Title
Algorithm development for sparse signal recovery and performance limits using multiple-user information theory

Permalink

Author
Jin, Yuzhe

Publication Date
2011

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Algorithm Development for Sparse Signal Recovery and Performance Limits Using Multiple-User Information Theory

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Yuzhe Jin

Committee in charge:

Bhaskar D. Rao, Chair
Sanjoy Dasgupta
William Hodgkiss
Young-Han Kim
Kenneth Kreutz-Delgado
Scott Makeig

2011

The dissertation of Yuzhe Jin is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2011

DEDICATION

To my parents and Yuan.

EPIGRAPH

Man’s mind is not a container to be filled

but rather a fire to be kindled.

—Dorothea Brande

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Working toward a Ph.D. is a journey.

First and foremost, I am deeply indebted to my advisor, Professor Bhaskar Rao, who has been extremely patient and helpful during my study. Professor Rao has taught me everything about doing research and, more importantly, being a professional. He also gave me lots of freedom in choosing the research area of my interest, allowing me to grow to the best of my capabilities. From his guidance, I learned that the value of research is the intellectual progression and expansion for us to push the boundaries of our knowledge. The expertise he generously shared with me, and industry experiences he helped me to obtain, have become my greatest assets.

Another great pleasure during my Ph.D. study is information theory. This is an area which I found difficult to grasp in the beginning but extremely powerful as my understanding developed. Professor Young-Han Kim demonstrated the power of network information theory in his lectures, where the techniques I learned became the foundation of my research. Professor Kim walked me into the domain of information theory and inspired me to appreciate its beauty. He taught me many useful mathematical tools, especially the large deviation theory which made possible my theoretical work. He also patiently helped me to improve my presentation skills, particularly on effectively communicating complex ideas.

I would like to thank other members on my Ph.D. committee, Professors Ken Kreutz-Delgado, Sanjoy Dasgupta, William Hodgkiss, and Dr. Scott Makeig for their insightful discussions during my thesis proposal and many research meetings. Meanwhile, I thank Professor Nuno Vasconcelos for generously permitting me access to the resources at Statistical Visual Computing Lab during my first months in UCSD. The servers named after mathematicians in his lab enabled my timely completion of the costly EM iterations for the Cheetah project. Professor Truong Nguyen and his group granted me the access to the printer in the Video Processing Lab, which has printed me several (indeed, many) big piles of papers over these years. Professor Rajesh Hegde helped me a great deal when I started my study as a Ph.D. student. Under his guidance, I coauthored my first research paper ever. Ananthapadmanabhan Kandhadai and Jeff (Pengjun) Huang at Qualcomm kindly offered me several summer internships to practice my expertise in signal processing on commercialized speech coders. Dr. Kuansan Wang at Microsoft Research guided me to solve web service problems using signal processing techniques and led me to a promising industry. I am very grateful for all their help, supports, and career advices.

I am very lucky to meet my girlfriend, Yuan Zhi, during my Ph.D. study. To all, she exudes kindness, laughter, and strength. To me, also the light in my difficult times and the determination in standing by all my decisions.

Lots of friends have helped me and entertained my life here. I gained lots of analytical perspectives from many senior students including Dashan Gao, Jun Zheng, Junwen Wu, Min Li, Yanhua Mao, Jing Zhu, Dayou Zhou, Wenyi Zhang, Chengmo Yang, Shankar Shivappa, Yogananda Isukapalli, Ling Zhang, Zhou Lu, Lingyun Zhang, Honghao Shan, Yushi Shen, Xinran Wu, Hairuo Zhuang, and Meng-Ping (Ben) Kao. I want to thank my lab members, Zhilin Zhang, Liwen Yu, Ali Masnadi-Shirazi, Ethan Duni, Bongyong Song, Anh Nguyen, Eddy (Hwan Joon) Kwon, for stimulating discussions. Meanwhile, I enjoyed lots of chats with Renshen Wang, Wanping Zhang, Bo Hu, Luo Gu, Yi Jing, Yan Gao, Mingjing Chen, Liang Feng, David Wipf, Ali Karimian, Matt Pugh, Natan Jacobson, Emanuele Coviello, Paul Cuff, Hamed Masnadi-Shirazi, Alex Rasmussen, and Ozan Koyluoglu. A reward for staying long enough here is that I got to know many friends who came after my joining UCSD. It is my pleasure to meet Fei Liu, Wensong Xu, Yu Xiang, Lele Wang, Minghai Qin, Chiao-Yi Chen, Shengjun Pan, Hao Zheng, Jianjian Gao, Sheusheu Tan, Peng Du, Daqian Jin, Peng Wang, Enming Luo, Yanqin Jin, Yupeng Fu, Jing Zheng, and Yujia Wang. Discussions with them are always entertaining and inspiring.

None of my career advances can be achieved without the selfless love and the unconditional support from my parents. This work is the best present that embodies my love and appreciation for them.

The material of this thesis has been published, is submitted for possible publication, or is in preparation for submission.

Chapter 2 is, in part, a reprint of material in the paper "Limits on Support Recovery of Sparse Signals via Multiple Access Communication Techniques," Y. Jin, Y.-H. Kim, B. D. Rao, *submitted to IEEE Transactions on Information Theory*, 2010, the paper "Performance Tradeoffs for Exact Support Recovery of Sparse Signals," Y. Jin, Y.-H. Kim, and B. D. Rao, *published in the proceedings of the International Symposium on Information Theory*, 2010, the paper "Sparse Signal Recovery in Presence of Multiple

Measurement Vectors," Y. Jin and B. D. Rao, *in preparation for IEEE Transactions on Information Theory*, and the paper "On the Role of the Properties of the Non-zero Entries on Sparse Signal Recovery," Y. Jin and B. D. Rao, *published in proceedings of the Asilomar Conference on Signals, Systems and Computers*, Nov. 2010. In all cases, I was the primary author, and B. D. Rao supervised the research. Y.-H. Kim contributed to the paper "Limits on Support Recovery of Sparse Signals via Multiple Access Communication Techniques" and the paper "Performance Tradeoffs for Exact Support Recovery of Sparse Signals."

Chapter 3 is, in part, a reprint of the paper "MultiPass Lasso Algorithms for Sparse Signal Recovery," Y. Jin and B. D. Rao, *published in the proceedings of the International Symposium on Information Theory*, 2011, and the paper "Performance Limit of Matching Pursuit Algorithms," Y. Jin and B. D. Rao, *published in the proceedings of the International Symposium on Information Theory*, July 2008. In both cases, I was the primary author, and B. D. Rao supervised the research.

Chapter 4 is, in part, a reprint of the material published as "Algorithms for Robust Linear Regression by Exploiting the Connection to Sparse Signal Recovery," Y. Jin, B. D. Rao, *International Conference on Acoustics, Speech, and Signal Processing*, March 2010. I was the primary author, and B. D. Rao supervised the research.

Chapter 5 is, in part, a reprint of the paper "LMS Type Adaptive Filtering Algorithms that Incorporate Sparsity," Y. Jin, B. Song, and B. D. Rao, *submitted to IEEE Transactions on Signal Processing*. I was the primary author, B. Song contributed to the research, and B. D. Rao supervised the research.

VITA AND PUBLICATIONS

2005	B. S. in Computer Science and Technology, Tsinghua University, Beijing, China.
2008	Research Intern, Internet Services Research Center, Microsoft Research, Redmond, WA.
2009	Intern, Multimedia R&D and Standard Group, Qualcomm Inc. San Diego, CA.
2005-2011	Research Assistant, University of California, San Diego, La Jolla, CA.
2011	Ph. D. in Electrical Engineering (Signal and Image Processing), University of California, San Diego, La Jolla, CA.

Y. Jin, Y.-H. Kim, and B. D. Rao, "Limits on Support Recovery of Sparse Signals via Multiple Access Communication Techniques," *submitted to IEEE Transactions on Information Theory*, 2010.

Y. Jin, B. Song, and B. D. Rao, "LMS Type Adaptive Filtering Algorithms that Incorporate Sparsity," *submitted to IEEE Transactions on Signal Processing*, 2011.

Y. Jin and B. D. Rao, "Sparse Signal Recovery in Presence of Multiple Measurement Vectors," *in preparation for IEEE Transactions on Information Theory*.

Y. Jin and B. D. Rao, "MultiPass Lasso Algorithms for Sparse Signal Recovery," *International Symposium on Information Theory*, July 2011.

Y. Jin and B. D. Rao,"On the Role of the Properties of the Non-zero Entries on Sparse Signal Recovery," *Asilomar Conference on Signals, Systems and Computers*, November 2010.

Y. Jin, Y.-H. Kim, and B. D. Rao, "Performance Tradeoffs for Exact Support Recovery of Sparse Signals," *International Symposium on Information Theory*, June 2010.

Y. Jin and B. D. Rao, "Algorithms for Robust Linear Regression by Exploiting the Connection to Sparse Signal Recovery," *International Conference on Acoustics, Speech, and Signal Processing*, March 2010.

Y. Jin and B. D. Rao, "Performance Limit of Matching Pursuit Algorithms," *International Symposium on Information Theory*, July 2008.

Y. Jin and B. D. Rao, “Insights into the Stable Recovery of Sparse Solutions in Overcomplete Representations Using Network Information Theory,” *International Conference on Acoustics, Speech, and Signal Processing*, April 2008.

R. M. Hegde, Y. Jin, and B. D. Rao, “Spectral Estimation of Voiced Speech Using a Family of MVDR Estimates,” *International Conference on Acoustics, Speech, and Signal Processing*, April 2007.

ABSTRACT OF THE DISSERTATION

Algorithm Development for Sparse Signal Recovery and Performance Limits Using Multiple-User Information Theory

by

Yuzhe Jin

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2011

Bhaskar D. Rao, Chair

The problem of sparse signal recovery can be formulated as a problem of finding a sparse solution $\mathbf{X} \in \mathbb{R}^m$ to an underdetermined system of equations

$$\mathbf{Y} = A\mathbf{X} + \mathbf{Z}$$

where $A \in \mathbb{R}^{n \times m}, n < m$, is the measurement matrix, $\mathbf{Z} \in \mathbb{R}^n$ is the measurement noise, and $\mathbf{Y} \in \mathbb{R}^n$ is the noisy measurement. Let k be the number of nonzero entries in

$\mathbf{X}$. Then, $\mathbf{X}$ is sparse when $k \ll m$.

First, we study the performance limits of support recovery, i.e., the recovery of the locations of the nonzero entries of $\mathbf{X}$. By connecting sparse signal recovery to communication over a multiple access channel (MAC), we show that MAC capacity region sheds light on the performance limits of support recovery. Sufficient and necessary conditions are derived to reveal the role of the model parameters, such as the dimension of the signal m, the number of measurements n, and especially the magnitudes of nonzero entries of $\mathbf{X}$, in ensuring asymptotically successful support recovery. By drawing an analogy with the single-input multiple-output MAC, the information theoretic results are extended to the multiple measurement vectors (MMV) problem to shed light on the role of multiple measurements.

Next, we propose a multiple-pass algorithmic framework which exploits the variety in the dynamic range of nonzero entries. Specifically, nonzero entries with similar magnitudes are jointly detected as a group and different groups are detected sequentially. The MultiPass Lasso algorithm and its variant are studied in detail, with experimental results demonstrating the performance improvement over their one-pass counterparts, respectively, in reconstruction accuracy and computational complexity.

We also examine two novel applications of sparse signal recovery. For robust linear regression, we model the outliers in the observations, which occur infrequently, as the nonzero entries of a sparse vector, and derive methods rooted in sparse signal recovery to tackle this problem. For adaptive filtering with sparse predictors, we employ a general diversity measure for sparse signal recovery to develop a prototype least-mean-squares (LMS) type algorithm, in which the optimization is performed in the affine scaling domain to expedite the convergence. Steady-state analysis is conducted to provide insight into the performance. As two instantiations, the pALMS and pANLMS algorithms are studied in detail. Experimental results support the potentials of sparse signal recovery techniques in both applications.

Chapter 1

Introduction

The problem of sparse signal recovery can be formulated as a problem of finding a sparse solution to an underdetermined systems of equations

$$\mathbf{Y} = A\mathbf{X} + \mathbf{Z} \tag{1.1}$$

where $\mathbf{X} \in \mathbb{R}^m$ is the signal of interest, $A \in \mathbb{R}^{n \times m}$ is the measurement matrix with $n < m$, and $\mathbf{Z} \in \mathbb{R}^n$ is the measurement noise, and $\mathbf{Y} \in \mathbb{R}^n$ is the noisy measurement. Let k denote the number of nonzero entries of $\mathbf{X}$, and $\mathbf{X}$ is said to be sparse when $k \ll m$. This problem with model (1.1) is usually referred to as sparse signal recovery with single measurement vector (SMV).

When being considered in an application context, the rows or columns of A usually form a physically meaningful model, and the nonzero elements of $\mathbf{X}$ are usually of interest and need to be identified. As one example, in the problem of the estimation of a wireless multipath channel [5, 29], each row of A represents a segment of the training sequence, and $\mathbf{X}$ represents the coefficients of the sparse multipath channel to be estimated. As another example, in the applications of electroencephalography (EEG) and magnetoencephalography (MEG) [59, 60], the columns of A can represent the instantaneous impulse responses of the brain sources on the EEG or MEG sensors, and identifying the nonzero entries in the vector $\mathbf{X}$ can be viewed as localizing the neuronal activities for a given observation $\mathbf{Y}$. Recently, the development of compressed sensing [16, 39] provides the flexibility in designing the matrix A. In this case, each row of A can be randomly generated according to certain probability distribution, and hence $\mathbf{Y}$ contains n random projections of $\mathbf{X}$ onto the rows of A. The ability to recover a sparse vector $\mathbf{X}$ from as few random projections as possible offers a universal approach for signal acquisition and compression, which are independent of the domain specific structures of the signals. Further, (1.1) also serves as the underlying mathematical model for many other applications, such as image processing [71], [40], robust face recognition [135], bandlimited extrapolation and spectral estimation [13], speech processing [27], echo cancellation [41], [101], body area networks [53], and wireless communication [62].

