paradigm. In [39], group sparsity cognizant RLS is proposed by using various mixed norms, including $l_{2,1}$ norm, $l_{1,1}$ norm, $l_{2,0}$ norm, and $l_{1,0}$ norm. Numerical experiments show that the novel group sparse RLS is effective and robust for the block-sparse system identification problem, and provides improved performance when compared to the references that only exploit sparsity.

2.4 Group Partition Selection

In some of above references [26–32, 34, 38, 39], it is assumed that the dispersive active regions are located randomly in known partition groups. One may see Fig. 1(d), where two actives regions are denoted by hatching. However, one may readily accept that this assumption is impracticable in real scenarios. In fact, the location of each cluster is arbitrary and totally unknown. In this paper, we utilize mixed $l_{2,0}$ norm in which all of the group partition sizes are the same for practice (Fig. 1(e, f)). Furthermore, in order to avoid the confusion of blocks in unknown system response and the partition blocks in adaptive tap-weights, we adopt *block* or *cluster* to indicate the system coefficient blocks and *group* to denote the partitions in adaptive tap-weights. Based on the theoretical analysis, we will further study the optimal group partition size and demonstrate that the proposed algorithm with an appropriate group partition size achieves superior performance than l_0-LMS.

3 Block-Sparse LMS

The unknown coefficients to be identified and the input signal at time instant n are denoted, respectively, by $\mathbf{s} = [s_1, s_2, \cdots, s_L]^{\mathrm{T}}$ and $\mathbf{x}_n = [x_n, x_{n-1}, \cdots, x_{n-L+1}]^{\mathrm{T}}$, where both $\mathbf{s}$ and $\mathbf{x}_n$ are real-valued vectors, L is the length of the unknown system, and $(\cdot)^{\mathrm{T}}$ represents the transposition. The observed output signal is

$$d_n = \mathbf{x}_n^{\mathrm{T}}\mathbf{s} + v_n, \tag{1}$$

where v_n denotes the measurement noise. The estimated error between the output of the unknown system and of the adaptive filter is

$$e_n = d_n - \mathbf{x}_n^{\mathrm{T}}\mathbf{w}_n, \tag{2}$$

where $\mathbf{w}_n = [w_{1,n}, w_{2,n}, \cdots, w_{L,n}]^{\mathrm{T}}$ denotes the adaptive tap-weights.

Motivated by the practical scenarios where the unknown coefficients appear in blocks rather than being arbitrarily spread, we adopt mixed $l_{2,0}$ norm to evaluate block-sparsity of a vector $\mathbf{w} = [w_1, w_2, \cdots, w_L]^{\mathrm{T}}$ as

$$\|\mathbf{w}\|_{2,0} \triangleq \left\| \begin{bmatrix} \|\mathbf{w}_{[1]}\|_2 \\ \|\mathbf{w}_{[2]}\|_2 \\ \vdots \\ \|\mathbf{w}_{[N]}\|_2 \end{bmatrix} \right\|_0, \tag{3}$$

where $\mathbf{w}_{[i]} = \left[w_{(i-1)P+1}, w_{(i-1)P+2}, \cdots, w_{iP}\right]^{\mathrm{T}}$ denotes the ith group of $\mathbf{w}$. N and P denote the number of groups and the group partition size, respectively. We further assume that L can always be divided evenly by P as several zero taps can be added to the tail of $\mathbf{w}$.

In order to learn the unknown system by utilizing the prior block-sparsity, we design a new cost function, which combines the expectation of the estimated error and mixed $l_{2,0}$ norm of tap-weight vector,

$$\xi_n \triangleq \mathrm{E}\left\{|e_n|^2\right\} + \lambda \|\mathbf{w}_n\|_{2,0}, \tag{4}$$

where λ is a positive factor to balance the mean square error and the penalty of block-sparsity. Following the similar derivations as those for l_0-LMS [17], we arrive at the new recursion of the adaptive tap-weights

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu e_n \mathbf{x}_n + \kappa \mathbf{g}(\mathbf{w}_n), \tag{5}$$

where μ denotes the step-size, $\kappa = \mu\lambda/2$ adjusts the intensity of block-sparse penalty for given step-size, group zero-point attraction (GZA)

$$\mathbf{g}(\mathbf{w}) \triangleq \left[g_1(\mathbf{w}), g_2(\mathbf{w}), \cdots, g_L(\mathbf{w})\right]^{\mathrm{T}},$$

and

$$g_j(\mathbf{w}) \triangleq \begin{cases} 2\alpha^2 w_j - \dfrac{2\alpha w_j}{\left\|\mathbf{w}_{[\lceil j/P \rceil]}\right\|_2}, & 0 < \left\|\mathbf{w}_{[\lceil j/P \rceil]}\right\|_2 \leqslant 1/\alpha; \\ 0, & \text{elsewhere}, \end{cases} \tag{6}$$

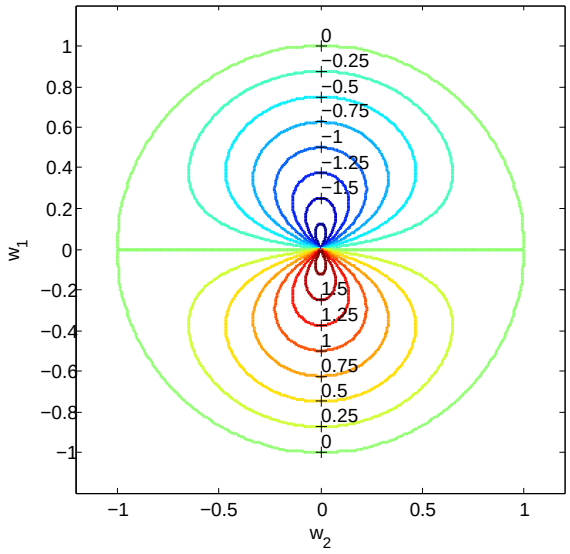

Figure 2: The visualization of an entry in group zero-point attraction, i.e., $g_1([w_1, w_2]^{\mathrm{T}})$, where $L = P = 2$ and $\alpha = 1$.

where α is a positive constant and $\lceil \cdot \rceil$ denotes ceiling function. For $L = P = 2$ and $\alpha = 1$, $g_1([w_1, w_2]^{\mathrm{T}})$ is plotted in Fig. 2 for visualization. One may read that when $w_2 = 0$, the group zero-point attraction for w_1 exactly behaves like its counterpart in l_0-LMS [17]. However, by the increasing of w_2, which means that the block that w_1 locates in seems to be a nonzero block, the attraction of w_1 decreases quickly. In order to avoid being divided by zero, a small positive constant δ is inserted into the denominator of (6) in real implementation. The detailed algorithm is described in Table 1.

Based on the procedure of BS-LMS listed in Table 1, we can see that the BS-LMS algorithm has a computational complexity of $O(L)$, where the detailed computations per iteration is listed and compared with LMS and l_0-LMS in Table 2. One may readily check that, roughly speaking, the complexity of BS-LMS is comparable to that of l_0-LMS. In real applications, partial update can be used to save computations because the slow varying of the unknown system. This is beyond the scope of this work and is not discussed. Readers may refer to [40] for partial updating the gradient item and [17] for the zero-point attraction.

4 Performance of BS-LMS for General Sparse Systems

In this section, we follow the study in [25] and generalize the theoretical results of l_0-LMS to that of the proposed BS-LMS. Most conclusions in this section share the similar formulation as their counterparts of l_0-LMS, though the constants inside the conclusions are quite different. To save space, the details of assumptions and derivations are omitted, while the *new* constants are listed in subsection 4.2 for reference. However, it should be

Table 1: The Procedure of BS-LMS.

Input: $\{x_n, d_n\}_{n=0,1,2,\cdots}$, L, P, μ, α, κ, δ;

Output: $\{\mathbf{w}_n\}_{n=0,1,2,\cdots}$.

Initialization: $\mathbf{w}_0 = \mathbf{0}, N = L/P$.

for $n = 0, 1, 2, \cdots$

$e_n = d_n - \mathbf{x}_n^{\mathrm{T}}\mathbf{w}_n$;

for $i = 1, 2, \cdots, N$

$$E_i = \left(\sum_{j=(i-1)P+1}^{iP} |w_{j,n}|^2 \right)^{1/2};$$

end for

for $j = 1, 2, \cdots, L$

$$g_j(\mathbf{w}_n) = 2\alpha^2 w_{j,n} - 2\alpha w_{j,n} \max\left(\frac{1}{E_{\lceil j/P \rceil} + \delta}, \alpha \right);$$

$$w_{j,n+1} = w_{j,n} + \mu e_n x_{n-j+1} + \kappa g_j(\mathbf{w}_n);$$

end for

end for

emphasized that conducting the complicated derivations where the non-unit partition size P is introduced is the main contribution of this section.

4.1 Assumptions

Following the approach in [25], we classify the unknown coefficients, correspondingly, the adaptive tap-weights, into three categories in group-partition-wise as

$$\begin{aligned}
\text{Large coefficients:} \quad & \mathcal{C}_{\mathrm{L}}(P) \triangleq \left\{ j \middle| \left\| \mathbf{s}_{[\lceil j/P \rceil]} \right\|_2 \geq 1/\alpha \right\}, \\
\text{Small coefficients:} \quad & \mathcal{C}_{\mathrm{S}}(P) \triangleq \left\{ j \middle| 0 < \left\| \mathbf{s}_{[\lceil j/P \rceil]} \right\|_2 < 1/\alpha \right\}, \\
\text{Zero coefficients:} \quad & \mathcal{C}_0(P) \triangleq \left\{ j \middle| \left\| \mathbf{s}_{[\lceil j/P \rceil]} \right\|_2 = 0 \right\}.
\end{aligned}$$

We further denote the number of tap-weights belonging to the nonzero group partitions by $Q(P) \triangleq |\mathcal{C}_{\mathrm{L}}(P) \cup \mathcal{C}_{\mathrm{S}}(P)|$, which is also termed the number of nonzero coefficients. However, one should recognize that some zero coefficients may be counted as nonzero coefficients, which is so called *border effect*, (Fig. 1(e, f), where the active regions and border effect are denoted by hatching and cross-hatching, respectively), and will be studied in next section. Comparing to those defined in [25], one may notice that the above introduced coefficients are closely dependent on the group partition size. Without confusion, however, they are sometimes abbreviated to $\mathcal{C}_{\mathrm{L}}, \mathcal{C}_{\mathrm{S}}, \mathcal{C}_0$, and Q.