It is worthwhile to note the growing recognition of the importance of the area of sparse signal recovery as evidenced by numerous special research journal issues, conference workshops and sessions. One of the earliest full sessions, if not the first session, on sparse signal recovery was organized by Yoram Bresler and Bhaskar Rao at ICASSP, 1998. Since then, many conference sessions and journal issues have been dedicated to the issue of sparsity. Some recent notable examples of such include several tutorial sessions at ITA Workshop, 2008; a panel session at ICASSP, 2008; a special issue on IEEE Journal of Selected Topics in Signal Processing, 2010; and a special issue IEEE Journal of Selected Topics in Signal Processing, 2011. Overall, sparse signal recovery has become an important and promising area with both theoretic and practical potentials.

1.1 Background

1.1.1 Sparse Signal Recovery with SMV

For the problem of sparse signal recovery with SMV, computationally efficient algorithms have been proposed to find or approximate the sparse solution $\mathbf{X} \in \mathbb{R}^m$ in various settings. A partial list includes matching pursuit [82], orthogonal matching pursuit (OMP) [94], Lasso [113], basis pursuit [24], FOCUSS [60], iteratively reweighted ℓ_1 minimization [20], iteratively reweighted ℓ_2 minimization [21], sparse Bayesian learning (SBL) [114, 130], finite rate of innovation [120], CoSaMP [88], and subspace pursuit [34]. Analysis has been developed to shed light on the performances of these practical algorithms. For example, Donoho [39], Donoho, Elad, and Temlyakov [36], Candès and Tao [19], and Candès, Romberg, and Tao [18] presented sufficient conditions for ℓ_1-norm minimization algorithms, including basis pursuit and its variant in the noisy setting, to successfully recover the sparse signals with respect to different performance metrics. Wainwright [123] and Zhao and Yu [140] provided sufficient and necessary conditions for Lasso to recover the support of the sparse signal, i.e., the set of indices of the nonzero entries. Tropp [115], Tropp and Gilbert [116], and Donoho, Tsaig, Drori, and Starck [37] studied the performances of greedy sequential selection methods such as matching pursuit and its variants. On the other hand, from an information theoretic perspective, a series of papers, for instance, Wainwright [122], Fletcher, Rangan, and Goyal [51], Wang, Wainwright, and Ramchandran [124], and Akçakaya and Tarokh [2], provided sufficient and necessary conditions to indicate the performance limits of optimal algorithms for support recovery, regardless of computational complexity.

In addition to the batch estimation techniques, which are based on blocks of measurements, the development of adaptive algorithms with sparsity concerns has recently become an area of growing interest. The advantages of adaptive algorithms include their relatively lower computational complexity and their ability to track the system dynamics. As a notable example, the proportionate NLMS (PNLMS) algorithm was developed by Duttweiler [41] to deal with the sparse channel estimation that arises in echo cancellers. Although this algorithm was not formally derived by minimizing an underlying objective function, it was well motivated with theoretical analysis and simulations to support its effectiveness in adaptively estimating the sparse predictor. Other recent adaptive filtering algorithms with sparsity concerns include PNLMS++ [54], improved PNLMS (IPNLMS) [7], improved IPNLMS (IIPNLMS) [33], zero attracting LMS (ZA-LMS) and the reweighted zero attracting LMS (RZA-LMS) [25, 26], and ℓ_0-LMS and ℓ_0-NLMS [61]. Performance analysis are derived for the algorithms above to support their usage in practical applications.

1.1.2 Sparse Signal Recovery with MMV

An emerging trend in the area of sparse signal recovery is the capability of collecting multiple measurement vectors in an increasing number of applications, such as EEG and MEG [128, 136], blind source separation [28], source localization [80], multivariate regression [89], and direction of arrival estimation [111]. This gives rise to the problem of sparse signal recovery with multiple measurement vectors (MMV) [30], for which the model is given by

$$Y = AX + Z \tag{1.2}$$

where $X \in \mathbb{R}^{m\times l}$ is the (matrix) signal of interest, $A \in \mathbb{R}^{n\times m}$ is again the measurement matrix, $Z \in \mathbb{R}^{n\times l}$ is the measurement noise, and $Y \in \mathbb{R}^{n\times l}$ is the noisy measurement, for $l > 1$. Let k be the number of nonzero rows in X, and X is sparse when $k \ll m$.

Practical algorithms have been developed to address the new challenges in this scenario. One class of algorithms for solving the MMV problem can be viewed as straightforward extensions based on their counterparts for the SMV problem. To sample a few, M-OMP [29, 117], M-FOCUSS [29], ℓ_1/ℓ_2 minimization method[1] [46], multivariate group Lasso [89], and M-SBL [134] can be all viewed as examples of this kind. Another class of algorithms additionally make explicit effort to exploit the structure underlying the sparse signal X, such as the temporal correlation or the autoregressive

[1]This method is sometimes referred to as ℓ_2/ℓ_1 minimization, due to the naming convention in a specific paper. In this thesis, we use ℓ_1/ℓ_p to indicate a cost of a matrix B which is define as $\sum_i |(\sum_j |b_{i,j}|^p)^{1/p}|$.

nature across the columns of X which would be otherwise unavailable when $l = 1$, to aim for better performance of sparse signal recovery. For instance, the improved M-FOCUSS algorithms [136], auto-regressive sparse Bayesian learning (AR-SBL) [138], and T-SBL [139] all have the capability of explicitly taking advantage of the structural properties of X to improve the recovery performance. Alone side the algorithmic advancement, a series of work have been focusing on the theoretical analysis to support the effectiveness of existing algorithms for the MMV problem. We briefly divide these results into two categories. The first category of theoretic analysis aims at practical algorithms for sparse signal recovery with MMV. For example, Chen and Huo [23] discovered the sufficient conditions for ℓ_1/ℓ_p norm minimization method and orthogonal matching pursuit to exactly recover every sparse signal within certain sparsity level in the noiseless setting. Eldar and Rauhut [47] also analyzed the performance of sparse signal recovery using the ℓ_1/ℓ_2 norm minimization method in the noiseless setting, but the sparse signal was assumed to be randomly distributed according to certain probability distribution and the performance was averaged over all possible realizations of the sparse signal. Obozinski, Wainwright, and Jordan [89] provided sufficient and necessary conditions for multivariate group Lasso to successfully recover the support of the sparse signal[2] in the presence of measurement noise. The second category of theoretic analysis are of an information theoretic nature, and explore the performance limits that any algorithm, regardless of computational complexity, could possibly achieve. In this regard, Tang and Nehorai [111] employed a hypothesis testing framework with the likelihood

[2]We refer to the support of a matrix X as the set of indices corresponding to the nonzero rows of X.

ratio test as the optimal decision rule to study how fast the error probability decays. Sufficient and necessary conditions are further identified in order to guarantee successful support recovery in the asymptotic sense.

1.2 Main Contributions of the Thesis

The main contributions of the thesis focus on the theory, algorithms, and applications of sparse signal recovery. We briefly preview these aspects in this section.

1.2.1 Performance Limits of Support Recovery

The problem of support recovery of sparse signals concerns the identification of the locations of the nonzero entries of $\mathbf{X}$ in the SMV problem (or the nonzero rows of X in the MMV problem). This problem arises in applications where the locations of the nonzero entries capture important information about the underlying signal activity. As an example, in medical imaging applications such as MEG and EEG, it is desirable to localize active neuronal activities which corresponds to the detection of the locations of the nonzero entries of $\mathbf{X}$.

To address this problem, we proposed a connection between the problems of sparse signal recovery and multiuser communication. Let us consider the SMV problem for the discussion to follow. Based on this connection, the columns of the measurement matrix A form a common codebook for all senders. Codewords from the senders are individually multiplied by unknown channel gains, which correspond to nonzero entries of $\mathbf{X}$. Then, the noise-corrupted linear combination of these codewords is observed. Thus, support recovery can be interpreted as decoding messages from multiple senders. Using the techniques rooted in multiuser communication but suitably modified for the support recovery problem, we develop sharp necessary and sufficient conditions for support recovery of sparse signals to be asymptotically successful. For example, when k is fixed, we show that $n = (\log m)/c(\mathbf{X})$ is sufficient and necessary. We give a complete characterization of $c(\mathbf{X})$ that depends on the values of all nonzero entries of $\mathbf{X}$. This result provides a clear insight into the role of nonzero entries in support recovery, which improves upon many existing results where only the minimum nonzero magnitude entered the performance tradeoffs. We also provide performance limits for the scenarios with increasing k, random nonzero activities, and MMV, respectively. Especially, the performance limit regarding the MMV model, which is enabled by a connection to the single-input multiple-output (SIMO) MAC communication problem, suggests the potential performance improvement enabled by having multiple measurement vectors.

1.2.2 Algorithm Design and Analysis Using Multiuser Information Theoretic Techniques

One can broadly classify the existing algorithms for sparse signal recovery into two categories: sequential selection methods, such as matching pursuit and orthogonal matching pursuit, and joint recovery methods, such as basis pursuit and Lasso. It has been observed that sequential selection methods can deal well with sparse signals whose nonzero entries have very different magnitudes, but perform poorly for sparse signal with similar magnitudes. In contrast, joint recovery methods can handle well sparse signals with similar magnitudes, but they may not exploit the disparity in nonzero magnitudes. In practice, the nonzero entries of a sparse signal can be modelled as clusters with each cluster comprising of a group of nonzero entries with comparable magnitudes. The variation in the dynamic range of the signals poses a new challenge for algorithmic development.

Motivated by the group detector for multiuser detection, we explore the opportunity of merging the different algorithmic design principles together to develop novel algorithms suitable for practical signals. We propose the MultiPass algorithmic framework, which sequentially detects different groups of nonzero entries. This is achieved by applying a joint recovery method with the goal of identifying a subset of the support, on which the nonzero entries have comparable magnitudes, with high probability at each iteration. The MultiPass Lasso (MPL) algorithm is developed as an instantiation under this algorithmic framework. Meanwhile, we propose the Reweighted MultiPass Lasso algorithm which utilizes MPL as an component. Experiment study shows that the proposed algorithms yield improved estimation accuracy and computational efficiency. Theoretical analysis is also provided to support the performance improvement of the MultiPass Lasso algorithm.

Then, we focus on the performance limit of orthogonal matching pursuit (OMP) using the connection between OMP and the successive interference cancellation (SIC) scheme for multiuser detection. Inspired by the decoding criterion for SIC, we propose an intuitive necessary condition for OMP to successfully recover the support of the sparse signal. Experiments and discussions are provided to shed light on the performance limit of OMP.

1.2.3 Robust Linear Regression

The problem of robust linear regression focuses on learning a linear model based on the data in presence of outliers, which are observations that "deviate so much from other observations as to arouse suspicion that they were generated by a different mechanism" [67]. Note that it is important to recognize the impact of outliers for modeling and analysis. Since some methods such as the least-squares (LS) estimation are very sensitive to outliers, the need of robust methods for regression analysis becomes evident.

We leverage the techniques for sparse signal recovery to design novel algorithms for robust regression. The key aspect in our approach is the employment of a two-component model for measurement noise, where one component accounts for the regular Gaussian noise and the other component captures the outliers in the data. Since the outliers occurs infrequently, the component vector representing outliers can be therefore viewed as a sparse vector. We propose a maximum-a-posteriori (MAP) based method and a sparse Bayesian learning (SBL) based method to jointly estimate the regression coefficients as well as the outlier component. Experimental results on simulated and real data sets support the effectiveness of the proposed algorithms.

1.2.4 Adaptive Filtering Algorithms with Sparsity Concerns

In the problem of adaptive filtering, the goal is to develop online learning algorithms that can adaptively estimate the underlying system coefficients based on the previous state and the current input of the system. We focus on the design of adaptive algorithms for the scenario where the system coefficients are sparse.

To tackle this problem, we utilize a general diversity measures, whose minimization leads to sparse vectors, in the cost function for deriving a prototype least-mean-squares (LMS) type algorithm and a normalized LMS (NLMS) type algorithm. Specifically, motivated by the ideas behind the interior point method in convex and non-linear optimization, we perform the steepest descent method in the affine scaling domain with the goal of expediting the speed of convergence of the algorithm. Then, we study in detail the pALMS and pANLMS algorithms, which instantiate the prototype algorithms with the ℓ_p norm diversity measure. Steady-state analysis are provided to shed light on their performances. Experiments are performed to support the effectiveness of the proposed algorithms.

1.3 Thesis Outline

The remainder of the thesis is organized as follows.

In Chapter 2, we study the performance limits of support recovery of sparse signals. Utilizing the connection between the sparse signal recovery problem and multiple-user communication, we explore the sufficient and necessary conditions for support recovery to be asymptotically successful in SMV and MMV cases, respectively. We further provide an intuitive analysis for the performance limits of matching pursuit algorithms.

Chapter 3 focuses on how to apply multiuser information theoretic techniques to facilitate the design and the analysis of practical algorithms for sparse signal recovery. First, we propose the MultiPass algorithmic framework that takes advantage of both sequential recovery methods and joint recovery methods for performance improvement. The MultiPass Lasso algorithm and the Reweighted MultiPass Lasso algorithm are proposed, and their performances are studied in detail. Then, we discuss the performance limit of OMP using its connection to the successive interference cancellation scheme for multiuser detection.

Starting from Chapter 4, we turn our attention to applications. In this chapter, we employ a two-component model for measurement noise in the problem of robust regression. Consequently, a sparse component corresponding to the outliers is naturally isolated. We show how to transform the techniques of sparse signal recovery to achieve robust regression.

In Chapter 5, we focus on the problem of adaptive filtering. Employing the diversity measures from sparse signal recovery, we derive a prototype adaptive filters using a steepest descent approach in the affine scaling domain, with the goal of expedite the convergence of the algorithm. The choice of using the ℓ_p norm as the diversity measure is studied in detail.

Chapter 6 concludes the thesis with discussion on future research directions.

Chapter 2

Performance Limits of Support Recovery Using Multiuser Information Theory

2.1 Introduction

The problem of support recovery of sparse signals focuses on the recovery of the set of (row) indices corresponding to the nonzero entries. In many applications, finding the exact support of the signal is important even in the noisy setting. For example, in applications of medical imaging, magnetoencephalography (MEG) and electroencephalography (EEG) are common approaches for collecting noninvasive measurements of external electromagnetic signals [4]. A relatively fine spatial resolution is required to localize the neural electrical activities from a huge number of potential locations [128].