Table 2: The computational cost of LMS, l_0-LMS, and BS-LMS.

Algorithm	LMS	l_0-LMS	BS-LMS
MAC	$2L$	$4L$	$5L$
Multiplication	-	L	L
Division	-	-	L/P
Comparison	-	$2L$	L/P
Square root	-	-	L/P

We proceed to use the assumptions on the system response and the predefined parameters that raised in [25] and listed as follows, where $\mathcal{C}_{\mathrm{L}}$ and $\mathcal{C}_{\mathrm{S}}$ here have different meanings from their counterparts.

(i) Input signal x_n follows *i.i.d.* zero-mean Gaussian distribution.

(ii) Tap-weights $\mathbf{w}_n$, input vector $\mathbf{x}_n$, and additive noise v_n are mutually independent.

(iii) The parameter κ is so small that $2\alpha^2\kappa \ll \mu\sigma_x^2$ can be guaranteed, where σ_x^2 denotes the variance of x_n.

(iv) All tap-weights, $\mathbf{w}_n$, may be modeled as Gaussian variables.

(v) For $j \in \mathcal{C}_{\mathrm{L}} \cup \mathcal{C}_{\mathrm{S}}$, the tap-weight $w_{j,n}$ is assumed to have the same sign with the corresponding unknown coefficient.

(vi) The adaptive weight $w_{j,n}$ is assumed out of the attraction range for $j \in \mathcal{C}_{\mathrm{L}}$, while in the attraction range elsewhere.

We could demonstrate that the above assumptions still hold because the new recursion does not destroy their validity. We further propose another assumption to make the analysis of BS-LMS feasible.

(vii) The difference between the relative strength of $w_{j,n}$ and that of s_j in $\mathcal{C}_{\mathrm{S}}$ is small enough to ratify the following approximation,

$$\frac{w_{j,n}}{\left\|\mathbf{w}_{[\lceil j/P \rceil],n}\right\|_2} \approx \frac{s_j}{\left\|\mathbf{s}_{[\lceil j/P \rceil]}\right\|_2}, \quad \forall j \in \mathcal{C}_{\mathrm{S}}.$$

This assumption is considered proper due to the following reason. It is readily accepted that in traditional LMS the tap-weights of $w_{j,n}$ uniformly converge to their optimal values with *i.i.d.* white Gaussian input. In the proposed BS-LMS, because of group zero-point attraction in (5), the uniform convergence may not exist in a global manner, but may be available inside each group. Therefore, the temporary tap-weight and the unknown

coefficient with respect to their strengths in group are supposed very close. In fact, the numerical experiment has verified that this assumption always remains valid, especially in high SNR scenarios.

4.2 Notations and Expressions

All through this paper, we have

$$\Delta_L \triangleq 2-(L+2)\mu\sigma_x^2, \qquad \Delta_Q \triangleq 2-(Q+2)\mu\sigma_x^2, \tag{7}$$

$$\Delta_0 \triangleq 1-\mu\sigma_x^2, \qquad \Delta_0' \triangleq 2-\mu\sigma_x^2, \tag{8}$$

$$G(\mathbf{s}) \triangleq \langle \mathbf{g}(\mathbf{s}), \mathbf{g}(\mathbf{s}) \rangle = \sum_{k\in\mathcal{C}_\mathrm{S}} g_k^2(\mathbf{s}), \tag{9}$$

$$G'(\mathbf{s}) \triangleq \langle \mathbf{s}, \mathbf{g}(\mathbf{s}) \rangle = \sum_{k\in\mathcal{C}_\mathrm{S}} s_k g_k(\mathbf{s}). \tag{10}$$

In Lemma 1, the constants $\{\beta_i\}$ are

$$\beta_0 \triangleq \mu\sigma_x^2\Delta_0'\Delta_L G(\mathbf{s}) + 4\alpha^2\Delta_Q\left(\frac{\mu\sigma_x^2\Delta_L}{P} + \frac{\Delta_0\Delta_Q\theta^2}{\pi}\right), \tag{11}$$

$$\beta_1 \triangleq \frac{\Delta_0' G(\mathbf{s}) + 4(L-Q)\alpha^2\left(\frac{\mu\sigma_x^2}{P} + \frac{2\Delta_0\Delta_Q\theta^2}{\pi\Delta_L}\right)}{\mu^2\sigma_x^4\Delta_L}, \tag{12}$$

$$\beta_2 \triangleq \frac{4\alpha\,(L-Q)\,\theta}{\mu^2\sigma_x^4\Delta_L^2}\sqrt{\frac{\Delta_0\beta_0}{\pi}}, \tag{13}$$

$$\beta_3 \triangleq 2\mu^3\sigma_x^4\sigma_v^2\Delta_0\Delta_L/\beta_0, \tag{14}$$

where θ is defined as

$$\theta \triangleq \begin{cases} \dfrac{[((P-1)/2)!]^2\,2^{P-1}}{P!}, & P \text{ is odd;} \\ \dfrac{(P-1)!\pi}{(P/2)!(P/2-1)!2^P}, & P \text{ is even.} \end{cases} \tag{15}$$

In Lemma 3, $\mathbf{A} = \{a_{ij}\}$ is defined as

$$\mathbf{A} \triangleq \begin{bmatrix} 1-\mu\sigma_x^2\Delta_L & -\theta\sqrt{\frac{8}{\pi}}\frac{\kappa\alpha\Delta_0}{\omega} \\ (L-Q)\,\mu^2\sigma_x^4 & 1-2\mu\sigma_x^2\Delta_0 - \theta\sqrt{\frac{8}{\pi}}\frac{\kappa\alpha\Delta_0}{\omega} \end{bmatrix} \tag{16}$$

and $\mathbf{b}_n$ is used in the proof (please refer to [25]),

$$\mathbf{b}_n \triangleq [b_{0,n}, b_{1,n}]^\mathrm{T}, \tag{17}$$

where

$$b_{0,n} \triangleq L\mu^2\sigma_x^2\sigma_v^2 + (L-Q)\left(\frac{4\alpha^2\kappa^2}{P} - \theta\sqrt{\frac{8}{\pi}}\kappa\Delta_0\alpha\omega\right) + \frac{\kappa^2\left(\Delta_0' - 2\Delta_0^{n+1}\right)}{\mu\sigma_x^2}G(\mathbf{s}) - 2\kappa\Delta_0^{n+1}G'(\mathbf{s}), \tag{18}$$

$$b_{1,n} \triangleq (L-Q)\left(\mu^2\sigma_x^2\sigma_v^2 + \frac{4\alpha^2\kappa^2}{P} - \theta\sqrt{\frac{8}{\pi}}\kappa\Delta_0\alpha\omega\right). \tag{19}$$

Please notice that in (16), (18), and (19), ω is the solution of

$$2\mu\sigma_x^2\Delta_0\Delta_L\omega^2+\frac{8\kappa\alpha\theta\Delta_0\Delta_Q}{\sqrt{2\pi}}\omega-2\mu^2\sigma_x^2\sigma_v^2\Delta_0-\kappa^2\left(\frac{4\alpha^2\Delta_Q}{P}+G(\mathbf{s})\Delta_0'\right)=0. \tag{20}$$

In Lemma 3, the constants λ_3 and c_3 are

$$\lambda_3 \triangleq \Delta_0, \tag{21}$$

$$c_3 \triangleq -\frac{2\kappa\Delta_0}{\mu\sigma_x^2}\frac{\mu\sigma_x^2-2\mu^2\sigma_x^4+\theta\sqrt{\frac{8}{\pi}}\frac{\kappa\alpha\Delta_0}{\omega}}{\det\left(\lambda_3\mathbf{I}-\mathbf{A}\right)}\left(\kappa G(\mathbf{s})+\mu\sigma_x^2G'(\mathbf{s})\right). \tag{22}$$

4.3 Steady-State Misalignment and Transient Behavior

Defining $\mathbf{h}_n \triangleq \mathbf{w}_n-\mathbf{s}$ as the misalignment of tap-weights and following the similar approach in [25], the bias in steady state can be derived,

$$\overline{h_{j,\infty}} \triangleq \lim_{n\to\infty}\overline{h_{j,n}}=\frac{\kappa}{\mu\sigma_x^2}g_j\left(\mathbf{s}\right), \quad \forall j=1,2,\cdots,L, \tag{23}$$

where *overline* denotes taking expectation. According to (6), one may find that the tap-weights are unbiased for large and zero group coefficients, while they are biased for small group coefficients.

Lemma 1 (the counterpart of Theorem 1 in [25]) *The steady-state mean square deviation (MSD) of BS-LMS is*

$$\begin{aligned} D_\infty &\triangleq \lim_{n\to\infty}D_n \triangleq \lim_{n\to\infty}\overline{\mathbf{h}_n^{\mathrm{T}}\mathbf{h}_n} \\ &=\frac{\mu\sigma_v^2L}{\Delta_L}+\beta_1\kappa^2-\beta_2\kappa\sqrt{\kappa^2+\beta_3}, \end{aligned} \tag{24}$$

where σ_v^2 denotes the variance of measurement noise, Δ_L is defined in (7), $\{\beta_i\}_{i=1,2,3}$ are defined in (12), (13), and (14) in Appendix 4.2, respectively. The step-size should satisfy

$$0<\mu<\mu_{\max}\triangleq\frac{2}{\left(L+2\right)\sigma_x^2} \tag{25}$$

to guarantee convergence .