In the domain of cognitive radio, spectrum sensing plays an important role in identifying available spectrum for communication, where estimating the number of active subbands and their locations becomes a nontrivial task [112]. In multiple-user communication systems such as a code-division multiple access (CDMA) system, the problem of neighbor discovery requires identification of active nodes from all potential nodes in a network based on a linear superposition of the signature waveforms of the active nodes [62]. In all these problems, finding the support of the sparse signal is more important than approximating the signal vector in the Euclidean distance. Hence, it is important to understand performance issues in the exact support recovery of sparse signals with noisy measurements.

Information theoretic tools have proven successful in this direction. In the single measurement vector case, Wainwright [121], [122] considered the problem of exact support recovery using the optimal maximum likelihood decoder. Necessary and sufficient conditions are established for different scalings between the sparsity level and signal dimension. Using the same decoder, Rad [98] derived sharp upper bounds on the error probability of exact support recovery. Meanwhile, Fletcher, Rangan, and Goyal [50], [51] improved the necessary condition with the same decoder. Wang, Wainwright, and Ramchandran [124], [125] also presented a set of necessary conditions for exact support recovery. Akçakaya and Tarokh [2] analyzed the performance of a joint typicality decoder and applied it to find a set of necessary and sufficient conditions under different performance metrics including the one for exact support recovery. In the multiple measurement vector case, Tang and Nehorai [111] employed a hypothesis testing framework with the likelihood ratio test as the optimal decision rule to study how fast the error probability decays. Sufficient and necessary conditions are further identified in order to guarantee successful support recovery in the asymptotic sense.

In addition, a series of papers have leveraged different information theoretic tools, including rate-distortion theory [105], [49], expander graphs [70], belief propagation and list decoding [96], and low-density parity-check codes [137], to design novel algorithms for sparse signal recovery and to analyze their performances.

In this chapter, a set is a collection of unique objects. Let $\mathbb{R}^m$ denote the m-dimensional real Euclidean space. Let $\mathbb{N} = \{1, 2, 3, ...\}$ denote the set of natural numbers. Let $[k]$ denote the set $\{1, 2, ..., k\}$. The notation $|S|$ denotes the cardinality of set S, $\|\mathbf{x}\|_p$ denotes the ℓ_p norm of a vector $\mathbf{x}$, and $\|A\|_F$ denotes the Frobenius norm of a matrix A. Let $A_{\mathcal{S}}$ denote the submatrix formed by the columns of A indexed by the elements in set $\mathcal{S}$, and let $\underline{A}_{\mathcal{S}}$ denote the submatrix formed by the rows of A indexed by the elements in set $\mathcal{S}$. The expression $f(x) = o(g(x))$ denotes $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$, $f(x) = O(g(x))$ denotes $|f(x)| \leq \alpha|g(x)|$ as $x \to \infty$ for some constant $\alpha > 0$, $f(x) = \Theta(g(x))$ denotes $f(x) = O(g(x))$ and $g(x) = O(f(x))$, $f(x) = \Omega(g(x))$ denotes $g(x) = O(f(x))$, and $f(x) = \omega(g(x))$ denotes $g(x) = o(f(x))$.

2.2 Formal Definition of the Problem

Let $W \in \mathbb{R}^{k\times l}$, where $w_{i,j} \neq 0$ for $i \in [k], j \in [l]$. Let $\mathbf{S} = [S_1, ..., S_k]^\intercal \in \mathbb{N}^k$ be such that S_1, ..., S_k are chosen uniformly at random from $[m]$ without replacement.

Then, the signal of interest $X = X(W, \mathbf{S})$ is generated as

$$X_{s,i} = \begin{cases} w_{j,i} & \text{if } s = S_j, \\ 0 & \text{if } s \notin \{S_1, ..., S_k\}. \end{cases} \tag{2.1}$$

The support of X, denoted by supp(X), is the set of indices corresponding to the nonzero rows of X, i.e., supp$(X) = \{S_1, ..., S_k\}$. According to the signal model (2.1), $|\text{supp}(X)| = k$. We assume k is known.

We measure X through the linear operation

$$Y = AX + Z \tag{2.2}$$

where $A \in \mathbb{R}^{n\times m}$ is the measurement matrix, $Z \in \mathbb{R}^{n\times l}$ is the measurement noise, and $Y \in \mathbb{R}^{n\times l}$ is the noisy measurement. We assume that the elements of A are generated independently and identically distributed (i.i.d.) according to the Gaussian distribution $\mathcal{N}(0, \sigma_a^2)$, and the noise $Z_{i,j}$ are i.i.d. according to $\mathcal{N}(0, \sigma_z^2)$. We assume σ_a^2 and σ_z^2 are known.

Upon observing the noisy measurement Y, the goal is to recover the indices of the nonzero rows of X. A support recovery map is defined as

$$d : \mathbb{R}^{n\times l} \longmapsto 2^{[m]}. \tag{2.3}$$

Given the signal model (2.1), the measurement model (2.2), and the support recovery map (2.3), we define the average probability of error as

$$\mathsf{P}\{d(Y) \neq \text{supp}(X(W, \mathbf{S}))\}$$

for each (unknown) signal value matrix $W \in \mathbb{R}^{k \times l}$. Note that the probability is averaged over the randomness of locations of the nonzero rows $\mathbf{S}$, the measurement matrix A, and the measurement noise Z.

2.3 A Multiuser Information Theoretic Perspective on Sparse Signal Recovery

In this section, we introduce an important interpretation of the problem of sparse signal recovery by relating it to a single-input multiple-output multiple access channel communication problem. This relationship motivates the intuition behind our main results and facilitates the development of the proof techniques.

2.3.1 Brief Review on the Single-Input Multiple-Output MAC

Consider the following wireless communication scenario. Suppose k senders wish to transmit information to a set of l common receivers. Each sender i has access to a codebook $\mathscr{C}^{(i)} = \{\mathbf{c}_1^{(i)}, \mathbf{c}_2^{(i)}, ..., \mathbf{c}_{m^{(i)}}^{(i)}\}$, where $\mathbf{c}_j^{(i)} \in \mathbb{R}^n$ is a codeword and $m^{(i)}$ is the number of codewords in the codebook. The rate for the ith sender is $R^{(i)} = (\log m^{(i)})/n$. To transmit information, each sender chooses a codeword from its codebook, and all senders transmit their codewords simultaneously to l receivers leading to the single-input multiple-output (SIMO) multiple access channel (MAC) communication problem:

$$Y_{j,i} = h_{j,1}X_{1,i} + h_{j,2}X_{2,i} + \cdots + h_{j,k}X_{k,i} + Z_{j,i}, \quad i \in [n],\ j \in [l] \tag{2.4}$$

where $X_{q,i}$ denotes the input symbol from sender q to the channel at time i, $h_{j,q}$ denotes the channel gain between sender q and receiver j, $Z_{j,i}$ is the channel noise i.i.d. according to $\mathcal{N}(0, \sigma_z^2)$, and $Y_{j,i}$ is the channel output at receiver j at time i.

After receiving $Y_{j,1,}, ..., Y_{j,n}$ at each receiver $j \in [l]$, the receivers work jointly to determine the codewords transmitted by each sender. Since the senders interfere with each other, there is an inherent tradeoff among their operating rates. The notion of capacity region is introduced to capture this tradeoff by characterizing all possible rate tuples $(R^{(1)}, R^{(2)}, ..., R^{(k)})$ at which reliable communication can be achieved with diminishing error probability of decoding. By assuming each sender obeys the power constraint $\|\mathbf{c}_j^{(i)}\|_2^2/n \leq \sigma_c^2$ for all $j \in [m^{(i)}]$ and all $i \in [k]$, the capacity region of a SIMO MAC with known channel gains [57] is

$$\left\{ (R^{(1)}, ..., R^{(k)}) : \sum_{i \in \mathcal{T}} R^{(i)} \leq \frac{1}{2} \log \det \left(I + \frac{\sigma_c^2}{\sigma_z^2} \sum_{i \in \mathcal{T}} \mathbf{h}_i \mathbf{h}_i^\intercal \right), \forall\, \mathcal{T} \subseteq [k] \right\} \tag{2.5}$$

where $\mathbf{h}_j \triangleq [h_{1,j}, ..., h_{l,j}]^\intercal$ for $j \in [k]$.

2.3.2 Similarities to the Problem of Support Recovery

Based on the measurement model (2.2), we can remove the columns in A which correspond to the zero rows of X, and obtain the following effective form of the measurement procedure

$$\mathbf{Y}_j = X_{S_1,j}\mathbf{A}_{S_1} + \cdots + X_{S_k,j}\mathbf{A}_{S_k} + \mathbf{Z}_j \tag{2.6}$$

for $j \in [l]$. By contrasting (2.6) to the SIMO MAC (2.4), we can draw the following key connections that relate the two problems.

i) **A nonzero row as a sender**: We can view the existence of a nonzero row index S_i as sender i that accesses the SIMO MAC.

ii) **A measurement vector as a receiver**: We can view the existence of a measurement vector $\mathbf{Y}_j$ as a measurement at receiver j. The multiple receivers lead to the multiple output (MO) part of the analogy.

iii) $X_{S_i,j}$ **as the channel gain**: The nonzero entry $X_{S_i,j}$, i.e., $w_{i,j}$, plays the role of the channel gain $h_{j,i}$ between sender i and receiver j.

iv) $\mathbf{A}_i$ **as the codeword**: We treat the measurement matrix A as a codebook with each column $\mathbf{A}_i$, $i \in [m]$, as a codeword. Each element of $\mathbf{A}_{S_i}$ is fed one by one through the channel as input symbols from sender i to the l receivers, resulting in n uses of the channel. Since a user transmits a single stream, this leads to the single input (SI) part of the analogy.

v) **Similarity of objectives**: In the problem of sparse signal recovery, we focus on finding the support $\{S_1, ..., S_k\}$ of the signal. In the problem of MAC communication, the receiver needs to determine the indices of codewords, i.e., $S_1, ..., S_k$, that are transmitted by senders.

Based on the abovementioned aspects, the two problems share significant similarities which enable the potential of leveraging multiuser information theoretic approaches for performance analysis of support recovery of sparse signals. However, as we shall see next, there are domain specific differences between the support recovery problem and the channel coding problem that should be addressed accordingly to rigorously apply the information theoretic approaches.

2.3.3 Key Differences

1. **Common codebook**: In MAC communication, each sender uses its own codebook. However, in sparse signal recovery, the "codebook" A is shared by all "senders." All senders choose their codewords from the same codebook and hence operate at the same rate. Different senders will not choose the same codeword, or they will collapse into one sender.

2. **Unknown channel gains**: In MAC communication, the capacity region (2.5) is valid assuming that the receivers know the channel gain $\mathbf{h}_j, j \in [k]$ [118]. In contrast, for sparse signal recovery, X_{S_i} is actually unknown and needs to be estimated. Although coding techniques and capacity results are available for commu-

nication with channel uncertainty, a closer examination indicates that those results are not directly applicable to our problem. For instance, channel training with pilot symbols is a common practice to combat channel uncertainty [66]. However, it is not obvious how to incorporate the training procedure into the measurement model (2.2), and hence the related results are not directly applicable.

Once these differences are properly accounted for, the connection between the problems of sparse signal recovery and multiuser communication makes available a variety of information theoretic tools for handling performance issues pertaining to the problem of support recovery. Based on techniques that are rooted in channel coding, but suitably modified to deal with the differences, we will present the our results in the sections to follow.

2.4 Support Recovery with SMV: Main Results and Implications

In this section, we discuss the performance limits of support recovery with single measurement vector (SMV), i.e., $l = 1$. Specifically, we use $\mathbf{X}$ to denote X, $\mathbf{Y}$ to denote Y, $\mathbf{Z}$ to denote Z, and $\mathbf{w}$ to denote W, for the reason that all these variables are columns vectors.

2.4.1 Fixed Number of Nonzero Entries

To discover the precise impact of the values of the nonzero entries on support recovery, we consider the support recovery of a sequence of sparse signals generated with the same signal value vector $\mathbf{w}$. In particular, we assume that k is fixed. Define the auxiliary quantity

$$c(\mathbf{w}) \triangleq \min_{\mathcal{T}\subseteq[k]} \left[\frac{1}{2|\mathcal{T}|} \log\left(1 + \frac{\sigma_a^2}{\sigma_z^2}\sum_{j\in\mathcal{T}} w_j^2\right)\right]. \tag{2.7}$$

For example, when $k = 2$,

$$\begin{aligned} c(w_1, w_2) = \min &\left[\frac{1}{2}\log\left(1 + \frac{\sigma_a^2 w_1^2}{\sigma_z^2}\right), \frac{1}{2}\log\left(1 + \frac{\sigma_a^2 w_2^2}{\sigma_z^2}\right),\right. \\ &\qquad\qquad\qquad \left.\frac{1}{4}\log\left(1 + \frac{\sigma_a^2(w_1^2 + w_2^2)}{\sigma_z^2}\right)\right]. \end{aligned}$$

We can see from Section 2.3 that this quantity is closely related to the 2-sender multiple access channel capacity with equal-rate constraint.

The following two theorems summarize our main results under this setup. The subscript in n_m denotes possible dependence between n and m. The proof of the theorems are presented in Sections 2.8.1 and 2.8.2, respectively.

Theorem 1. *If*

$$\limsup_{m\to\infty} \frac{\log m}{n_m} < c(\mathbf{w}) \tag{2.8}$$

then there exists a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{[m]}$, *such that*

$$\lim_{m\to\infty} \mathsf{P}\{d^{(m)}(\mathbf{Y}) \neq \mathit{supp}(\mathbf{X}(\mathbf{w},\mathbf{S}))\} = 0. \tag{2.9}$$

Theorem 2. *If*

$$\limsup_{m\to\infty} \frac{\log m}{n_m} > c(\mathbf{w}) \tag{2.10}$$

then for any sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{[m]}$, *we have*

$$\liminf_{m\to\infty} \mathsf{P}\{d^{(m)}(\mathbf{Y}) \neq \mathit{supp}(\mathbf{X}(\mathbf{w},\mathbf{S}))\} > 0. \tag{2.11}$$

We provide the following observations. First, Theorems 1 and 2 together indicate that $n = (\log m)/(c(\mathbf{w}) \pm \epsilon)$ is sufficient and necessary for exact support recovery. The constant $c(\mathbf{w})$ is explicitly characterized, capturing the role of all nonzero entries of a sparse signal in support recovery. Second, the proof of Theorem 2 for the necessary condition employs the assumption that the values of the nonzero entries are known. Immediately, it follows that even if the values of the nonzero entries are known, the sufficient condition for successfully recovering the support is still given by (2.8). This observation indicates that the unknown channel gain problem indeed does not pose a serious obstacle in support recovery for the case of fixed k. Further, the benefit of exploiting the connection between sparse signal recovery and multiple access communication is also supported by the theorems. Resorting to channel capacity results enables us to explicitly extract the constant $c(\mathbf{w})$ and obtain the tight sufficient and necessary conditions.