Lemma 2 (the counterpart of Corollary 1 in [25]) *In order to make the steady-state MSD be as small as possible, the best choice for κ is*

$$\kappa_{\mathrm{opt}}=\frac{\sqrt{\beta_3}}{2}\left(\sqrt[4]{\frac{\beta_1+\beta_2}{\beta_1-\beta_2}}-\sqrt[4]{\frac{\beta_1-\beta_2}{\beta_1+\beta_2}}\right) \tag{26}$$

and the minimum steady-state MSD is

$$D_\infty^{\min}=\frac{\mu\sigma_v^2L}{\Delta_L}+\frac{\beta_3}{2}\left(\sqrt{\beta_1^2-\beta_2^2}-\beta_1\right). \tag{27}$$

Lemma 3 (the counterpart of Theorem 2 in [25]) *For a given unknown system, the closed form of instantaneous MSD is*

$$D_n = c_1\lambda_1^n + c_2\lambda_2^n + c_3\lambda_3^n + D_\infty, \tag{28}$$

where λ_1 and λ_2 are the eigenvalues of matrix $\mathbf{A}$, which is defined in (16). c_1 and c_2 are coefficients defined by initial value (please refer to Lemma 1 in [25]). The expressions of constants λ_3 and c_3 are listed in (21) *and* (22)*, respectively, in Appendix 4.2.*

Remark 1 *Based on the above lemmas, we have successfully generalize the theoretical results of l_0-LMS to BS-LMS. As we have mentioned, most of their formulations are exactly identical, whereas the constants included are rather different. To totally understand the above contents, the readers are recommended to refer to [25] and compare those constants in Appendix A of [25] with those in subsection 4.2 of this paper.*

Because of the sophisticated formulation of (27), the statistical characteristic of steady-state MSD is difficult to analyze, when the unknown block-sparse impulse response follows specific probability distribution. After providing an approximated but simple expression of (27) in the following lemma, we have finally prepared to comprehensively study the performance of BS-LMS.

Lemma 4 *When κ_{opt} in* (26) *is adopted, the minimum steady-state MSD in* (27) *can be approximated by*

$$D_\infty^{\min} \approx \frac{\mu\sigma_v^2}{\Delta_Q}\left(Q + \frac{\sqrt{2\pi(L-Q)G(\mathbf{s})}}{\alpha\theta\Delta_Q}\right), \tag{29}$$

where $\Delta_Q, G(\mathbf{s})$, and θ are defined in (7)*,* (9)*, and* (15)*, respectively.*

PROOF The proof is postponed to Appendix 9.1. ∎

5 Markov-Gaussian Model for Generating Block-Sparse Systems

Inspired by the characteristic of nonzero (or zero) coefficients clustering in blocks, we propose a Markov-Gaussian (M-G) model with parameter set, $\mathcal{M}(L, p_1, p_2, \sigma_s^2)$, to generate a wide range of block-sparse systems. One will notice that the proposed one is a simplified Ising model [42] that fits to the scenario in this study.

Utilizing the proposed model, the impulse response $\mathbf{s}$ of a block-sparse system is generated in two steps. In the first step, the zero and nonzero sets which contain the index of zero coefficients and nonzero coefficients, respectively, are produced by a Markov process. From 1 to L, index j is iteratively and stochastically determined to fall into zero or nonzero sets

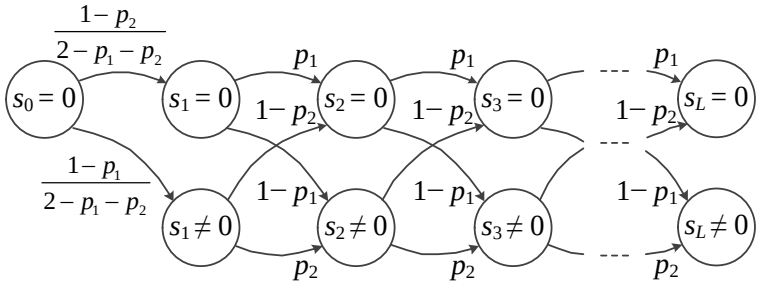

Figure 3: The proposed model for generating block-sparse impulse response.

based on the class of index $(j-1)$. Please refer to Fig. 3 for detail, where for $j = 2, 3, \cdots, L$

$$\mathrm{P}\{s_j = 0 | s_{j-1} = 0\} = p_1,$$
$$\mathrm{P}\{s_j \neq 0 | s_{j-1} \neq 0\} = p_2,$$

and the category of s_1 is decided by s_0, which is an imaginative scaler and fixed to zero. In the second step, after the nonzero set is determined, the amplitudes of nonzero coefficients are independently and identically drawn from a Gaussian distribution with zero mean and variance σ_s^2.

According to Fig. 3, for the sake of producing a sparse system response, it should be guaranteed that $(1-p_2)$ is far larger than $(1-p_1)$. According to this model, a coefficient probably has the same property, i.e., be zero or be nonzero, as its precursor when both p_1 and p_2 are large. Therefore, both p_1 and p_2 need to be chosen close to 1 for generating a clustering impulse response. In the following text, several properties of the proposed model are provided, which will be used in the following section to simplify the theoretical analysis, especially for analyzing the effect of P on the steady state MSD.

Property 1 *(*Sparsity and block size*) For given $\mathcal{M}(L, p_1, p_2, \sigma_s^2)$, the average percentage of nonzero coefficients, the average block size of nonzero and zero coefficients of the generated impulse responses, which are denoted by $\overline{S}, \overline{B_{\mathrm{nz}}}$, and $\overline{B_{\mathrm{z}}}$, respectively, follow*

$$\overline{S} = \frac{1-p_1}{2-p_1-p_2}, \quad \overline{B_{\mathrm{nz}}} = \frac{1}{1-p_2}, \quad \textit{and} \quad \overline{B_{\mathrm{z}}} = \frac{1}{1-p_1}.$$

Several examples of block-sparse systems generated by $L = 800, \sigma_s^2 = 1$, and various (p_1, p_2) are showed in Fig. 4. Inside every row and every column, the average block size of zero and nonzero coefficients increase, respectively, with respect to p_1 and p_2. Moreover, the three responses located on the diagonal subplots satisfy $\overline{S} = 0.1$ and have 80 expected nonzero coefficients.

Property 2 *(*Border effect*) For given $\mathcal{M}(L, p_1, p_2, \sigma_s^2)$ and a predefined group partition size P, we can calculate the average number of tap-weights belonging to nonzero groups, $\overline{Q}$, to describe the intensity of the border effect,*

$$\overline{Q} = L\left(1 - (1-\overline{S})p_1^{P-1}\right). \tag{30}$$

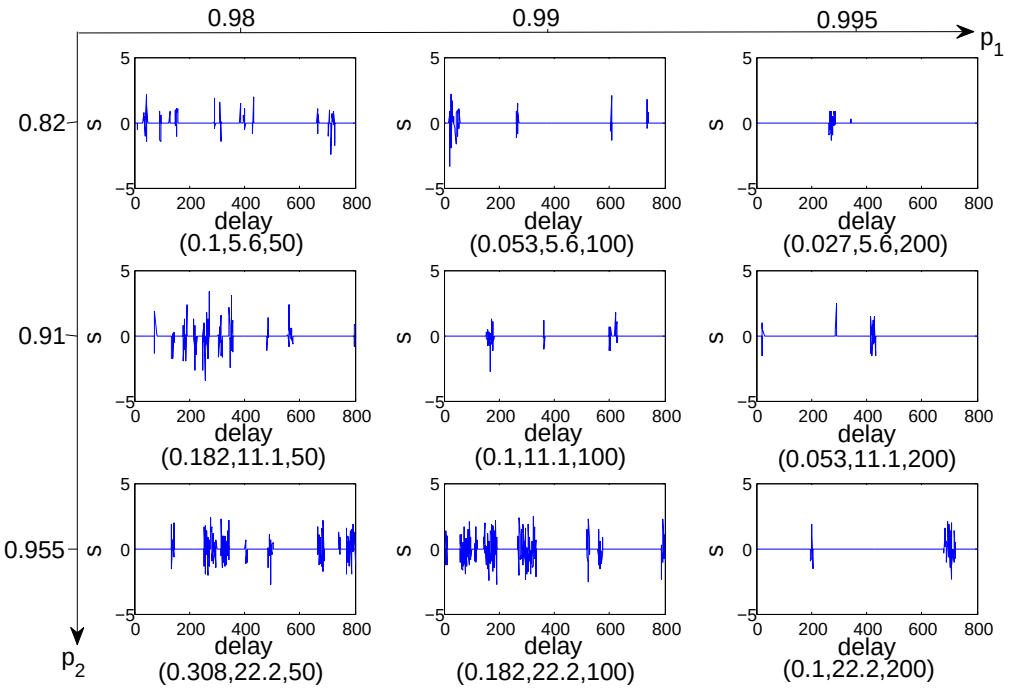

Figure 4: The examples of block-sparse impulse responses generated by the proposed M-G model with $L = 800, \sigma_s^2 = 1$ and various (p_1, p_2). $(\overline{S}, \overline{B_{\text{nz}}}, \overline{B_{\text{z}}})$ with corresponding parameter set is listed below each subfigure. The three systems located on the diagonal subplots share the same sparsity.

PROOF The proof is postponed to Appendix 9.2. ∎

According to this property, one may find that $\overline{Q}$ becomes larger when P increases, which shows the border effect is heavier.

Property 3 *(*Relation with Ising model*) The proposed Markov model for determining zero and nonzero coefficients sets is a special case of the Ising model, which is a prototypical Markov random field.*

As far as we know, this is the first time that Markov process is used to describe block-sparsity for adaptive filtering. The proposed M-G model may be utilized in various research area to generate arbitrary system response with given block-sparse constraint. There are still topics need to be studied on the M-G model. For example, it is important to build the relation between the model and the real-world applications, and to estimate the parameters from specific scenario. However, this is beyond the scope of this work and is not included.

6 Performance of BS-LMS for Block-Sparse Systems

The behavior of BS-LMS in block-sparse scenario is further studied in this section by utilizing the proposed M-G model. New assumptions are adopted as follows.

(viii) For a given unknown system, which is supposed to be long and sparse, the partition size P is small with respect to the filter length to guarantee that the system response in group-partition-wise is still sparse, i.e., $2 \ll Q(P) \ll L$.