2.4.2 Growing Number of Nonzero Entries

Next, we consider the support recovery for the case where the number of nonzero entries k grows with the dimension of the signal m. We assume that the magnitude of a nonzero entry is bounded from both below and above.

First, we present a sufficient condition for exact support recovery. The proof can be found in Section 2.8.3.

Theorem 3. *Let* $\{\mathbf{w}^{(m)}\}_{m=1}^{\infty}$ *be a sequence of vectors satisfying* $\mathbf{w}^{(m)} \in \mathbb{R}^{k_m}$ *and* $0 < w_{\mathit{min}} \leq |w_j^{(m)}| \leq w_{\mathit{max}} < \infty$ *for all* $j \in [k_m], m \geq 1$. *If*

$$\limsup_{m\to\infty} \frac{1}{n_m} \max_{j\in[k_m]} \left[\frac{6k_m \log k_m + 2j \log \frac{me}{j}}{\log\left(\frac{j w_{\mathit{min}}^2 \sigma_a^2}{\sigma_z^2} + 1\right)} \right] < 1 \tag{2.12}$$

then there exists a sequence of support recovery maps $\{d^{(m)}\}_{m=1}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{[m]}$, *such that*

$$\lim_{m\to\infty} \mathsf{P}\{d^{(m)}(A\mathbf{X}(\mathbf{w}^{(m)}, \mathbf{S}) + \mathbf{Z}) \neq \mathit{supp}(\mathbf{X}(\mathbf{w}^{(m)}, \mathbf{S}))\} = 0.$$

Note that, according to our proof technique, the upper bound $w_{\max}$ is not needed for performing support recovery, and it does not appear in the sufficient condition above. In the proof, however, we use the assumption that the nonzero signal values are uniformly bounded from above to show that the probability of error tends to zero as $m \to \infty$. To better understand Theorem 3, we present the following implication of (2.12) that shows the tradeoffs between the order of n versus m and k.

Corollary 1. *Under the assumption of Theorem 3,*

$$\lim_{m\to\infty} \mathsf{P}\{d^{(m)}(A\mathbf{X}(\mathbf{w}^{(m)}, \mathbf{S}) + \mathbf{Z}) \neq \textit{supp}(\mathbf{X}(\mathbf{w}^{(m)}, \mathbf{S}))\} = 0$$

provided that

$$n = \max\left\{\Omega(k \log k), \Omega\left(\frac{k}{\log k} \log \frac{m}{k}\right)\right\}.$$

In particular, we have the following:

1. *When $k = O(e^{\sqrt{\log m}})$, the sufficient number of measurements is $n = \Omega(\frac{k}{\log k} \log \frac{m}{k})$.*

2. *When $\omega(e^{\sqrt{\log m}}) \leq k \leq \Theta(m)$, the sufficient number of measurements is $n = \Omega(k \log m)$.*

Table 2.1 summarizes the sufficient conditions on n paired with different relations between k and m in Corollary 1.

In the existing literature, Wainwright [122], Akçakaya and Tarokh [2], and Rad [98] derived sufficient conditions for exact support recovery. Under the same assumption of Theorem 3, the sufficient conditions presented in these papers, respectively, are summarized in Table 2.2:

Table 2.1: Sufficient Conditions for Support Recovery in Different Sparsity Regions

Relation between m and k		Sufficient n
$k = o(m)$	$k = O(e^{\sqrt{\log m}})$	$n = \Omega(\frac{k}{\log k} \log \frac{m}{k})$
	$\omega(e^{\sqrt{\log m}}) \leq k \leq o(m)$	$n = \Omega(k \log m)$
$k = \Theta(m)$		$n = \Omega(k \log m)$

Table 2.2: Sufficient Conditions for Support Recovery in Existing Literature

	$k = o(m)$	$k = \Theta(m)$
Wainwright [122]	$n = \Omega(k \log \frac{m}{k})$	$n = \Omega(m)$
Akçakaya et al. [2]	$n = \Omega(k \log(m - k))$	$n = \Omega(m)$
Rad [98][1]	$n = \max\{\Omega(\frac{k}{\log k} \log \frac{m}{k}), \Omega(k)\}$	$n = \Omega(m)$

To compare the results, we first examine the case of $k = o(m)$ (i.e., sublinear sparsity). Note that in the regime where $k = O(e^{\sqrt{\log m}})$, our sufficient condition on n is among the best existing results. In the remaining sublinear regime and in the linear regime, i.e., $\omega(e^{\sqrt{\log m}}) \leq k \leq \Theta(m)$, our results are not as tight as the best existing results. More discussions will be provided in Section 2.4.3.

Next, we present a necessary condition, the proof of which can be found in Section 2.8.4.

Theorem 4. *Let* $\{\mathbf{w}^{(m)}\}_{m=1}^{\infty}$ *be a sequence of vectors satisfying* $\mathbf{w}^{(m)} \in \mathbb{R}^{k_m}$ *and* $0 < w_{min} \leq |w_j^{(m)}| \leq w_{max} < \infty$ *for all* $j \in [k_m], m \geq 1$. *If*

$$\limsup_{m \to \infty} \frac{2k_m \log(m/k_m)}{n_m \log\left(\frac{2k_m w_{max}^2 \sigma_a^2}{\sigma_z^2} + 1\right)} > 1 \tag{2.13}$$

then for any sequence of support recovery maps $\{d^{(m)}\}_{m=1}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{[m]}$, *we*

[1]The original result in Corollary 6-6) of [98] may suffer from a typo since it gave two conditions on n. We use Theorem 5 therein to figure out the actual sufficient condition.

have

$$\liminf_{m\to\infty} \mathsf{P}\{d^{(m)}(A\mathbf{X}(\mathbf{w}^{(m)},\mathbf{S})+\mathbf{Z}) \neq \mathit{supp}(\mathbf{X}(\mathbf{w}^{(m)},\mathbf{S}))\} > 0.$$

To compare with existing results under the same assumption[2] of Theorem 4, we first note that when $k = \Theta(m)$ (linear sparsity), Theorem 4 indicates $n = \Omega(\frac{k}{\log k}\log\frac{m}{k})$ as the necessary condition. Compared to the best known sufficient condition $n = \Omega(m)$ (see Table 2.2), there is a nontrivial gap. When $k = o(m)$ (sublinear sparsity), we summarize the necessary conditions developed in previous papers in Table 2.3:

Table 2.3: Necessary Conditions for Support Recovery

	$k = o(m)$
Wainwright [122]	$n = \Omega(\log m)$
Wang et al. [125]	$n = \Omega(\frac{k}{\log k}\log\frac{m}{k})$
Akçakaya et al. [2] [3]	$n = \Omega(\frac{k}{\log k}\log\frac{m}{k})$
Theorem 4	$n = \Omega(\frac{k}{\log k}\log\frac{m}{k})$

In this case, $n = \Omega(\frac{k}{\log k}\log\frac{m}{k})$ is the best known necessary condition.[4]

2.4.3 Further Discussions

We offer more insights into the analytical framework and proof techniques.

[2]The necessary conditions derived in [122], [125], and [2] were originally derived under slightly different assumptions. Here we adapted them to compare the asymptotic orders of n.

[3]This result is implied in [2], by identifying C_4' in Thm. 1.6 therein, and clarifying the order of n. The proof of Thm. 1.6 states that (below its (25)) asymptotically reliable support recovery is not possible if $n < [\log(1+\|\mathbf{w}^{(m)}\|_2^2/\sigma_z^2)]^{-1} mH(k/m) - \log(m+1)$. Note that $mH(k/m) = \Theta(k\log(m/k))$. Hence, we consider $n = \Omega(\frac{k\log(m/k)}{\log k})$ an appropriate necessary condition resulting from the proof in [2].

[4]Note that when $w_{\max} = \infty$, we can show that $n \geq k$ is necessary for both linear and sublinear sparsity [125]. Hence, when $w_{\max} = \infty$, $n = \max\left\{\Omega(\frac{k}{\log k}\log\frac{m}{k}), \Omega(k)\right\}$ is the best known necessary condition.

The sufficient conditions in this paper are derived based on the distance decoding technique which was used in channel decoding problem [76]. In order to perform the distance decoding, the channel gains need to be known or can be estimated. This is in contrary to the fact that the nonzero entries of a sparse signal are unknown, and therefore raises the unknown channel gain problem in Section 2.3.3. To tackle this problem, we employ the following procedure in the proofs for sufficient conditions.

1. Find an estimate of $\|\mathbf{w}\|_2$, and denote it by $\hat{\rho}$.

2. Find a set $\mathcal{Q}$ of points which can be viewed as ϵ-covering of the k-dimensional hypersphere of radius $\hat{\rho}$. By construction of $\mathcal{Q}$, there exists a $\hat{\mathbf{W}} \in \mathcal{Q}$ such that $\|\hat{\mathbf{W}} - \mathbf{w}\|_2 \leq \epsilon$ with high probability.

3. Find $\{\hat{s}_1, \hat{s}_2, ..., \hat{s}_k\} \subseteq [m]$ such that

$$\frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{\hat{s}_j}\right\|_2^2 \leq \sigma_z^2 + \epsilon^2 \sigma_a^2 \tag{2.14}$$

   for some $\hat{\mathbf{W}} \in \mathcal{Q}$. We declare $\{\hat{s}_1, \hat{s}_2, ..., \hat{s}_k\}$ as the estimated support of the sparse signal. As a byproduct, the elements of the corresponding $\hat{\mathbf{W}}$ can be viewed as estimates of the values of the nonzero entries.

The success of this support recovery procedure is closely related to the estimation quality of $\|\mathbf{w}\|_2$ and the cardinality of the set $\mathcal{Q}$. Accordingly, our methodology shows different strength in different regions of sparsity levels. First, in the case for fixed number of nonzero entries, consistent estimation of $\|\mathbf{w}\|_2$ can be obtained, and the cardinality of $\mathcal{Q}$ can be bounded from above. This provides the opportunity to discover the exact sufficient and necessary conditions for successful support recovery. Next, in the case with growing number of nonzero entries, the estimation quality of $\|\mathbf{w}\|_2$ and the cardinality of $\mathcal{Q}$ must be carefully controlled. To this end, the constraint $k = o(n)$, which is implied by Theorem 3, is needed for the estimation of $\|\mathbf{w}\|_2$ to be consistent, and w_{max} as the upper bound for the nonzero magnitudes is needed for controlling the cardinality of $\mathcal{Q}$. Note that for the sublinear sparsity with $k = O(e^{\sqrt{\log m}})$, our sufficient and necessary conditions both indicate $n = \Omega(\frac{k}{\log k} \log \frac{m}{k})$, and hence are tight in terms of order. As k increases with m at a faster rate, our sufficient and necessary conditions have gaps, which is a consequence of the difficulty in consistently estimating $\|\mathbf{w}\|_2$ and handling the large size of $\mathcal{Q}$.

Another interesting region which has been extensively discussed in previous work is the case where $w_{\text{min}} = O(1/\sqrt{k})$ [51, 122, 124]. Although Theorem 4 can be extended to provide a necessary condition for this case, it does not offer improvement upon existing results. Theorem 3 may not be extended to this scenario, which indicates that our analytical technique for proving sufficient conditions is not suited for this scaling.

2.5 Support Recovery with MMV: Main Results and Implications

In this section, we focus on the performance limits of support recovery of sparse signals with MMV. We use the general notation as defined in Section 2.2.

2.5.1 Main Results

Similarly to the case with SMV, we consider the recovery of the nonzero rows of a sequence of sparse signals generated with the same signal value matrix W. In particular, we assume that k and l are fixed. Define the auxiliary quantity

$$c(W) \triangleq \min_{\mathcal{T}\subseteq[k]} \left[\frac{1}{2|\mathcal{T}|} \log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}\right)\right]. \tag{2.15}$$

The following theorems provide the performance limit of support recovery for the MMV problem. The proofs are presented in Sections 2.8.5 and 2.8.6, respectively.

Theorem 5. *If*

$$\limsup_{m\to\infty} \frac{\log m}{n_m} < c(W) \tag{2.16}$$

then there exists a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m\times l} \mapsto 2^{[m]}$, *such that*

$$\lim_{m\to\infty} \mathsf{P}\{d(Y) \neq \textit{supp}(X(W,\mathbf{S}))\} = 0. \tag{2.17}$$

Theorem 6. *If*

$$\limsup_{m\to\infty} \frac{\log m}{n_m} > c(W) \tag{2.18}$$

then for any sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m \times l} \mapsto 2^{[m]}$,

$$\liminf_{m\to\infty} \mathsf{P}\{d(Y) \neq \textit{supp}(X(W,\mathbf{S}))\} > 0. \tag{2.19}$$

Theorems 5 and 6 together indicate that $n = \frac{1}{c(W)\pm\epsilon}\log m$ is the sufficient and necessary number of measurements per measurement vector to ensure asymptotically successful support recovery. The constant $c(W)$ explicitly captures the role of the nonzero entries in the performance limit.

Next, we explore the implications of having multiple measurement vectors. Due to the complicated expression of $c(W)$, we will employ different approximations to make the interpretations more accessible without loss of sufficient insight.

2.5.2 The Low-Noise-Level Scenario

We consider the case where σ_z^2 is sufficiently small. Let $\lambda_{\mathcal{T},i}, \mathcal{T} \subseteq [k]$, denote the ith largest eigenvalue of $\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}$. For a SIMO MAC problem, the sum capacity grows as $\min(k,l)$ leading to significant gains in the task of support recovery. This is captured in the following corollary.

Corollary 2. *For a given* W*, suppose* $\operatorname{rank}(\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}) = \min(|\mathcal{T}|, l)$ *for all* $\mathcal{T} \subseteq [k]$*. For sufficiently small* $\sigma_z^2 > 0$*, there exists a constant* $\alpha \in (0,1)$ *such that if*

$$\lim_{m\to\infty} \frac{\log m}{n_m} < \alpha \cdot \frac{\min(k,l)}{2k} \cdot \log \frac{\sigma_a^2}{\sigma_z^2} \tag{2.20}$$

then there exists a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}, d^{(m)} : \mathbb{R}^{n_m \times l} \mapsto 2^{[m]}$*, such that*

$$\lim_{m\to\infty} \mathsf{P}\{d(Y) \neq \textit{supp}(X(W,\mathbf{S}))\} = 0.$$

Proof. Note that for $\mathcal{T} \subseteq [k]$ with $|\mathcal{T}| \leq l$, $\lambda_{\mathcal{T},i} > 0$ for $i = 1, 2, ..., |\mathcal{T}|$. Thus

$$\begin{aligned}
\frac{1}{2|\mathcal{T}|} \log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}\right) &= \frac{1}{2|\mathcal{T}|} \log \prod_{i=1}^{|\mathcal{T}|}\left(1 + \frac{\sigma_a^2}{\sigma_z^2}\lambda_{\mathcal{T},i}\right) \\
&= \frac{1}{2|\mathcal{T}|} \log \prod_{i=1}^{|\mathcal{T}|}\left(\frac{\sigma_a^2}{\sigma_z^2}\left(\frac{\sigma_z^2}{\sigma_a^2} + \lambda_{\mathcal{T},i}\right)\right) \\
&= \frac{1}{2|\mathcal{T}|}\left[|\mathcal{T}| \cdot \log\frac{\sigma_a^2}{\sigma_z^2} + \sum_{i=1}^{|\mathcal{T}|}\log\left(\frac{\sigma_z^2}{\sigma_a^2} + \lambda_{\mathcal{T},i}\right)\right] \\
&= \frac{1}{2}\log\frac{\sigma_a^2}{\sigma_z^2} \cdot \left(1 + \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\frac{\log\left(\frac{\sigma_z^2}{\sigma_a^2} + \lambda_{\mathcal{T},i}\right)}{\log\frac{\sigma_a^2}{\sigma_z^2}}\right) \\
&= \frac{1}{2}\log\frac{\sigma_a^2}{\sigma_z^2} \cdot \left(1 + O\left(\frac{1}{-\log\sigma_z^2}\right)\right) \\
&\geq \frac{1}{2}\log\frac{\sigma_a^2}{\sigma_z^2} \cdot \alpha_{\mathcal{T}}
\end{aligned}$$

for some $\alpha_{\mathcal{T}} \in (0, 1)$. For any possible $\mathcal{T} \subseteq [k]$ with $|\mathcal{T}| > l$, $\lambda_{\mathcal{T},i} > 0$ for $i = 1, 2, ..., l$. Then, we have similarly

$$\begin{aligned}
\frac{1}{2|\mathcal{T}|} \log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}\right) &= \frac{l}{2|\mathcal{T}|}\log\frac{\sigma_a^2}{\sigma_z^2} \cdot \left(1 + O\left(\frac{1}{-\log\sigma_z^2}\right)\right). \\
&\geq \frac{l}{2|\mathcal{T}|}\log\frac{\sigma_a^2}{\sigma_z^2} \cdot \alpha_{\mathcal{T}}
\end{aligned}$$

Thus, if $k \leq l$

$$\min_{\mathcal{T}\subseteq[k]}\left[\frac{1}{2|\mathcal{T}|}\log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}\right)\right] \geq \frac{1}{2}\log\frac{\sigma_a^2}{\sigma_z^2} \cdot \min_{\mathcal{T}\subseteq[k]}\alpha_{\mathcal{T}} \tag{2.21}$$

and if $k > l$

$$\min_{\mathcal{T}\subseteq[k]} \left[\frac{1}{2|\mathcal{T}|} \log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2} \underline{W}_{\mathcal{T}}^\intercal \underline{W}_{\mathcal{T}}\right)\right] \geq \frac{l}{2k} \log\frac{\sigma_a^2}{\sigma_z^2} \cdot \min_{\mathcal{T}\subseteq[k]} \alpha_{\mathcal{T}}. \tag{2.22}$$

Combining (2.21) and (2.22) and applying Theorem 5 conclude the proof. □

Corollary 2 indicates the following observations. First, as the measurement noise level σ_z^2 approaches zero, the term $\frac{\min(k,l)}{2k} \log \frac{\sigma_a^2}{\sigma_z^2}$ exerts a major influence on the sufficient condition (2.20). The nonzero signal matrix W plays its role mainly through the ranks of its row-wise submatrices, which are ensured to be full rank according to the technical assumption that $\operatorname{rank}(\underline{W}_{\mathcal{T}}^\intercal \underline{W}_{\mathcal{T}}) = \min(|\mathcal{T}|, l)$ for any $\mathcal{T} \subseteq [k]$.

Second, by rearranging the terms in (2.20), we obtain

$$m = \left(\frac{\sigma_a^2}{\sigma_z^2}\right)^{\alpha\cdot\min(k,l)\cdot\frac{n}{2k}}$$

which corresponds to the maximum number of columns of A that still yields a diminishing error probability in support recovery. Specifically, the term $\min(k, l)$ reveals the following insight. In the scenario with sufficiently small σ_z^2, for the challenging problem where the number of measurement vectors is less than the number of nonzero rows, i.e., $l < k$, adding one more measurement vector can lead to a much larger upper bound on the manageable number of columns of A. On the other hand, when $k \leq l$, the problem is much simpler and adding more measurement vectors may not significantly increase the manageable size of A. From an algorithmic point of view, subspace based methods can be used to recover the support.

2.5.3 The Role of the Nonzero Signal Matrix

Next, we take a closer examination on the role of the nonzero signal matrix W in support recovery with MMV. We consider two different cases. In the first case, W consists of l identical columns. The following corollary states the corresponding sufficient condition for support recovery.

Corollary 3. *Suppose $W \in \mathbb{R}^{k \times l}$ has identical columns, i.e., $W = [\mathbf{w}, ..., \mathbf{w}]$, for some $\mathbf{w} \in \mathbb{R}^k$ with all entries being nonzero. If*

$$\lim_{m \to \infty} \frac{\log m}{n_m} < \min_{\mathcal{T} \subseteq [k]} \frac{1}{2|\mathcal{T}|} \log \left(1 + l \cdot \frac{\sigma_a^2}{\sigma_z^2} \|\mathbf{w}_{\mathcal{T}}\|_2^2 \right) \tag{2.23}$$

then there exists a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m \times l} \mapsto 2^{[m]}$, such that

$$\lim_{m \to \infty} \mathsf{P}\{d(Y) \neq \mathit{supp}(X(W, \mathbf{S}))\} = 0.$$

Proof. Note that, for any $\mathcal{T} \subseteq [k]$,

$$\begin{aligned}
\log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}\right) &= \log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2}[\mathbf{w}_{\mathcal{T}}, ..., \mathbf{w}_{\mathcal{T}}]^{\intercal}[\mathbf{w}_{\mathcal{T}}, ..., \mathbf{w}_{\mathcal{T}}]\right) \\
&= \log\det\left(I + \frac{\sigma_a^2}{\sigma_z^2}\|\mathbf{w}_{\mathcal{T}}\|_2^2\mathbf{1}\cdot\mathbf{1}^{\intercal}\right) \\
&= \log\left(1 + l\cdot\frac{\sigma_a^2}{\sigma_z^2}\|\mathbf{w}_{\mathcal{T}}\|_2^2\right)
\end{aligned}$$

where $\mathbf{1} \in \mathbb{R}^l$ is a vector of all ones. Applying Theorem 5 completes the proof. □

Based on (2.23), the effect of having l identical nonzero signal vectors is equivalent to decreasing the noise level by a factor of l, compared to the problem with SMV. This is in accordance with the intuition that when the underlying signals remain the same, taking more measurement vectors provides an opportunity to average down the measurement noise level. We hasten to add that identical columns are unlikely in practice. Even small changes in the coefficients can lead to a full rank matrix, leading to significant benefits in the high signal-to-noise ratio (SNR) case.

In the second case, we consider a more optimistic situation, which is summarized in the following corollary.

Corollary 4. *Suppose $W = [\mathbf{w}_1, \mathbf{w}_2] \in \mathbb{R}^{k\times 2}$, where k is even, $\mathbf{w}_1 = \mathbf{1} \in \mathbb{R}^k$, and $\mathbf{w}_2$ is defined as*

$$w_{i,2} = \begin{cases} 1 & \textit{if } 1 < i \leq \frac{k}{2}, \\ -1 & \textit{if } \frac{k}{2} < i \leq k. \end{cases} \tag{2.24}$$

If

$$\lim_{m\to\infty} \frac{\log m}{n} < \frac{1}{k} \log\left(1 + k \cdot \frac{\sigma_a^2}{\sigma_z^2}\right) \tag{2.25}$$

then there exists a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}, d^{(m)} : \mathbb{R}^{n_m \times l} \mapsto 2^{[m]}$, *such that*

$$\lim_{m\to\infty} \mathsf{P}\{d(Y) \neq \textit{supp}(X(W, \mathbf{S}))\} = 0.$$

Proof. Please see Section 2.8.7. □

For the ease of illustration, we compare the performances among the problems with (i) SMV where $W = \mathbf{1} \in \mathbb{R}^{k\times 1}$, (ii) MMV where $W = [\mathbf{1}, \mathbf{1}] \in \mathbb{R}^{k\times 2}$, and (iii) MMV where W is defined in Corollary 4, for an even k.[5] Table 2.4 summarizes the results.

Table 2.4: Bounds under different scenarios.

	lower bound on n	upper bound on m
(i) SMV ($W = \mathbf{1}$)	$n > \frac{\log m}{\frac{1}{2k}\log\left(1+\frac{k\sigma_a^2}{\sigma_z^2}\right)}$	$m < \left(1 + \frac{k\sigma_a^2}{\sigma_z^2}\right)^{\frac{n}{2k}}$
(ii) MMV ($W = [\mathbf{1}, \mathbf{1}]$)	$n > \frac{\log m}{\frac{1}{2k}\log\left(1+2\cdot\frac{k\sigma_a^2}{\sigma_z^2}\right)}$	$m < \left(1 + 2\cdot\frac{k\sigma_a^2}{\sigma_z^2}\right)^{\frac{n}{2k}}$
(iii) MMV (W as defined in Cor. 4)	$n > \frac{\log m}{\frac{1}{k}\log\left(1+\frac{k\sigma_a^2}{\sigma_z^2}\right)}$	$m < \left(1 + \frac{k\sigma_a^2}{\sigma_z^2}\right)^{\frac{n}{k}}$

Based on this table, we have the following observations for this specific setup.

[5] Note that $\|\mathbf{w}_1\|_2 = \|\mathbf{w}_2\|_2$, which can be viewed as a way of normalization to make comparison meaningful.

First, compared with the SMV problem, having MMV can improve the performance of support recovery by enabling a relaxed condition on the number of measurements n. Equivalently, for the same number of measurements per measurement vector, the MMV setup permits a measurement matrix A with more columns. Second, the performance improved by having MMV is closely related to $c(W)$, and it can be quite different for different nonzero signal value matrices. In case (ii), we achieve a moderate performance gain which is equivalent to reducing the noise level by half. On the contrary, in example (iii), a larger performance gain can be achieved due to the structure of the nonzero signal value matrix. In summary, these examples are specially constructed as representative cases to illustrate the effect of the nonzero signal value matrix W in support recovery. Generally, the difficulty of a support recovery problem is inherently determined by the model parameters, especially the nonzero value matrix W, and Theorems 5 and 6 together characterize their exact roles.

2.5.4 A Generalization of W

Thus far, we have assumed $w_{i,j} \neq 0$ for all $i \in [k], j \in [l]$ in the discussion above. Now, we wish to generalize W in the following manner: for each $i \in [k]$, there exist a $j \in [l]$ such that $w_{i,j} \neq 0$; meanwhile, for each $j \in [l]$, there exist a $i \in [k]$ such that $w_{i,j} \neq 0$. This relaxed assumption indicates that neither a zero row nor a zero column exists but zero elements are allowed in W, as opposed to the original assumption that all elements of W are nonzeros. Accordingly,

$$\text{supp}(X) = \bigcup_{j=1}^{l} \text{supp}(\mathbf{X}_j)$$

which means the support of X is equal to the union of the supports of all columns of X. This general scenario is also considered in [47] and [89]. Following the proofs for Theorems 5 and 6, one can readily see that the two theorems still hold in this case.

It is worthwhile to note that having more measurement vectors does not necessarily result in performance improvement. To illustrate this point, we construct a simple example. Let $W^{(1)} = [0.1, 5]^\intercal$, $W^{(2)} = \begin{bmatrix} 0.1 & 0 \\ 5 & 6 \end{bmatrix}$, and $\frac{\sigma_a^2}{\sigma_z^2} = 10$. As a result, $c(W^{(1)}) = c(W^{(2)}) = \frac{1}{2}\log 1.1$. This means that the performance limits for these two setups are the same. Intuitively, by inspecting the definition of $c(W)$, it can be seen that if a submatrix composed of certain rows of W is ill-conditioned, the minimization inside $c(W)$ may likely be determined by that submatrix. Hence, for an extra measurement vector to benefit support recovery, this measurement vector should correspond to a column of W whose presence improves the small eigenvalues of the previous worst-case submatrix that causes the performance bottleneck. The observations are reminiscent of some of the intuition developed in space time wireless communication systems [95]. The l receivers can be viewed an a l antenna receiver and it is known that the rank of the channel matrix plays an important role in the high SNR case. The correlation between the channel gains is not as harmful in this context. The gains of having multiple receive antennas is lower at low SNR.

2.5.5 Relation to Performance Limit of Multivariate Group Lasso

We compare our performance limits to the performance limits of the multivariate group Lasso algorithm [89]. Our goal is to understand the potential gap between the performance of a practical algorithm and the fundamental performance limit, and to suggest improvement in algorithm design.