(ix) The unknown system response to be identified is generated by the proposed M-G model, $\mathcal{M}(L, p_1, p_2, \sigma_s^2)$.

Assumption (viii) makes sense because P is an important predefined parameter that need to be elaborately selected. Furthermore, it can be accepted that BS-LMS penalizes sparsity in group-partition-wise and an overlarge P definitely destroys the sparsity. One may refer to Fig. 1, where an overlarge P in (e) aggravates the border effect, denoted by cross-hatching, compared to that in (f).

The introduction of M-G model by Assumption (ix) makes it feasible to analyze BS-LMS with respect to block-sparse systems. As a consequence, we will study the performance of BS-LMS in the sense of *expected* unknown system response, which is generated by the given M-G model with specified parameters. Therefore, the average minimum steady-state MSD (AMS-MSD), the optimal group partition size, and the average minimum transient MSD (AMT-MSD), denoted as $\overline{D_\infty^{\min}}$, P_{opt}, and $\overline{D_n^{\min}}$, respectively, will be derived in the following text.

Finally we demonstrate that BS-LMS outperforms l_0-LMS in convergence rate significantly when the group partition size is chosen close to its optimum.

6.1 Steady-State Performance and Optimal Group Partition Size

The following theorem presents the effect of the group partition size on the AMS-MSD and the selection of the optimal partition size.

Theorem 1 *For given block-sparse systems generated by the proposed M-G model, the AMS-MSD of BS-LMS is*

$$\overline{D_\infty^{\min}} \approx \frac{\mu\sigma_v^2}{\Delta_{\overline{Q}}} \left(\overline{Q} + \frac{\sqrt{2\pi(L-\overline{Q})\overline{G(\mathbf{s})}}}{\alpha\theta\Delta_{\overline{Q}}} \right). \tag{31}$$

where $\overline{Q}, \Delta_{\overline{Q}}$, and $\overline{G(\mathbf{s})}$ are defined in (30)*, (46), and (49), respectively. The optimal group partition size can be numerically found by*

$$P_{\text{opt}} = \arg\min_P \overline{D_\infty^{\min}}. \tag{32}$$

Proof The proof is postponed to Appendix 9.3. ∎

Corollary 1 *The AMS-MSD monotonically increases with respect to the step-size.*

Corollary 1 is coincident with the intuition and the theory on l_0-LMS. This can be readily seen from (31) because $\Delta_{\overline{Q}}$ monotonically decreases with respect to μ.

Remark 2 *As l_0-LMS is a special case of BS-LMS when P equals 1, we expect that the AMS-MSD of BS-LMS with P_{opt} is no larger than that of l_0-LMS when all the other parameters are same.*

Considering the high complexity of the closed form of P_{opt}, it is not derived here for the sake of simplicity. Nonetheless, the qualitative relationship between P_{opt} and parameters p_1 and p_2 can be obtained.

Remark 3 *In general, P_{opt} increases monotonously with p_1 and p_2. For the convenience of analysis, here we only discuss the condition that a family of system coefficients have the same sparsity. From Property 1 and 2, we know that the strengthening rate of border effect with respect to P becomes smaller as p_1 and p_2 increase. Besides, we know that the increase of P makes the improper effect, which means that some tap-weights corresponding to small coefficients may be wrongly led to an unnecessary bias and undergo a quite large deviation because of (23), weaker and the border effect stronger. Both of the effects will make the average steady-state MSD larger. When P increases from 1, the reduction of the average steady-state MSD brought about by weaker improper effect exceeds the increase caused by stronger border effect, thus decreasing the mean steady-state MSD. However, if P is larger than P_{opt}, the reduction becomes smaller than the increase. Therefore P_{opt} is decided by the balance between these two effects. Because the increase of p_1 and p_2 makes the strengthening rate of border effect smaller, we can expect that the two effects are balanced when P is chosen larger, e.g., P_{opt} becomes larger.*

6.2 Superior Performance of BS-LMS

Based on Lemma 3 and Theorem 1, we could demonstrate that the averaged transient behavior of BS-LMS with an appropriate group partition size is better than that of l_0-LMS.

Theorem 2 *For given block-sparse systems generated by the proposed M-G model, the closed form of the AMT-MSD of BS-LMS is*

$$\overline{D_n^{\min}} = c_1'(\lambda_1')^n + c_2'(\lambda_2')^n + c_3'(\lambda_3')^n + \overline{D_\infty^{\min}}, \tag{33}$$

where

$$\lambda_1' \triangleq 1 - 2\mu\sigma_x^2, \tag{34}$$

$$\lambda_2' \triangleq 1 - 2\mu\sigma_x^2\alpha\theta\sqrt{\frac{2\left(L-\overline{Q}\right)}{\pi\overline{G(\mathbf{s})}}}, \tag{35}$$

$$\lambda_3' \triangleq 1 - \mu\sigma_x^2, \tag{36}$$

and the expressions of $\{c_i'\}_{i=1,2,3}$ *share the same forms with* $\{c_i\}_{i=1,2,3}$ *in Lemma 3 except for that* $Q, G(\mathbf{s}), \|\mathbf{s}\|_2^2$, *and* $G'(\mathbf{s})$ *are replaced by their means defined by* (30), (49), (64), *and* (65), *respectively.*

PROOF The proof is postponed to Appendix 9.4. ∎

The difference of Theorem 2 from Lemma 3 is that the former provides an averaged minimum behavior of the best κ with respect to a given unknown system generation model. As a consequence, the close form of $\{\lambda_i'\}$ are provided to reveal the detail of convergence in BS-LMS.

In order to compare the convergence behavior of BS-LMS and l_0-LMS fairly, it is assumed that the final steady-state MSDs of both algorithms are equal. According to Corollary 1 and Remark 2, we know that the step-size in BS-LMS is larger than that in l_0-LMS when the two algorithms demonstrate the same steady-state performance. Then we have the following corollary.

Corollary 2 *For a given M-G model* $\mathcal{M}(L, p_1, p_2, \sigma_s^2)$ *satisfying*

$$\frac{1-p_1}{(1-p_2)^2} \geq \frac{1}{3(1-\mathrm{e}^{-1})}, \tag{37}$$

$\{\lambda_i'\}_{i=1,2,3}$ *in* (33) *with* P_{opt} *are smaller than, respectively, those of* l_0*-LMS, which means that BS-LMS with* P_{opt} *always converges more quickly than* l_0*-LMS.*

PROOF The proof is postponed to Appendix 9.5. ∎

Though we further restrict the selection of p_1 and p_2 to facilitate the proof, it is found that (37) is not necessary. In fact, Corollary 2 is usually valid even when (37) is violated, as shown in numerical results.

7 Numerical Simulations

Five experiments are designed to verify the contents in this paper, where the first two demonstrate the performance of BS-LMS and the M-G model, and the last three test the theoretical analysis. The reference algorithms include STNLMS [11], SELQUE [14], M-SELQUE [15], and l_0-LMS [17]. In all the experiments, the unknown systems are of length $L = 800$. For those generated with the proposed M-G model, σ_s^2 is set as 1. The input signal and measurement noise are independent zero mean Gaussian series, where the variance of the former is unit, i.e., $\sigma_x^2 = 1$. The Signal-to-Noise Ratio (SNR) is 40dB and 20dB in the fifth experiment, while it is 40dB in the others. For both BS-LMS and l_0-LMS, α is chosen as 1. For BS-LMS, δ = 1e-8. Simulation results are averaged by 10 independent trials for each unknown system. To evaluate the performance on a specific M-G model, 100 unknown systems are generated and identified, and then the learning curves of identifying these systems are averaged.

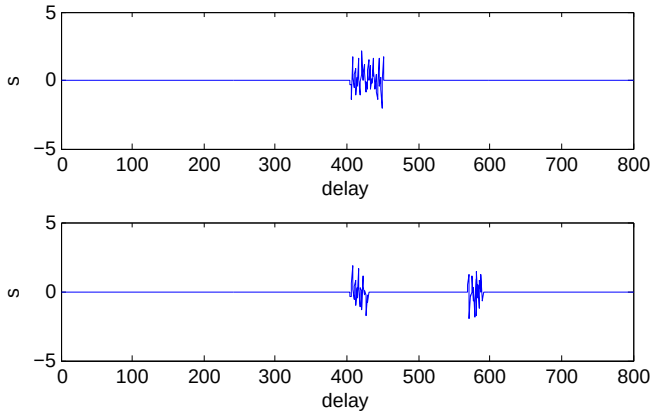

Figure 5: The block-sparse systems tested in the first experiment.

7.1 On the Performance of BS-LMS and M-G Model

In the first experiment, the proposed algorithm is tested and compared with the references by identifying two block-sparse systems which have the same sparsity. The impulse responses of these systems are displayed in Fig. 5, where the first has a single cluster of nonzero coefficients at $[405, 451]$ and the second has two clusters at $[405, 429]$ and $[569, 590]$, respectively. The simulation results are plotted in Fig. 6. For the proposed algorithm, μ is set close to half its largest value, i.e., $1/(L\sigma_x^2)$, κ and P are chosen, respectively, as their optimum values, 1.55e-6 and 5. For all reference algorithms, the step-sizes are selected to make their steady-state MSDs equal to that of BS-LMS, and other parameters are elaborately tuned to produce their fastest convergence.

According to Fig. 6, when there is only one cluster in the system response, SELQUE gets to the steady state the fastest and the convergence performance of BS-LMS is slightly inferior to other block algorithms. Nonetheless, when there exists two clusters, BS-LMS shows its advantage. Because BS-LMS need not detect the active regions, its performance is nearly not affected by the number of clusters. The convergence rates of STNLMS and SELQUE deteriorate significantly, because all of the active regions, along with flat delays between the two clusters, are considered as a long active region. Although M-SELQUE can obviate this problem, its convergence behavior still becomes worse when the unknown system has multi-clusters.