We note that the model employed in [89] is similar to the model assumptions underlying the measurement model (2.2). Sufficient and necessary conditions are derived therein for multivariate group Lasso to successfully recover the support of the signal in the presence of noise, as m, n, and k grow to infinity in certain manner.[6] This is different from our assumption that k is fixed. Although a direct comparison seems difficult, we provide the following intuitive discussion. Note that Example 1 in [89, Section 2.3] considered the case for identical regression, which means the nonzero signal matrix W has identical columns. The conclusion therein is that multivariate group Lasso offered no performance improvement under the MMV formulation compared with using ordinary Lasso on an alternative SMV formulation with each measurement vector. However, our Corollary 3 indicates that the effect of having l identical columns in W is equivalent to lowering the noise level by a factor of l.[7] The different performances

[6]Note that in the preprint version [89], it is stated, at the end of its Section 3.3, that the requirement on k growing to infinity can be removed. The remark therein provided an alternative probability upper bound for the intermediate term T_1 such that this bound can drop to zero even for a fixed k. However, it seems that the other intermediate term T_2 still relies on a probability upper bound that involves a term scaling as $\exp(-\frac{k}{2})$, which requires an increasing k to drive it to zero.

[7]Note that this noise scaling behavior can be achieved by practical algorithms. As a simple example,

indicated by multivariate group Lasso and the information theoretic analysis lead to the following observation. In general, if the sparse signal possesses strong structural property, an algorithm needs to take advantage of this characteristic to achieve performance improvement. As an example, AR-SBL [138] is developed based on the assumption that the columns of W are drawn from an auto-regressive process, and it explicitly attempts to learn this correlation structure. Based on the experimental study presented in [138], notable performance improvement in support recovery was observed when such correlation is present at different degrees, including the case where the columns of W were almost identical.

2.6 Random Nonzero Entries

We have discussed the performance limits of support recovery of sparse signals in the scenario where the signal value matrix W is fixed. In this section, we extend our discussion to the case where W is randomly generated according to certain probability distribution. For simplicity, we consider the SMV case, i.e., $l = 1$. The notation follows Section 2.4, except $\mathbf{W}$ is used to denote the random column vector of nonzero signal values.

Interestingly, the model (2.2) (for $l = 1$) with this new assumption can now be

one can average over all measurements to obtain $\mathbf{Y}_{\text{avg}} = \frac{1}{l}\sum_{i=1}^{l}\mathbf{Y}_i$. We solve a SMV problem based on A and $\mathbf{Y}_{\text{avg}}$. Then, we treat the support of the solution as the set of indices of the nonzero rows for the signal matrix in the original MMV problem. This algorithm will outperform the multivariate group lasso in the case for identical regression, but not necessarily in other cases.

contrasted to a MAC with random channel gains

$$Y_i = H_1 X_{1,i} + H_2 X_{2,i} + \cdots + H_k X_{k,i} + Z_i,\ i \in [n]. \tag{2.26}$$

The difference between (2.26) and (2.4) with $l = 1$ is that the channel gains H_i are random variables in this case. Specifically, in order to contrast the problem of support recovery of sparse signals, H_i should be considered as being realized once and then kept fixed during the entire channel use. This channel model is usually termed as a slow fading channel [118].

The following theorem states the performance limit of support recovery of sparse signals with random signal activities.

Theorem 7. *Suppose* $\mathbf{W}$ *has bounded support, and* $\limsup_{m\to\infty} \frac{\log m}{n_m} = r$. *Then, there exists a sequence of support recovery maps* $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{[m]}$, *such that*

$$\limsup_{m\to\infty} \mathsf{P}\{d^{(m)}(A^{(m)}\mathbf{X}(\mathbf{W},\mathbf{S}) + \mathbf{Z}) \neq \mathit{supp}(\mathbf{X})\} \leq \mathsf{P}\{c(\mathbf{W}) \leq r\}$$

where $c(\mathbf{W})$ *is defined as in (2.7).*

Proof. Note that

$$
\begin{aligned}
&\limsup_{m\to\infty} \mathsf{P}\{d^{(m)}(A^{(m)}\mathbf{X}(\mathbf{W},\mathbf{S})+\mathbf{Z}) \neq \text{supp}(\mathbf{X})\} \\
&= \limsup_{m\to\infty} \int_{\mathbf{w}} \mathsf{P}\{d^{(m)}(A^{(m)}\mathbf{X}(\mathbf{w},\mathbf{S})+\mathbf{Z}) \neq \text{supp}(\mathbf{X})\} \cdot dF(\mathbf{w}) \\
&= \limsup_{m\to\infty} \int_{\mathbf{w}:c(\mathbf{w})>r} \mathsf{P}\{d^{(m)}(A^{(m)}\mathbf{X}+\mathbf{Z}) \neq \text{supp}(\mathbf{X})\} \cdot dF(\mathbf{w}) \\
&\quad + \limsup_{m\to\infty} \int_{\mathbf{w}:c(\mathbf{w})\le r} \mathsf{P}\{d^{(m)}(A^{(m)}\mathbf{X}+\mathbf{Z}) \neq \text{supp}(\mathbf{X})\} \cdot dF(\mathbf{w}) \\
&\le \int_{\mathbf{w}:c(\mathbf{w})>r} \limsup_{m\to\infty} \mathsf{P}\{d^{(m)}(A^{(m)}\mathbf{X}+\mathbf{Z}) \neq \text{supp}(\mathbf{X})\} \cdot dF(\mathbf{w}) + \int_{\mathbf{w}:c(\mathbf{w})\le r} dF(\mathbf{w}) \qquad (2.27) \\
&\le \mathsf{P}\{c(\mathbf{W}) \le r\} \qquad (2.28)
\end{aligned}
$$

where (2.27) follows from Fatou's lemma [102] and (2.28) follows by applying the proof of Theorem 1 to the integrand. □

Theorem 7 implies that generally, rather than having a diminishing error probability, we have to tolerate certain error probability which is upper-bounded by $\mathsf{P}(c(\mathbf{W}) \le r)$, when the nonzero values are randomly generated. Conversely, in order to design a system with probability of success at least $(1-p)$, one can find r that satisfies $\mathsf{P}(c(\mathbf{W}) \le r) \le p$. Note that $\mathsf{P}\{c(\mathbf{W}) \le r\}$ can be viewed as the outage probability of a slow fading MAC given the target rate r of each sender [118]. Thus, $\mathsf{P}\{c(\mathbf{W}) \le r\}$ represents the probability that the channel gains are realized too poorly to support the target rate.

Note that similar results can also be derived for the MMV case. We would not pursue this direction in detail here.

2.7 Acknowledgements

Chapter 2 is, in part, a reprint of material in the paper "Limits on Support Recovery of Sparse Signals via Multiple Access Communication Techniques," Y. Jin, Y.-H. Kim, B. D. Rao, *submitted to IEEE Transactions on Information Theory*, 2010, the paper "Performance Tradeoffs for Exact Support Recovery of Sparse Signals," Y. Jin, Y.-H. Kim, and B. D. Rao, *published in the proceedings of the International Symposium on Information Theory*, 2010, the paper "Sparse Signal Recovery in Presence of Multiple Measurement Vectors," Y. Jin and B. D. Rao, *in preparation for IEEE Transactions on Information Theory*, and the paper "On the Role of the Properties of the Non-zero Entries on Sparse Signal Recovery," Y. Jin and B. D. Rao, *published in proceedings of the Asilomar Conference on Signals, Systems and Computers*, Nov. 2010. In all cases, I was the primary author, and B. D. Rao supervised the research. Y.-H. Kim contributed to the paper "Limits on Support Recovery of Sparse Signals via Multiple Access Communication Techniques" and the paper "Performance Tradeoffs for Exact Support Recovery of Sparse Signals."

2.8 Appendices

2.8.1 Proof of Theorem 1

The proof of Theorem 1 employs the distance decoding technique [76]. Let $\mathbf{A}_j^{(m)}$ denote the jth column of $A^{(m)}$.

For simplicity of exposition, we describe the support recovery procedure for two distinct cases on the number of nonzero entries.

Case 1: $k = 1$. In this case, the signal of interest is $\mathbf{X} = \mathbf{X}(w_1, S_1)$. Consider the following support recovery procedures. Fix $\epsilon > 0$. First form an estimate $\hat{\rho}$ of $|w_1|$ as

$$\hat{\rho} \triangleq \sqrt{\frac{\left|\frac{1}{n_m}\|\mathbf{Y}\|_2^2 - \sigma_z^2\right|}{\sigma_a^2}}. \tag{2.29}$$

Declare that $\hat{s}_1 \in [m]$ is the estimated location for the nonzero entry, i.e., $d^{(m)}(\mathbf{Y}) = \{\hat{s}_1\}$, if it is the unique index such that

$$\frac{1}{n_m}\|\mathbf{Y} - (-1)^q \hat{\rho}\mathbf{A}_{\hat{s}_1}^{(m)}\|_2^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2 \tag{2.30}$$

for either $q = 1$ or $q = 2$. If there is none or more than one, pick an arbitrary index.

We now analyze the average probability of error

$$\mathsf{P}(\mathcal{E}) = \mathsf{P}\{d^{(m)}(\mathbf{Y}) \neq \{S_1\}\}$$

where the expectation is taken with respect to the random measurement matrix $A^{(m)}$. Due to the symmetry in the problem and the measurement matrix generation, we assume without loss of generality $S_1 = 1$, that is,

$$\mathbf{Y} = w_1 \mathbf{A}_1^{(m)} + \mathbf{Z}$$

for some w_1. In the following analysis, we drop superscripts and subscripts on m for notational simplicity when no ambiguity arises. Define the events for $s \in [m]$

$$\mathcal{E}_s \triangleq \left\{ \exists q \in \{1, 2\} \text{ such that } \frac{1}{n} \|\mathbf{Y} - (-1)^q \hat{\rho} \mathbf{A}_s\|_2^2 \leq \sigma_z^2 + \epsilon^2 \sigma_a^2 \right\}.$$

Then

$$\mathsf{P}(\mathcal{E}) \leq \mathsf{P}\left(\mathcal{E}_1^c \cup (\cup_{s=2}^m \mathcal{E}_s)\right). \tag{2.31}$$

Let

$$\mathcal{E}_{\text{aux}} \triangleq \{\hat{\rho} - |w_1| \in (-\epsilon, \epsilon)\} \cap \left\{ \frac{1}{n} \|\mathbf{Y}\|_2^2 - [w_1^2 \sigma_a^2 + \sigma_z^2] \in (-\epsilon, \epsilon) \right\}.$$

Then, by the union of events bound and the fact that $\mathcal{A}^c \cup \mathcal{B} = \mathcal{A}^c \cup (\mathcal{B} \cap \mathcal{A})$,

$$\mathsf{P}(\mathcal{E}) \leq \mathsf{P}(\mathcal{E}_{\text{aux}}^c) + \mathsf{P}(\mathcal{E}_1^c) + \sum_{s=2}^{m} \mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}}). \tag{2.32}$$

We bound each term in (2.32). First, by the weak law of large numbers (LLN), $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_{\text{aux}}^c) = 0$. Next, we consider $\mathsf{P}(\mathcal{E}_1^c)$. If $w_1 > 0$,

$$\begin{aligned}\frac{1}{n}\|\mathbf{Y} - \hat{\rho}\mathbf{A}_1\|_2^2 &= \frac{1}{n}\|w_1\mathbf{A}_1 + \mathbf{Z} - \hat{\rho}\mathbf{A}_1\|_2^2 \\ &= (w_1 - \hat{W})^2\frac{\|\mathbf{A}_1\|_2^2}{n} + 2(w_1 - \hat{W})\frac{\mathbf{A}_1^\intercal\mathbf{Z}}{n} + \frac{\|\mathbf{Z}\|_2^2}{n}. \qquad (2.33)\end{aligned}$$

For any $\epsilon_1 > 0$, as $m \to \infty$, by the LLN,

$$\mathsf{P}\left(\{w_1 - \hat{\rho} \in (-\epsilon_1, \epsilon_1)\} \cap \left\{ \frac{\|\mathbf{A}_1\|_2^2}{n} - \sigma_a^2 \in (-\epsilon_1, \epsilon_1) \right\} \right) \to 1.$$

Hence, we have for the first term in (2.33)

$$\mathsf{P}\left((w_1 - \hat{\rho})^2 \frac{\|\mathbf{A}_1\|_2^2}{n} \in [0, \epsilon_1^2\sigma_a^2 + \epsilon_1^3) \right) \to 1.$$

Following a similar reasoning using LLN, for the second term in (2.33),

$$\mathsf{P}\left((w_1 - \hat{\rho})\frac{\mathbf{A}_1^\intercal\mathbf{Z}}{n} \in (-\epsilon_1^2, \epsilon_1^2) \right) \to 1$$

and for the third term,

$$\mathsf{P}\left(\frac{\|\mathbf{Z}\|_2^2}{n} \in (\sigma_z^2 - \epsilon_1, \sigma_z^2 + \epsilon_1) \right) \to 1.$$

Therefore, for any $\epsilon_1 > 0$,

$$\lim_{m\to\infty} \mathsf{P}\left(\frac{1}{n}\|\mathbf{Y} - \hat{\rho}\mathbf{A}_1\|_2^2 \in (\sigma_z^2 - \epsilon_1, \sigma_z^2 + \epsilon_1)\right) = 1$$

which implies that

$$\lim_{m\to\infty} \mathsf{P}\left(\frac{1}{n}\|\mathbf{Y} - \hat{\rho}\mathbf{A}_1\|_2^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2\right) = 1.$$

Similarly, if $w_1 < 0$,

$$\lim_{m\to\infty} \mathsf{P}\left(\frac{1}{n}\|\mathbf{Y} + \hat{\rho}\mathbf{A}_1\|_2^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2\right) = 1.$$

Hence, $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_1^c) = 0$.