In the second experiment, BS-LMS and the proposed M-G model are tested. The simulation results are plotted in Fig. 7, where (top), (middle), and (bottom) correspond to the unknown systems plotted in the diagonal subfigures of Fig. 4, respectively, from left-top to right-bottom. For BS-LMS, μ is always set close to half the largest value $1/(L\sigma_x^2)$ and other parameters κ and P are set as, respectively, their optimum values 8.66e-7 and 3 for (top), 1.07e-6 and 4 for (middle), 1.60e-6 and 5 for (bottom). For the reference algorithms, the step-size and other parameters are properly adjusted to get their fastest convergence

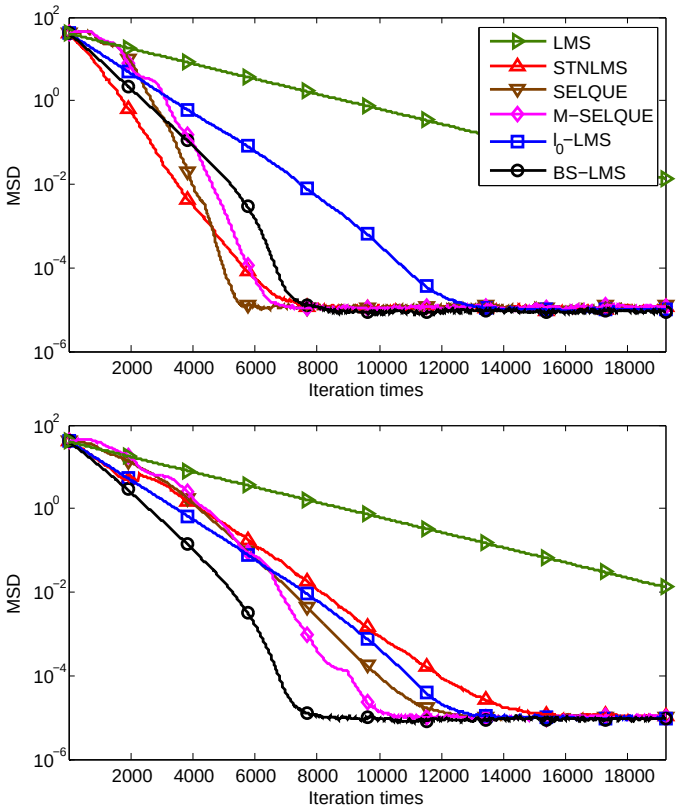

Figure 6: The learning curves of the proposed algorithm and the references when identifying the two block-sparse systems displayed in Fig. 5, where (top) and (bottom) correspond to the single-cluster and multi-cluster system respectively.

rate and equal steady-state MSD with BS-LMS.

According to the simulation results, BS-LMS and l_0-LMS are always among the best when identifying various block-sparse systems, which are generated by the proposed M-G model with various parameter sets. BS-LMS converges faster than l_0-LMS, because the former utilizes the block-sparsity prior. Among all the algorithms based on active region detection, M-SELQUE behaves the best but still converges slower than l_0-LMS, because it takes more iterations to identify the locations of nonzero coefficients and thus reduces the convergence rate when there are more and dispersed clusters, which is highly likely to be produced by utilizing the M-G model. SELQUE and STNLMS get the worst performance because they are not suitable to the multi-cluster system, which has been demonstrated in the first experiment. Above all, we can conclude that BS-LMS has a superior robustness than all reference algorithms, especially in the scenarios of multiple-scattered-cluster sparse systems.

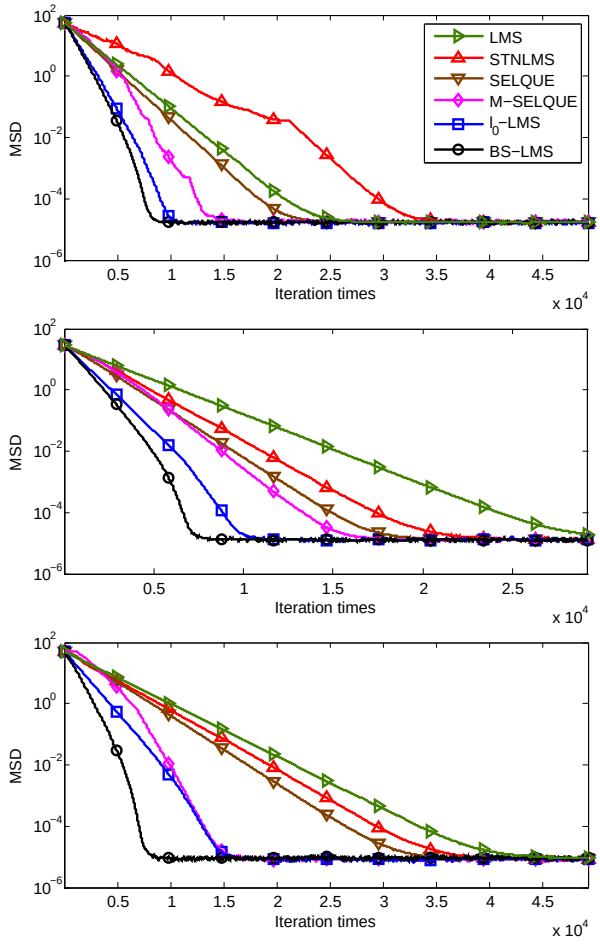

Figure 7: The learning curves of the proposed algorithm and the references when identifying three different unknown systems whose impulse response are plotted in the diagonal subfigures of Fig. 4, where (top), (middle), and (bottom) corresponding to, respectively, the left-top, the middle, and the right-bottom.

7.2 On the Theoretical Results

In the third experiment, the steady-state performance of BS-LMS of different group partition size P with respect to κ is tested. The unknown system response is shown in the center of Fig. 4. The stepsize is fixed to $0.8/(L\sigma_x^2)$. P is chosen as 1, 5, and 20, respectively. For each P, κ varies from 10^{-9} to 10^{-5}.

Referring to Fig. 8, we can see that the theoretical steady-state MSD of BS-LMS agrees well with the simulation result. For every group partition size, as κ increases from 10^{-9}, the steady-state MSD decreases at first, which means that proper GZA is useful to reduce the amplitudes of coefficients in $\mathcal{C}_0$. However, when κ continues to increase, more intense GZA enhances the bias of coefficients in $\mathcal{C}_S$. For different group partition size, the minimum steady-state MSD and its corresponding optimal κ varies. One may recognize that the

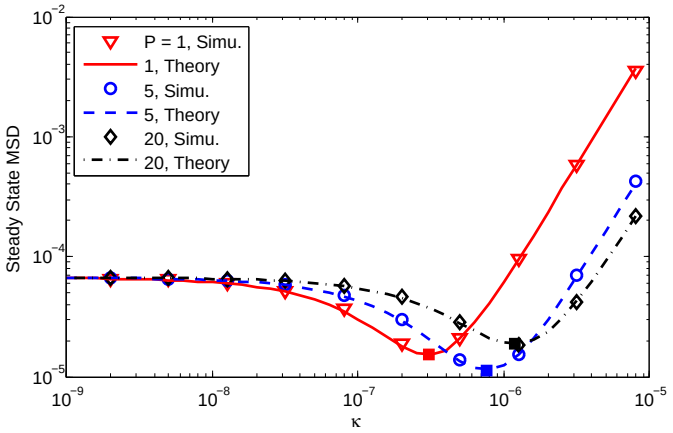

Figure 8: Steady-state MSD of BS-LMS of different group partition size with respect to κ. The solid square denotes the theoretical κ_{opt}.

simulation result of the optimal κ tallies with theoretical κ_{opt} very well.

In the fourth experiment, for given M-G model, the effect of group partition size on the average steady-state MSD is investigated. The model parameter set (p_1, p_2) is chosen as $(0.98, 0.82)$, $(0.99, 0.91)$, and $(0.995, 0.955)$, respectively. Three typical system responses are plotted in the diagonal subfigures of Fig. 4. In this test, μ is chosen as $0.4/(L\sigma_x^2)$ and P varies from 1 to 50. For each P, κ is chosen as the theoretical optimal value.

Please refer to Fig. 9 for the result. According to Fig. 4, with the growth of p_1 and p_2, the average block size of nonzero coefficients increase from 5.6, 11.1, to 22.2. As a consequence, the optimal group partition size P_{opt} also increases from 3, 4, to 5. One may find that for all block-sparse systems, when P initially increases from 1, the minimum MSD decreases quickly at first, which means that treating nonzero coefficients in groups really improves the identification of block-sparse system. However, the minimum MSD increases after P exceeds its optimum, which is much smaller than the average block size, and becomes larger than that of $P = 1$, which demonstrates the severe consequence of border effect. The above results accord well with the intuition. Furthermore, simulation results tally with analytical values, especially when P is small.

In the last experiment, for given M-G model, the theoretical transient behavior of BS-LMS with the optimal group partition size is verified by simulation and compared with that of l_0-LMS. The model parameter set (p_1, p_2) is chosen as $(0.99, 0.91)$. A typical unknown system response is shown in the center of Fig. 4. The SNR are selected as 40dB and 20dB to test the performance in various noisy scenarios. For BS-LMS and l_0-LMS, μ is chosen as $0.637/(L\sigma_x^2)$ and $0.4/(L\sigma_x^2)$, respectively, to make their average steady-state MSDs equal. For both algorithms, P and κ are chosen as their corresponding optimal values.

Please refer to Fig. 10 and Fig. 11. One may readily see that the convergence rate of BS-LMS is always faster than that of l_0-LMS. Furthermore, the theoretical analysis of

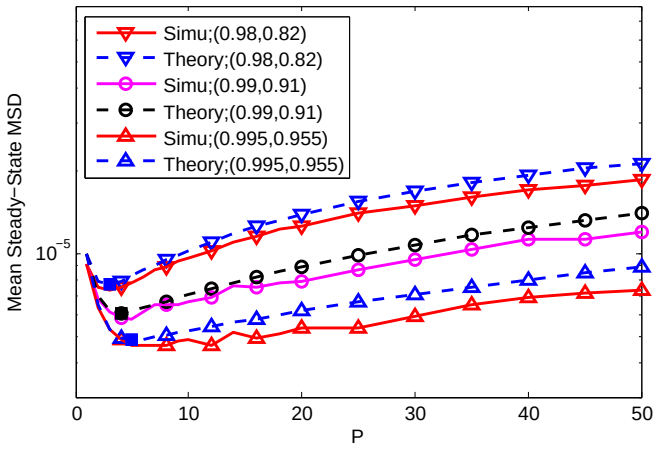

Figure 9: Average steady-state MSD of BS-LMS with respect to different group partition size for given M-G model. κ is chosen as the theoretical optimal and the solid square denotes P_{opt}.

transient behavior accords with simulation in a tolerable error, which origins mainly from large step-size and Assumption (ii), i.e., the famous independence assumption [1].

8 Conclusion

In order to improve the performance of block-sparse system identification, a new algorithm based on l_0-LMS is proposed in this paper by changing l_0 norm to mixed $l_{2,0}$ norm with equal group partition sizes in the cost function. Also, a M-G model is put forward to describe the block-sparse system. Furthermore, the theoretical analysis on performance of BS-LMS compared to l_0-LMS is presented based on the expressions of mean square misalignment, which shows that BS-LMS is better than l_0-LMS theoretically. Finally, simulations are designed to verify the theoretical results and confirm superior performance of our proposed algorithm.

9 Appendix

9.1 Proof of Lemma 4

Proof In order to get the mean steady-state MSD of BS-LMS, we need to simplify the result in Lemma 2. Before giving out approximations, we will prove a useful inequality

$$G(\mathbf{s}) < 4\alpha^2 Q/P. \tag{38}$$

For small group coefficients, according to (6), we know that

$$|g_j(\mathbf{s})| < \frac{2\alpha|s_j|}{\left\|\mathbf{s}_{[\lceil j/P \rceil]}\right\|_2}. \tag{39}$$

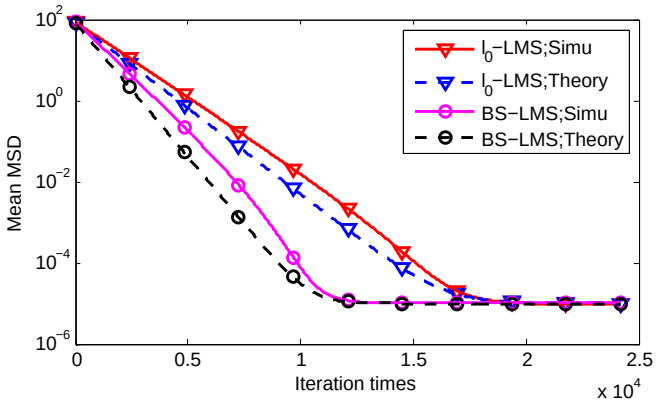

Figure 10: Average learning curves of BS-LMS and l_0-LMS with given M-G model ($p_1 = 0.99$ and $p_2 = 0.91$). P and κ are chosen as the optimal. The SNR is 40dB.

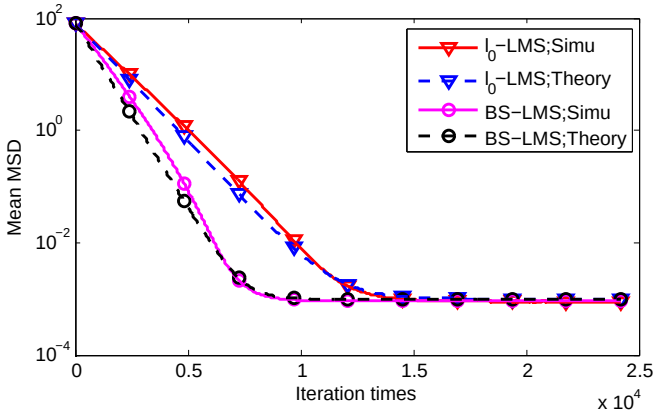

Figure 11: Average learning curves of BS-LMS and l_0-LMS with given M-G model ($p_1 = 0.99$ and $p_2 = 0.91$). P and κ are chosen as the optimal. The SNR is 20dB.

As the number of small groups is no larger than Q/P and the sum of squares of $g_j(\mathbf{s})$, which belongs to the same small group, is less than $4\alpha^2$, we get the inequality (38).

Then we will present some approximations. Utilizing $\Delta_0 \approx 1$, $\Delta_0' \approx 2$, which are derived

from Assumption (viii), and (38) in (11), (12), (13), and (14), we have

$$\begin{aligned}
\beta_0 \approx & 4\alpha^2\Delta_0\Delta_Q^2\theta^2/\pi \approx 4\alpha^2\Delta_Q^2\theta^2/\pi, && (40)\\
\beta_1 \approx & \frac{1}{\pi\mu^2\sigma_x^4\Delta_L^2}8(L-Q)\alpha^2\Delta_0\Delta_Q\theta^2 \\
\approx & \frac{1}{\pi\mu^2\sigma_x^4\Delta_L^2}8(L-Q)\alpha^2\Delta_Q\theta^2, && (41)\\
\sqrt{\beta_1^2-\beta_2^2} \approx & \frac{1}{\mu^2\sigma_x^4\Delta_L^2}\Bigg[\left(8(L-Q)\alpha^2\Delta_Q\theta^2/\pi+2\Delta_L G(\mathbf{s})\right)^2 \\
& -16\alpha^2(L-Q)^2\theta^2\left(4\alpha^2\Delta_Q^2\theta^2/\pi+2\mu\sigma_x^2\Delta_L G(\mathbf{s})\right)/\pi\Bigg]^{\frac{1}{2}} \\
\approx & \frac{4\alpha\theta\sqrt{2(L-Q)G(\mathbf{s})/\pi}}{\mu^2\sigma_x^4\Delta_L}, && (42)\\
\beta_3 \approx & \frac{\pi\mu^3\sigma_x^4\sigma_v^2\Delta_L}{2\alpha^2\Delta_Q^2\theta^2}. && (43)
\end{aligned}$$

In deriving (42), we ignore the term of $4\Delta_L^2G^2(\mathbf{s})$ because $2\Delta_L G(\mathbf{s})$ is much smaller than $8(L-Q)\alpha^2\Delta_Q\theta^2/\pi$. Then by utilizing the equality that $\Delta_L = \Delta_Q - (L-Q)\mu\sigma_x^2$, we get (42).

Utilizing (41), (42), and (43) in (27), one achieves a temporary result that

$$D_\infty^{\min} \approx \frac{\mu\sigma_v^2}{\Delta_L}\left(L-\frac{2(L-Q)}{\Delta_Q}\right)+\frac{\mu\sigma_v^2\sqrt{2\pi(L-Q)G(\mathbf{s})}}{\alpha\theta\Delta_Q^2}. \quad (44)$$

The first item in the RHS of (44) can be further approximated by adopting Assumption (viii) and we finally arrive (29). ∎

9.2 Proof of Property 2

When we choose a group partition size P to divide the unknown system coefficients, L/P groups are obtained. For simplicity, we consider every group independently. Here we denote the number of nonzero coefficients in a group as a random variable M and the probability that M equals $m, 0 \le m \le L$, as $\mathrm{P}\{M=m\}$. Based on the definition of transfer matrix and the fact that the Markov process is in steady-state distribution, we can get the solution of $m=0$ as

$$\mathrm{P}\{M=0\} = \frac{1-p_2}{2-p_1-p_2}p_1^{P-1}.$$

Utilizing Property 1, we then have

$$\overline{Q} = \frac{L}{P}(P\cdot\mathrm{P}\{M>0\}) = L\left(1-(1-\overline{S})p_1^{P-1}\right). \quad (45)$$

9.3 Proof of Theorem 1

Proof For the sake of mathematical tractability, we replace Q and $G(\mathbf{s})$ in (29) with their means, respectively, to yield the final average steady-state MSD of (31). One may accept

that there is no other choice because the formula of (29), which is simplified from (27), is still highly sophisticated. However, simulation result will verify that the approximation produces acceptable errors.

What remains in finishing the proof is to derive the expression of $\overline{Q}$ and $\overline{G(\mathbf{s})}$. The former can be gotten based on Property 2 of M-G model, and we define

$$\Delta_{\overline{Q}} \triangleq 2 - (\overline{Q} + 2)\mu\sigma_x^2. \tag{46}$$

In the following, we will conduct the derivations of $\overline{G(\mathbf{s})}$.

Assuming that there are m nonzero unknown coefficients in a certain small group, we have

$$\sum_{j=1}^{m} g_j^2(\mathbf{s}) = 4\alpha^4 \sum_{j=1}^{m} s_j^2 - 8\alpha^3 \left(\sum_{j=1}^{m} s_j^2 \right)^{\frac{1}{2}} + 4\alpha^2. \tag{47}$$

We denote the mean of the LHS of (47) as $F_\alpha(m)$. According to the M-G model that the nonzero coefficients follow *i.i.d.* Gaussian distribution and the property of χ^2 distribution, $F_\alpha(m)$ is obtained as

$$F_\alpha(m) = \frac{4\alpha^2}{\Gamma(m/2)} \left[2\alpha^2\sigma_s^2 \gamma\left(\frac{m+2}{2}, \frac{1}{2\alpha^2} \right) + \gamma\left(\frac{m}{2}, \frac{1}{2\alpha^2} \right) - 2\sqrt{2}\alpha\sigma_s \gamma\left(\frac{m+1}{2}, \frac{1}{2\alpha^2} \right) \right], \tag{48}$$

where $\Gamma(\cdot)$ and $\gamma(\cdot,\cdot)$ denote the ordinary gamma function and the lower incomplete function, respectively. Then we know that

$$\overline{G(\mathbf{s})} = \frac{L}{P} \sum_{m=1}^{P} \mathrm{P}\{M = m\} F_\alpha(m). \tag{49}$$

where $\mathrm{P}\{M = m\}$ follows the definition in Appendix 9.2 and is further solved as [43]

$$\mathrm{P}\{M{=}m\} = \begin{cases} \dfrac{1-p_2}{2-p_1-p_2} p_1^{P-1}, & m{=}0; \\ \dfrac{1}{(2{-}p_1{-}p_2)(1{-}p_1)} \Big(R_{P-m} {+} 2(1{-}p_1{-}p_2) R_{P-m-1} + (1{-}p_1{-}p_2)^2 R_{P-m-2} \Big), & 0{<}m{<}P; \\ \dfrac{1-p_1}{2-p_1-p_2} p_2^{P-1}, & m{=}P. \end{cases} \tag{50}$$

where

$$R_i = \begin{cases} (1{-}p_1)^{m+1} \sum_{n=0}^{m-1} \binom{m-1}{n} \binom{m+i-n}{m-n} \left(\frac{p_1+p_2-1}{1-p_1} \right)^n (1{-}p_2)^{m-1-n} p_1^{i-m+1}, & i \geq 0; \\ 0, & i < 0. \end{cases} \tag{51}$$

Then the AMS-MSD is finally achieved.