For the third term in (2.32), we need the following lemma, whose proof is presented at the end of this appendix:

Lemma 1. *Let $0 < \beta < \alpha$. Let $\{u_i\}_{i=1}^n$ be a real sequence satisfying*

$$\frac{1}{n}\sum_{i=1}^n u_i^2 \in (\alpha - \beta, \alpha + \beta).$$

Let $\{V_i\}_{i=1}^n$ be an i.i.d. random sequence where $V_i \sim \mathcal{N}(0, \sigma_v^2)$. Then, for any $\gamma \in (0, \alpha - \beta)$,

$$\mathsf{P}\left(\frac{1}{n}\sum_{i=1}^n (u_i - V_i)^2 \leq \gamma\right) \leq 2^{-\frac{n}{2}\log\left(\frac{\alpha-\beta}{\gamma}\right)}.$$

Continuing the proof of Theorem 1, we consider $\mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}})$ for $s \neq 1$. Then

$$\mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}}) \leq \mathsf{P}(\mathcal{E}_s | \mathcal{E}_{\text{aux}}) = \int_{\mathbf{y} \in \mathcal{E}_{\text{aux}}} \mathsf{P}(\mathcal{E}_s | \{\mathbf{Y} = \mathbf{y}\} \cap \mathcal{E}_{\text{aux}}) f(\mathbf{y} | \mathcal{E}_{\text{aux}}) d\mathbf{y}.$$

Since $\mathbf{A}_s$ is independent of $\mathbf{Y}$ and $\hat{\rho}$, it follows from the definition of $\mathcal{E}_{\text{aux}}$ and Lemma 1 (with $\alpha = w_1^2 \sigma_a^2 + \sigma_z^2$ and $\gamma = \sigma_z^2 + \epsilon^2 \sigma_a^2$) that

$$\mathsf{P}\left(\frac{1}{n} \|\mathbf{Y} - (-1)^q \hat{\rho} \mathbf{A}_s\|_2^2 \leq \sigma_z^2 + \epsilon^2 \sigma_a^2 \,\middle|\, \{\mathbf{Y} = \mathbf{y}\} \cap \mathcal{E}_{\text{aux}} \right) \leq 2^{-\frac{n}{2} \log\left(\frac{w_1^2 \sigma_a^2 + \sigma_z^2 - \epsilon}{\sigma_z^2 + \epsilon^2 \sigma_a^2} \right)}$$

for $q = 1, 2$, if ϵ is sufficiently small. Thus,

$$\mathsf{P}(\mathcal{E}_s | \{\mathbf{Y} = \mathbf{y}\} \cap \mathcal{E}_{\text{aux}}) \leq 2 \cdot 2^{-\frac{n}{2} \log\left(\frac{w_1^2 \sigma_a^2 + \sigma_z^2 - \epsilon}{\sigma_z^2 + \epsilon^2 \sigma_a^2} \right)}$$

and therefore

$$\sum_{s=2}^{m} \mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}}) \leq 2m \cdot 2^{-\frac{n}{2} \log\left(\frac{w_1^2 \sigma_a^2 + \sigma_z^2 - \epsilon}{\sigma_z^2 + \epsilon^2 \sigma_a^2} \right)}$$

which tends to zero as $m \to \infty$, if

$$\limsup_{m \to \infty} \frac{\log m}{n_m} < \frac{1}{2} \log\left(\frac{w_1^2 \sigma_a^2 + \sigma_z^2 - \epsilon}{\sigma_z^2 + \epsilon^2 \sigma_a^2} \right). \tag{2.34}$$

Therefore, by (2.32), the probability of error $\mathsf{P}(\mathcal{E})$ tends to zero as $m \to \infty$, if (2.34) is satisfied. Finally, since $\epsilon > 0$ is chosen arbitrarily, we have the desired proof of Theorem 1.

Case 2: $k \geq 2$. In this case, the signal of interest is $\mathbf{X} = \mathbf{X}(\mathbf{w}, \mathbf{S})$, where $\mathbf{w} = [w_1, ..., w_k]^\intercal$ and $\mathbf{S} = [S_1, ..., S_k]^\intercal$. Consider the following support recovery procedures. Fix $\epsilon > 0$. First, form an estimate $\hat{\rho}$ of $\|\mathbf{w}\|$ as

$$\hat{\rho} \triangleq \sqrt{\frac{\left|\frac{1}{n}\|\mathbf{Y}\|_2^2 - \sigma_z^2\right|}{\sigma_a^2}}. \tag{2.35}$$

For $r, \zeta > 0$, let $\mathcal{Q} = \mathcal{Q}(r, \zeta)$ be a minimal set of points in $\mathbb{R}^k$ satisfying the following properties:

i) $\mathcal{Q} \subseteq \mathcal{B}_k(r)$, where $\mathcal{B}_k(r)$ is the k-dimensional hypersphere of radius r, i.e., $\mathcal{B}_k(r) \triangleq \{\mathbf{b} : \mathbf{b} \in \mathbb{R}^k, \|\mathbf{b}\|_2 = r\}$,

ii) For any $\mathbf{b} \in \mathcal{B}_k(r)$, there exists $\hat{\mathbf{w}} \in \mathcal{Q}$ such that $\|\hat{\mathbf{w}} - \mathbf{b}\|_2 \leq \frac{\zeta}{2}$.

The following properties are useful:

Lemma 2. *1)*

$$\lim_{m \to \infty} \mathsf{P}\left(\exists \hat{\mathbf{W}} \in \mathcal{Q}(\hat{\rho}, \zeta) \textit{ such that } \|\hat{\mathbf{W}} - \mathbf{w}\|_2 < \zeta\right) = 1.$$

2) $q(r, \zeta) \triangleq |\mathcal{Q}(r, \zeta)|$ *is monotonically non-decreasing in* r *for fixed* ζ.

Lemma 2-1) will be proved at the end of this appendix, whereas Lemma 2-2) is obvious.

Given $\hat{\rho}$ and ϵ, fix $\mathcal{Q} = \mathcal{Q}(\hat{\rho}, \epsilon)$. Declare $d(\mathbf{Y}) = \{\hat{s}_1, \hat{s}_2, ..., \hat{s}_k\} \subseteq [m]$ is the recovered support of the signal, if it is the unique set of indices such that

$$\frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{\hat{s}_j}\right\|_2^2 \le \sigma_z^2 + \epsilon^2\sigma_a^2 \tag{2.36}$$

for some $\hat{\mathbf{W}} \in \mathcal{Q}$. If there is none or more than one such set, pick an arbitrary set of k indices.

Next, we analyze the average probability of error

$$\mathsf{P}(\mathcal{E}) = \mathsf{P}\{d^{(m)}(\mathbf{Y}) \neq \{S_1, ..., S_k\}\}$$

where the expectation is taken with respect to A. As before, we assume without loss of generality that $S_j = j$ for $j = 1, 2, ..., k$, which gives

$$\mathbf{Y} = \sum_{j=1}^{k} w_j \mathbf{A}_j + \mathbf{Z}$$

for some $\mathbf{w}$. Define the event

$$\begin{aligned}\mathcal{E}_{s_1,s_2,...,s_k} \triangleq \Bigg\{ &\exists \hat{\mathbf{W}} \in \mathcal{Q} \text{ and } \{s_1', s_2', ..., s_k'\} = \{s_1, s_2, ..., s_k\} \\ &\text{such that } \frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{s_j'}\right\|_2^2 \le \sigma_z^2 + \epsilon^2\sigma_a^2 \Bigg\}.\end{aligned}$$

Then

$$
\begin{aligned}
\mathsf{P}(\mathcal{E}) &= \mathsf{P}\left(\mathcal{E}^c_{1,2,\dots,k} \cup \left(\bigcup_{s_1<\dots<s_k:\{s_1,\dots,s_k\}\neq[k]} \mathcal{E}_{s_1,s_2,\dots,s_k}\right)\right) \\
&\leq \mathsf{P}\left(\mathcal{E}^c_{\text{aux}} \cup \mathcal{E}^c_{1,2,\dots,k} \cup \left(\bigcup_{s_1<\dots<s_k:\{s_1,\dots,s_k\}\neq[k]} (\mathcal{E}_{s_1,s_2,\dots,s_k} \cap \mathcal{E}_{\text{aux}})\right)\right) \\
&\leq \mathsf{P}(\mathcal{E}^c_{\text{aux}}) + \mathsf{P}(\mathcal{E}^c_{1,2,\dots,k}) + \sum_{s_1<\dots<s_k:\{s_1,\dots,s_k\}\neq[k]} \mathsf{P}(\mathcal{E}_{s_1,s_2,\dots,s_k} \cap \mathcal{E}_{\text{aux}}) \qquad (2.37)
\end{aligned}
$$

where in this case

$$
\begin{aligned}
\mathcal{E}_{\text{aux}} \triangleq \{\hat{\rho} - \|\mathbf{w}\|_2 \in (-\epsilon, \epsilon)\} \\
&\cap \left(\bigcap_{j=1}^{k} \left\{\frac{1}{n}\|\mathbf{A}_j\|_2^2 - \sigma_a^2 \in (-\epsilon, \epsilon)\right\}\right) \\
&\cap \left(\bigcap_{j=1}^{k} \bigcap_{l=j+1}^{k} \left\{\frac{1}{n}\mathbf{A}_j^{\intercal}\mathbf{A}_l \in (-\epsilon, \epsilon)\right\}\right) \\
&\cap \left(\bigcap_{j=1}^{k} \left\{\frac{1}{n}\mathbf{A}_j^{\intercal}\mathbf{Z} \in (-\epsilon, \epsilon)\right\}\right) \\
&\cap \left\{\frac{1}{n}\|\mathbf{Z}\|_2^2 - \sigma_z^2 \in (-\epsilon, \epsilon)\right\}.
\end{aligned}
$$

We now bound the terms in (2.37). First, by the LLN, $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}^c_{\text{aux}}) = 0$.

Next, we consider $\mathsf{P}(\mathcal{E}^c_{1,2,...,k})$. Note that, for any $\hat{\mathbf{W}} \in \mathcal{Q}$,

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_j\right\|_2^2 &= \frac{1}{n}\left\|\sum_{j=1}^{k} w_j \mathbf{A}_j + \mathbf{Z} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_j\right\|_2^2 \\ &= \frac{1}{n}\sum_{j=1}^{k}\sum_{l=1}^{k}(w_j - \hat{W}_j)(w_l - \hat{W}_l)\mathbf{A}_j^{\intercal}\mathbf{A}_l \\ &\quad + \frac{2}{n}\sum_{j=1}^{k}(w_j - \hat{W}_j)\mathbf{A}_j^{\intercal}\mathbf{Z} + \frac{1}{n}\|\mathbf{Z}\|_2^2. \end{aligned} \tag{2.38}$$

By applying the LLN to each term in (2.38), as similarly done in case 1, and using Lemma 2-1), we have

$$\lim_{m\to\infty} \mathsf{P}\left(\exists \hat{\mathbf{W}} \in \mathcal{Q} \text{ s.t. } \frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_j\right\|_2^2 \le \sigma_z^2 + \epsilon^2\sigma_a^2\right) = 1$$

which implies that $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}^c_{1,2,...,k}) = 0$.

Next, we consider $\mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} \cap \mathcal{E}_{\text{aux}})$ for $\{s_1, s_2, ..., s_k\} \neq [k]$. Note that

$$\begin{aligned}&\mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} \cap \mathcal{E}_{\text{aux}}) \\ &\le \mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} | \mathcal{E}_{\text{aux}}) \\ &= \int \cdots \int_{\{\mathbf{a}_1,...,\mathbf{a}_k,\mathbf{z}\}\in\mathcal{E}_{\text{aux}}} \mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} | \{\mathbf{A}_1 = \mathbf{a}_1\} \cap \\ &\quad \cdots \cap \{\mathbf{A}_k = \mathbf{a}_k\} \cap \{\mathbf{Z} = \mathbf{z}\} \cap \mathcal{E}_{\text{aux}}) \times f(\mathbf{a}_1, ..., \mathbf{a}_k, \mathbf{z} | \mathcal{E}_{\text{aux}}) d\mathbf{a}_1 \cdots d\mathbf{a}_k d\mathbf{z}. \end{aligned} \tag{2.39}$$

For notational simplicity, define $\xi \triangleq \sigma_z^2 + \epsilon^2\sigma_a^2$, $\mathcal{T} \triangleq \{s_1, s_2, ..., s_k\} \cap [k]$, $\mathcal{T}^c \triangleq \{s_1, s_2, ..., s_k\} \backslash \mathcal{T}$, and $\mathcal{E}_{\text{cond}} \triangleq \{\mathbf{A}_1 = \mathbf{a}_1\} \cap \cdots \cap \{\mathbf{A}_k = \mathbf{a}_k\} \cap \{\mathbf{Z} = \mathbf{z}\} \cap \mathcal{E}_{\text{aux}}$.

For any permutation $(s'_1, s'_2, ..., s'_k)$ of $\{s_1, s_2, ..., s_k\}$ and any $\hat{\mathbf{W}} \in \mathcal{Q}$,

$$\begin{aligned}
&\mathsf{P}\left(\frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{s'_j}\right\|_2^2 \le \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&= \mathsf{P}\left(\frac{1}{n}\left\|\sum_{j=1}^{k} w_j \mathbf{A}_j + \mathbf{Z} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{s'_j}\right\|_2^2 \le \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&= \mathsf{P}\left(\frac{1}{n}\left\|\left[\sum_{j=1}^{k} w_j \mathbf{A}_j - \sum_{s'_j \in \mathcal{T}} \hat{W}_j \mathbf{A}_{s'_j} + \mathbf{Z}\right] - \sum_{s'_j \in \mathcal{T}^c} \hat{W}_j \mathbf{A}_{s'_j}\right\|_2^2 \le \xi \middle| \mathcal{E}_{\text{cond}}\right). \qquad (2.40)
\end{aligned}$$