Considering its sophisticated shape of expression (31), we prefer to numerically solve the optimal parameter by (32). Thus the proof of Theorem 1 is completed. ∎

9.4 Proof of Theorem 2

PROOF From Lemma 3, we have already gotten the close forms of instantaneous MSD, c_3, and λ_3. Here we will show how to get the approximate close forms of $\lambda_1, \lambda_2, c_1$, and c_2. Note that all expressions of $\{c_i, \lambda_i\}$ are attached by $(\cdot)'$ here to distinguish them from those in Lemma 3.

We will first present the sketch of our proof. Based on Lemma 3, we know that λ_1 and λ_2 are the eigenvalues of matrix $\mathbf{A}$, which is defined by (16). Utilizing Assumption (viii), which derives $\Delta_0 \approx 1$ and $\Delta_Q \approx 2$, in (16), we could simplify $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$ by

$$(\lambda')^2 - \left(2 - 2\mu\sigma_x^2 - \mu C(\mu,\omega)\right)\lambda' + 1 - 2\mu\sigma_x^2 - \mu C(\mu,\omega) + 2\mu^2\sigma_x^2 C(\mu,\omega) = 0, \tag{52}$$

where

$$C(\mu,\omega) \triangleq \sigma_x^2\Delta_L + \sqrt{\frac{8}{\pi}}\frac{\kappa\alpha}{\mu\omega}\theta. \tag{53}$$

We then solve the quadratic equation of (52) and get the close forms of

$$\lambda_1' \triangleq 1 - 2\mu\sigma_x^2, \qquad \lambda_2' \triangleq 1 - \mu C(\mu,\omega).$$

As a consequence, c_1' and c_2' are obtained by satisfying initial values,

$$c_{1,2}' \triangleq \frac{(1-\mu\sigma_x^2\Delta_L)\,\|\mathbf{s}\|_2^2 + b_{0,0} - D_\infty - c_3'\lambda_3'}{\lambda_{1,2}' - \lambda_{2,1}'} - \frac{\lambda_{2,1}'\left(\|\mathbf{s}\|_2^2 - c_3' - D_\infty\right)}{\lambda_{1,2}' - \lambda_{2,1}'}, \tag{54}$$

where λ_3' and c_3' are defined in (21) and (22), respectively. If we could prove that

$$C(\mu,\omega) \approx 2\sigma_x^2\alpha\theta\sqrt{\frac{2\,(L-Q)}{\pi G(\mathbf{s})}}, \tag{55}$$

Thoerem 2 is ready to be proved.

Next we will prove (55). By taking the approximation of $\Delta_0 \approx 1, \Delta_0' \approx \Delta_Q \approx 2$ and utilizing the optimal κ in (20), we get

$$\mu\sigma_x^2\Delta_L\omega^2 + \frac{8\kappa_{\text{opt}}\alpha\theta\omega}{\sqrt{2\pi}} - \mu^2\sigma_x^2\sigma_v^2 - \kappa_{\text{opt}}^2\left(\frac{4\alpha^2}{P} + G(\mathbf{s})\right) = 0. \tag{56}$$

Solving (56) to get the closed form of ω and inserting it into $\kappa_{\text{opt}}/(\mu\omega)$, we get

$$\frac{\kappa_{\text{opt}}}{\mu\omega} = \frac{\sigma_x^2\Delta_L}{-\sqrt{\frac{8}{\pi}}\alpha\theta + \sqrt{t_1 + t_2 + t_3}}, \tag{57}$$

where

$$t_1 \triangleq 8\alpha^2\theta^2/\pi, \tag{58}$$

$$t_2 \triangleq \mu^3\sigma_x^4\sigma_v^2\Delta_L/\kappa_{\text{opt}}^2 \approx 16\alpha^2\theta^2/(\pi t_4), \tag{59}$$

$$t_3 \triangleq \mu\sigma_x^2\Delta_L\left(4\alpha^2/P + G(s)\right), \tag{60}$$

$$t_4 \triangleq \sqrt{\frac{L-Q}{2\pi G(\mathbf{s})}}\frac{4\alpha\theta}{\Delta_L} - 1. \tag{61}$$

Please note that the approximation in (59) is produced by utilizing (40), (41), (42), (43), and $\Delta_Q \approx 2$ in (26).

If it can be proved that

$$G(s) \ll L\alpha^2/P, \tag{62}$$

we see that t_1 is much larger than t_3. Inserting (58) and (59) in (57) and omitting t_3, we get

$$\frac{\kappa_{\text{opt}}}{\mu\omega} = \frac{\sigma_x^2\Delta_L}{\sqrt{\frac{8}{\pi}}\alpha\theta\left(-1+\sqrt{1+2/t_4}\right)} \approx \sqrt{\frac{\pi}{8}}\frac{\sigma_x^2\Delta_L t_4}{\alpha\theta}, \tag{63}$$

■

where $t_4 \gg 1$, which can be derived based on (61) and (62), is adopted in (63). Utilizing (61) and (63) in (53), we finally prove (55).

Now we will prove (62). According to the empirical hypothesis that $\max(F_\alpha(m)/\alpha^2) \sim O(1)$ and two properties that $\mathrm{P}\{M=m\} \ll 1$ with $1 \le M \le P$ and that $F_\alpha(m)$ decreases dramatically with the increase of m based on the property of $\gamma(\cdot,\cdot)$, we get that $\overline{G(s)} \ll L\alpha^2/P$. Finally we adopt the method used in the proof of Theorem 1 by replacing the random variables in Lemma 3 by their expectations to produce an average result.

In Theorem 2, $\overline{\|\mathbf{s}\|_2^2}$ and $\overline{G'(\mathbf{s})}$ are defined as, respectively,

$$\overline{\|\mathbf{s}\|_2^2} = L\overline{S}\sigma_s^2, \tag{64}$$

$$\overline{G'(\mathbf{s})} = \frac{L}{P}\sum_{m=1}^{P}\mathrm{P}\{M=m\}F'_\alpha(m), \tag{65}$$

where $F'_\alpha(m)$ is defined as

$$\frac{1}{\Gamma(m/2)}\left[4\alpha^2\sigma_s^2\gamma\left(\frac{m+2}{2},\frac{1}{2\alpha^2}\right) - 2\sqrt{2}\alpha\sigma_s\gamma\left(\frac{m+1}{2},\frac{1}{2\alpha^2}\right)\right].$$

9.5 Proof of Corollary 2

Proof Because the step-size in BS-LMS with P_{opt} is larger than that in l_0-LMS, it is obvious that λ'_1 and λ'_3 in BS-LMS with P_{opt} is smaller. In order to compare λ'_2 between BS-LMS and l_0-LMS, we investigate $(L-\overline{Q})\theta^2/\overline{G(\mathbf{s})}$ in (35).

According to the property of $\gamma(a,x)$, $F'_\alpha(n)$ decreases dramatically with the increase of n. Thus $\sum_{m=1}^{P}\mathrm{P}\{M=m\}F'_\alpha(m)$ is mainly determined by its first item. We then have

$$\frac{(L-\overline{Q})\theta^2}{\overline{G(\mathbf{s})}} \approx \frac{(1-p_2)\theta^2 P f(P)}{(1-p_2/p_1)F_\alpha(1)},$$

where $f(P)$ is defined as

$$f(P) \triangleq \begin{cases} \dfrac{1-p_2/p_1}{1-p_1}, & P=1; \\ \dfrac{p_1^{P-1}-p_2^{P-1}}{(1-p_2)(1-p_1^{P-1})+(1-p_1)(1-p_2^{P-1})}, & P \ge 2. \end{cases}$$

First, we show that $\theta^2 P_{P>1}$ is no less than $\theta^2 P_{P=1}$. When P is odd, $\theta^2 P$ can be expressed as

$$\theta^2 P = \begin{cases} \dfrac{(P-1)!!(P-1)!!}{P!!(P-2)!!}, & P \text{ is odd and } P > 1; \\ 1, & P = 1. \end{cases}$$

We see that $\theta^2 P$ increases when P becomes larger. Similarly, the same conclusion is reached when P is even. Moreover, we can see that $\theta^2 P_{P=2} = \pi^2/8 > 1 = \theta^2 P_{P=1}$. Therefore $\theta^2 P_{P>1} \geq \theta^2 P_{P=1}$ is gotten.

Next, we prove $f(P_{\text{opt}}) \geq f(1)$ by utilizing the condition that $\overline{Q(P_{\text{opt}})}/L = 1 - (1 - p_2)p_1^{P_{\text{opt}}-1}/(2 - p_1 - p_2) \ll 1$. $f(P_{\text{opt}}) \geq f(1)$ is equivalent to

$$\begin{aligned} L(P_{\text{opt}}) &= \frac{1-p_1}{1-p_2}\left(1 - (p_2/p_1)^{P_{\text{opt}}}\right) \\ &\geq R(P_{\text{opt}}) = \frac{\overline{Q(P_{\text{opt}})}/L}{1 - \overline{Q(P_{\text{opt}})}/L}(1 - p_2/p_1). \end{aligned} \tag{66}$$

If $2 \leq P_{\text{opt}} \leq p_1/(p_1 - p_2)$, we just need to prove $L(2) \geq R(2)$ and $L(p_1/(p_1 - p_2)) \geq R(p_1/(p_1 - p_2))$ because $L(P)$ and $R(P)$ are concave and convex respectively. When $P = 2$, we have

$$\begin{aligned} L(2) &= \frac{1-p_1}{1-p_2}\left(1 - \frac{p_2}{p_1}\right)\left(1 + \frac{p_2}{p_1}\right) \\ &> R(2) = \frac{1-p_1}{1-p_2}\left(1 - \frac{p_2}{p_1}\right)\frac{2-p_2}{p_1}, \end{aligned}$$

based on that p_1 and p_2 are close to 1. When $P = p_1/(p_1 - p_2)$, we take the first-order Taylor expansion of $L(P)$ and $R(P)$. If the Taylor expansion of $L(P)$ is far larger than that of $R(P)$, (66) is valid. we know that

$$\begin{aligned} L\left(\frac{p_1}{p_1 - p_2}\right) &\approx \frac{1-p_1}{1-p_2}\left(1 - \frac{p_2}{p_1}\right)\left(1 + \frac{p_2}{p_1 - p_2}\right) \\ &\gg R\left(\frac{p_1}{p_1 - p_2}\right) \approx \frac{1-p_1}{1-p_2}\left(1 - \frac{p_2}{p_1}\right)\left(1 + \frac{(2 - p_1 - p_2)\frac{p_2}{p_1 - p_2}}{1 - \frac{1-p_1}{p_1-p_2}p_2}\right), \end{aligned} \tag{67}$$

based on that $p_2/(p_1 - p_2) \gg 1$, $2 - p_1 - p_2 \ll 1$ and $(1 - p_1)/(p_1 - p_2) \ll 1$. Therefore (66) is satisfied when $2 \leq P_{\text{opt}} \leq p_1/(p_1 - p_2)$. If $P_{\text{opt}} > p_1/(p_1 - p_2)$, we have that

$$\begin{aligned} L(P_{\text{opt}}) &> \frac{1-p_1}{1-p_2}\left(1 - (p_2/p_1)^{p_1/(p_1-p_2)}\right) \\ &> \frac{1-p_1}{1-p_2}(1 - 1/\text{e}) \geq \frac{1}{3}(1 - p_2) > \frac{1}{3}(1 - p_2/p_1) \\ &> \frac{\overline{Q(P_{\text{opt}})}/L}{1 - \overline{Q(P_{\text{opt}})}/L}(1 - p_2/p_1) = R(P_{\text{opt}}). \end{aligned}$$

■

Thus the conclusion that $f(P_{\text{opt}}) \geq f(1)$ is reached.

Above all, it can be seen that $(L - \overline{Q})\theta^2/\overline{G(\mathbf{s})}$ is larger when P is P_{opt}. Therefore, λ_2' in BS-LMS with optimal P is smaller than that in l_0-LMS. Thus the proof of Corollary 2 is arrived.

References

[1] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

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

[3] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancellers," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 508-518, Sep. 2000.

[4] "Transmission systems and media, digital systems and networks", recommendation ITU-T G.168 (2002).

[5] A. Steingass, A. Lehner, F. Perez-Fontan, E. Kubista, and B. Arbesser-Rastburg, "Characterization of the aeronautical satellite navigation channel through high-resolution measurement and physical optics simulation", *Int. J. Satell. Commun. Network.* 26:1-30, 2008.

[6] S. Wyne, N. Czink, J. Karedal, P. Almers, F. Tufvesson, and A. Molisch, "A Cluster-Based Analysis of Outdoor-to-Indoor Office MIMO Measurements at 5.2 GHz", IEEE 64th Vehicular Technology Conference (VTC-Fall), 1-5, Montreal, Canada, 2006.

[7] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[8] J. H. Gross, D. M. Etter, V. A. Margo, and N. C. Carlson, "A block selection adaptive delay filter algorithm for echo cancellation," in *Midwest Conf. Circuits Syst.*, Aug. 1992, pp. 895-898.

[9] V. A. Margo, D. M. Etter, N. C. Carlson, and J. H. Gross, "Multiple short-length adaptive filters for time-varying echo cancellation," *IEEE ICASSP*, 1993, pp. I161-I164.

[10] M. Berggren, M. Borgh, C. Schuldt, F. Lindstrom and I. Claesson, "Low-Complexity Network echo cancellation approach for systems equipped with external memory," *IEEE Trans. Audio, Speech and Language Process.*, vol. 19, no. 8, pp. 2506-2515, Nov. 2011.

[11] Y. Gu, Y. Chen, and K. Tang, "Network echo canceller with active taps stochastic localization," *IEEE ISCIT*, pp. 556-559, 2005.

[12] Y. Li, Y. Gu, and K. Tang, "Parallel NLMS filters with stochastic active taps and step-sizes for sparse system identification," *IEEE ICASSP*, vol. 3, pp. 109-112, Toulouse, France, 2006.

[13] X. Liu, Y. Li, Y. Gu, and K. Tang, "Enhanced stochastic taps NLMS filter with efficient sparse taps localization," *IEEE ICSP*, vol. 4, pp. 16-20, 2006.

[14] A. Sugiyama, H. Sato, A. Hirano, and S. Ikeda, "A fast convergence algorithm for adaptive FIR filters under computational constraint for adaptive tap-position control," *IEEE Trans. Circuits Syst. II*, vol. 43, pp. 629-636, Sept. 1996.

[15] A. Sugiyama, S. Ikeda, and A. Hirano, "A fast convergence algorithm for sparse-tap adaptive FIR filters identifying an unknown number of dispersive regions," *IEEE Trans. Signal Process.*, vol. 50, no. 12, pp. 3008-3017, December 2002.

[16] O. A. Noskoski, J. C. M. Bermudez, S. J. M. Almeida, "Region-based wavelet-packet adaptive algorithm for identification of sparse impulse responses," *IEEE Trans. Signal Process*, vol. 61, no. 13, pp. 3321-3333, July, 2013.

[17] Y. Gu, J. Jin, and S. Mei, "l_0 norm constraint LMS algorithm for sparse system identification," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 774-777, Sep. 2009.

[18] Y. Chen, Y. Gu, and A. O. Hero, "Sparse LMS for system identification," *IEEE ICASSP*, pp. 3125-3128, Taiwan, Apr. 2009.

[19] O. Taheri and S. A. Vorobyov, "Reweighted l_1-norm penalized LMS for sparse channel estimation and its analysis," Submitted to *IEEE Trans. Signal Process.*.

[20] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed L0 norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289-301, Jan. 2009.

[21] H. Mohimani, M. Babaie-Zadeh, I. Gorodnitsky, and C. Jutten, "Sparse recovery using smoothed L0 (SL0): convergence analysis," *ArXiv preprint arXiv:1001.5073*, 2010.

[22] F. Wu and F. Tong, "Gradient optimization p-norm-like constraint LMS algorithm for sparse system estimation," *Signal Process.*, 93, 967-971, 2013.

[23] F. Wu, Y. Zhou, F. Tong and R. Kastner, "Simplified p-norm-like constraint LMS algorithm for efficient estimation of underwater acoustic channels," *Journal of Marine Science and Application*, Volume 12, Issue 2, pp. 228-234, June 2013.

[24] Y. Chen, Y. Gu, and A. O. Hero, "Regularized least-mean-square algorithms," *ArXiv e-prints Dec. 2010* [Online]. Available: http://arxiv.org/abs/1012.5066v2.

[25] G. Su, J. Jin, Y. Gu, and J. Wang, "Performance analysis of l_0 norm constraint least mean square algorithm," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2223-2235, May 2012.

[26] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302-5316, Nov. 2009.

[27] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the Reconstruction of Block-Sparse Signals With an Optimal Number of Measurements," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3075-3085, 2009.

[28] M. Stojnic, "l_2/l_1-Optimization in Block-Sparse Compressed Sensing and Its Strong Thresholds," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 350-357, Apr. 2010.

[29] J. Liu, J. Jin, and Y. Gu, "Efficient Recovery of Block Sparse Signals via Zero-point Attracting Projection," *IEEE ICASSP*, pp. 3333-3336, Mar. 25-30, 2012, Kyoto, Japan.

[30] E. Elhamifar and R. Vidal, "Block-Sparse Recovery via Convex Optimization," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4094-4107, Aug. 2012.

[31] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982-2001, 2010.

[32] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042-3054, 2010.

[33] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using Markov random fields," *NIPS*, Vancouver, BC, Canada, Dec. 2008.

[34] Z. Ben-Haim and Y. C. Eldar, "Near-Oracle Performance of Greedy Block-Sparse Estimation Techniques From Noisy Measurements," *IEEE Trans. Signal Process.*, vol. 5, no. 5, pp. 1032-1047, 2011.

[35] L. Yu, H. Sun, J. P. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Process.*, vol. 92, no. 1, pp. 259-269, 2012.

[36] Z. Zhang and B. D. Rao, "Extension of SBL Algorithms for the Recovery of Block Sparse Signals With Intra-Block Correlation," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2009-2015, 2013.

[37] V. Cevher, P. Indyk, C. Hegde, and R. G. Baraniuk, "Recovery of clustered sparse signals from compressive measurements," *SAMPTA*, Marseille, France, May. 2009.

[38] F. Parvaresh and B. Hassibi, "Explicit measurements with almost optimal thresholds for compressed sensing," *IEEE ICASSP*, Mar-Apr 2008.

[39] E. M. Eksioglu, "Group sparse RLS algorithms", *International Journal of Adaptive Control and Signal Processing*, Dec. 11, 2013.

[40] C. D. Scott, "Adaptive Filters Employing Partial Updates", *IEEE Trans. Circuit and Systems*, 44(3):209-216, 1997.

[41] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," *ICML*, 1998, pp. 82-90.

[42] B. M. McCoy and T. T. Wu. *The two-dimensional Ising model.* Harvard Univ. Press, 1973.

[43] B. Brainerd and S. M. Chang, "Number of Occurrences in Two-State Markov Chains, with an Application in Linguistics," The Canadian Journal of Statistics, 10(3):225-231, 1982.