Conditioned on $\mathcal{E}_{\text{cond}}$ and the chosen $\mathcal{Q}$, $\frac{1}{n}\left\| \sum_{j=1}^{k} w_j \mathbf{A}_j - \sum_{s'_j \in \mathcal{T}} \hat{W}_j \mathbf{A}_{s'_j} + \mathbf{Z}\right\|_2^2$ is a fixed quantity satisfying

$$\begin{aligned}
\frac{1}{n}\left\|\sum_{j=1}^{k} w_j \mathbf{A}_j - \sum_{s'_j \in \mathcal{T}} \hat{W}_j \mathbf{A}_{s'_j} + \mathbf{Z}\right\|_2^2 \in &\left(\left[\sum_{j \in [k]\setminus\mathcal{T}} w_j^2 + \sum_{s'_j \in \mathcal{T}} (w_{s'_j} - \hat{W}_j)^2\right]\sigma_a^2 + \sigma_z^2 - \delta_1 \epsilon,\right. \\
&\left.\left[\sum_{j \in [k]\setminus\mathcal{T}} w_j^2 + \sum_{s'_j \in \mathcal{T}} (w_{s'_j} - \hat{W}_j)^2\right]\sigma_a^2 + \sigma_z^2 + \delta_1 \epsilon\right)
\end{aligned}$$

for some positive δ_1 that depends on $\mathbf{w}$ and ϵ only, and is non-decreasing in ϵ. Meanwhile, $\mathbf{A}_{s'_j}$ is independent of $\mathbf{A}_1, ..., \mathbf{A}_k$, and $\mathbf{Z}$ for $s'_j \in \mathcal{T}^c$. Hence, by Lemma 1 (with $\alpha = \left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2 + \sum_{s'_j\in\mathcal{T}} (w_{s'_j} - \hat{W}_j)^2\right)\sigma_a^2 + \sigma_z^2$ and $\gamma = \sigma_z^2 + \epsilon^2\sigma_a^2$), (2.40) is upper-bounded by

$$2^{-\frac{n}{2}\log\frac{\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2 + \sum_{s'_j\in\mathcal{T}}\left(w_{s'_j} - \hat{W}_j\right)^2\right)\sigma_a^2 + \sigma_z^2 - \delta_1\epsilon}{\sigma_z^2 + \epsilon^2\sigma_a^2}} \le 2^{-\frac{n}{2}\log\frac{\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2\right)\sigma_a^2 + \sigma_z^2 - \delta_1\epsilon}{\sigma_z^2 + \epsilon^2\sigma_a^2}}.$$

Hence, by the union of events bound,

$$
\begin{aligned}
&\mathsf{P}(\mathcal{E}_{s_1,s_2,\ldots,s_k}|\mathcal{E}_{\text{cond}}) \\
&\leq \sum_{\{s'_1,\ldots,s'_k\}=\{s_1,\ldots,s_k\}} \mathsf{P}\left(\exists \hat{\mathbf{W}} \in \mathcal{Q} \text{ s.t. } \frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{s'_j}\right\|_2^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&\leq \sum_{\{s'_1,\ldots,s'_k\}=\{s_1,\ldots,s_k\}} \sum_{\hat{\mathbf{W}} \in \mathcal{Q}} \mathsf{P}\left(\frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{s'_j}\right\|_2^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&\leq k! \cdot |\mathcal{Q}| \cdot 2^{-\frac{n}{2}\log \frac{\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2\right)\sigma_a^2+\sigma_z^2-\delta_1\epsilon}{\sigma_z^2+\epsilon^2\sigma_a^2}}.
\end{aligned}
$$

Furthermore, conditioned on $\mathcal{E}_{\text{aux}}$, $\hat{\rho} < \|\mathbf{w}\|_2 + \epsilon$ and hence $|\mathcal{Q}| \leq q(\|\mathbf{w}\|_2 + \epsilon, \epsilon)$ by Lemma 2-2). Thus,

$$
\mathsf{P}(\mathcal{E}_{s_1,s_2,\ldots,s_k} \cap \mathcal{E}_{\text{aux}}) \leq k! \cdot q(\|\mathbf{w}\|_2 + \epsilon, \epsilon) \cdot 2^{-\frac{n}{2}\log \frac{\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2\right)\sigma_a^2+\sigma_z^2-\delta_1\epsilon}{\sigma_z^2+\epsilon^2\sigma_a^2}}. \tag{2.41}
$$

Note that the probability upper-bound (2.41) depends on $s_1, \ldots, s_k$ only through $\mathcal{T}$.

Grouping the $\binom{m-k}{k-|\mathcal{T}|}$ events $\{\mathcal{E}_{s_1,s_2,\ldots,s_k} \cap \mathcal{E}_{\text{aux}}\}$ with the same $\mathcal{T}$,

$$
\begin{aligned}
\mathsf{P}(\mathcal{E}) &\leq \mathsf{P}(\mathcal{E}_{\text{aux}}^c) + \mathsf{P}(\mathcal{E}_{1,2,\ldots,k}^c) + \sum_{\mathcal{T}\subset[k]} \binom{m-k}{k-|\mathcal{T}|} \\
&\qquad \cdot k! \cdot q(\|\mathbf{w}\|_2+\epsilon,\epsilon) \cdot 2^{-\frac{n}{2}\log\frac{\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2\right)\sigma_a^2+\sigma_z^2-\delta_1\epsilon}{\sigma_z^2+\epsilon^2\sigma_a^2}} \\
&\leq \mathsf{P}(\mathcal{E}_{\text{aux}}^c) + \mathsf{P}(\mathcal{E}_{1,2,\ldots,k}^c) + k!\cdot q(\|\mathbf{w}\|_2+\epsilon,\epsilon) \\
&\qquad \cdot \sum_{\mathcal{T}\subset[k]} 2^{(k-|\mathcal{T}|)\log m} \cdot 2^{-\frac{n}{2}\log\frac{\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2\right)\sigma_a^2+\sigma_z^2-\delta_1\epsilon}{\sigma_z^2+\epsilon^2\sigma_a^2}} \\
&= \mathsf{P}(\mathcal{E}_{\text{aux}}^c) + \mathsf{P}(\mathcal{E}_{1,2,\ldots,k}^c) + k!\cdot q(\|\mathbf{w}\|_2+\epsilon,\epsilon) \\
&\qquad \cdot \sum_{\mathcal{T}\subseteq[k]} 2^{|\mathcal{T}|\log m} \cdot 2^{-\frac{n}{2}\log\frac{\left(\sum_{j\in\mathcal{T}} w_j^2\right)\sigma_a^2+\sigma_z^2-\delta_1\epsilon}{\sigma_z^2+\epsilon^2\sigma_a^2}}
\end{aligned}
$$

which tends to zero as $m\to\infty$, if

$$
\limsup_{m\to\infty}\frac{\log m}{n_m} < \frac{1}{2|\mathcal{T}|}\log\frac{\left(\sum_{j\in\mathcal{T}} w_j^2\right)\sigma_a^2+\sigma_z^2-\delta_1\epsilon}{\sigma_z^2+\epsilon^2\sigma_a^2} \tag{2.42}
$$

for all $\mathcal{T}\subseteq[k]$. Since $\epsilon>0$ is arbitrarily chosen, the proof of Theorem 1 is complete.

Now, it only remains to prove Lemma 1. For simplicity, let $\theta\equiv\sigma_v^2$. Denote $S_n=\frac{1}{n}\sum_{i=1}^n(u_i-V_i)^2$. The moment generating function of S_n is

$$
\mathsf{E}[e^{tS_n}] = \mathsf{E}[e^{\frac{t}{n}\sum_{i=1}^n(u_i-V_i)^2}] = \prod_{i=1}^n \mathsf{E}[e^{\frac{t}{n}(u_i-V_i)^2}]. \tag{2.43}
$$

Note that $(u_i-V_i)^2/\theta$ is a noncentral χ^2 random variable. Its moment generating function is given by [74] as $\mathsf{E}[e^{t(u_i-V_i)^2/\theta}] = \exp(\frac{tu_i^2/\theta}{1-2t})/(1-2t)^{\frac{1}{2}}$, for $t \leq 1/2$. By changing variable $\theta t/n \to t$, we have

$$\mathsf{E}[e^{t(u_i-V_i)^2/n}] = \frac{e^{\frac{\frac{t}{n}u_i^2}{1-2\theta t/n}}}{(1-2\theta t/n)^{\frac{1}{2}}}.$$

Back to (2.43), we obtain

$$\begin{aligned}
\mathsf{E}[e^{tS_n}] &= \prod_{i=1}^{n} \mathsf{E}[e^{\frac{t}{n}(u_i-V_i)^2}] \\
&= \prod_{i=1}^{n} \frac{e^{\frac{\frac{t}{n}u_i^2}{1-2\theta t/n}}}{(1-2\theta t/n)^{\frac{1}{2}}} \\
&= \frac{e^{\frac{\frac{t}{n}\sum_{i=1}^{n}u_i^2}{1-2\theta t/n}}}{(1-2\theta t/n)^{\frac{n}{2}}}.
\end{aligned}$$

The Chernoff bound implies

$$\begin{aligned}
\mathsf{P}(S_n \leq \gamma) &\leq \min_{s>0} e^{s\gamma}\, \mathsf{E}[e^{-sS_n}] \\
&= \min_{s>0} e^{s\gamma} \frac{e^{\frac{-\frac{s}{n}\sum_{i=1}^{n}u_i^2}{1+2\theta s/n}}}{(1+2\theta s/n)^{\frac{n}{2}}} \\
&= \min_{p<0} e^{-p\gamma} \frac{e^{\frac{\frac{p}{n}\sum_{i=1}^{n}u_i^2}{1-2\theta p/n}}}{(1-2\theta p/n)^{\frac{n}{2}}} \\
&= \exp\left\{\min_{p<0}\left\{\log e^{-p\gamma}\frac{e^{\frac{\frac{p}{n}\sum_{i=1}^{n}u_i^2}{1-2\theta p/n}}}{(1-2\theta p/n)^{\frac{n}{2}}}\right\}\right\} \\
&= \exp\left\{\min_{p<0}\left\{-p\gamma + \frac{\frac{p}{n}\sum_{i=1}^{n}u_i^2}{1-2\theta p/n} - \frac{n}{2}\log\left(1-2\theta p/n\right)\right\}\right\}.
\end{aligned}$$

Define

$$f(p) \triangleq -p\gamma + \frac{\frac{p}{n}\sum_{i=1}^{n} u_i^2}{1-2\theta p/n} - \frac{n}{2}\log\left(1-2\theta p/n\right)$$
$$g(\lambda) \triangleq f(n\lambda) = -n\lambda\gamma + \frac{\lambda\sum_{i=1}^{n} u_i^2}{1-2\theta\lambda} - \frac{n}{2}\log\left(1-2\theta\lambda\right).$$

Clearly, $\min_{p<0} f(p) = \min_{\lambda<0} g(\lambda)$. Denote

$$\alpha_s \triangleq \frac{1}{n}\sum_{i=1}^{n} u_i^2.$$

Then, let us focus on the minimization problem

$$\begin{aligned}
\min_{\lambda<0} g(\lambda) &= \min_{\lambda<0}\left\{-n\lambda\gamma + \frac{n\lambda\alpha_s}{1-2\theta\lambda} - \frac{n}{2}\log\left(1-2\theta\lambda\right)\right\} \\
&= n\cdot\min_{\lambda<0}\left\{-\lambda\gamma + \frac{\lambda\alpha_s}{1-2\theta\lambda} - \frac{1}{2}\log\left(1-2\theta\lambda\right)\right\} \\
&= -n\cdot\underbrace{\max_{\lambda<0}\left\{\lambda\gamma - \frac{\lambda\alpha_s}{1-2\theta\lambda} + \frac{1}{2}\log\left(1-2\theta\lambda\right)\right\}}_{\triangleq\,\Lambda(\alpha_s,\theta,\gamma)}.
\end{aligned}$$

It can be shown that the minimizing λ is

$$\lambda^* = \frac{2\gamma-\theta-\sqrt{\theta^2+4\alpha_s\gamma}}{4\theta\gamma} < 0$$

and hence

$$\Lambda(\alpha_s,\theta,\gamma) = \lambda^*\gamma - \frac{\lambda^*\alpha_s}{1-2\lambda^*\theta} + \frac{1}{2}\log(1-2\lambda^*\theta)$$
$$= \frac{\alpha_s+\gamma}{2\theta} - \frac{1}{2} - \frac{2\alpha_s\gamma}{\theta(\theta+\sqrt{\theta^2+4\alpha_s\gamma})} + \frac{1}{2}\log\frac{\theta+\sqrt{\theta^2+4\alpha_s\gamma}}{2\gamma}.$$

Next, for fixed α_s and γ,

$$\frac{\partial\Lambda(\alpha_s,\theta,\gamma)}{\partial\theta} = -\frac{\alpha_s+\gamma}{2\theta^2}$$
$$+ \frac{2\alpha_s\gamma\left[\theta+\sqrt{\theta^2+4\alpha_s\gamma}+\theta\left(1+\frac{2\theta}{2\sqrt{\theta^2+4\alpha_s\gamma}}\right)\right]}{\theta^2(\theta+\sqrt{\theta^2+4\alpha_s\gamma})^2}$$
$$+ \frac{1}{2(\theta+\sqrt{\theta^2+4\alpha_s\gamma})}\left(1+\frac{2\theta}{2\sqrt{\theta^2+4\alpha_s\gamma}}\right)$$
$$= -\frac{\alpha_s+\gamma}{2\theta^2} + \frac{\sqrt{4\alpha_s\gamma+\theta^2}}{2\theta^2}.$$

For $\theta > 0$, there is only one stationary point $\theta' = \alpha_s - \gamma$, which is a solution to $\frac{\partial\Lambda(\alpha_s,\theta,\gamma)}{\partial\theta} = 0$. Check the second derivative,

$$\left.\frac{\partial^2\Lambda(\alpha_s,\theta,\gamma)}{\partial\theta^2}\right|_{\theta=\alpha_s-\gamma} = \frac{1}{2(\alpha_s+\gamma)(\alpha_s-\gamma)} > 0.$$

This confirms that $\theta' = \alpha_s - \gamma$ is the minimum point of $\Lambda(\alpha_s,\theta,\gamma)$, for $\theta > 0$. Hence, for fixed α_s and γ with $\gamma < \alpha_s$,

$$\Lambda(\alpha_s,\theta,\gamma) \geq \Lambda(\alpha_s,\theta',\gamma) = \frac{1}{2}\log\frac{\alpha_s}{\gamma}.$$

As a result,

$$\begin{aligned}
\mathsf{P}(S_n \le \gamma) &\le \exp\left\{\min_{p<0}\left\{-p\gamma + \frac{\frac{p}{n}\sum_{i=1}^n u_i^2}{1-2\theta p/n} - \frac{n}{2}(1-2\theta p/n)\right\}\right\} \\
&= \exp\left\{\min_{\lambda<0} g(\lambda)\right\} \\
&= \exp\left\{-n\Lambda(\alpha_s, \theta, \gamma)\right\} \\
&\le \exp\left\{-n\Lambda(\alpha_s, \theta', \gamma)\right\} \\
&= \exp\left\{-\frac{n}{2}\log\left(\frac{\alpha_s}{\gamma}\right)\right\} \\
&\le \exp\left\{-\frac{n}{2}\log\left(\frac{\alpha-\beta}{\gamma}\right)\right\}.
\end{aligned}$$

Hence, by changing the base of logarithm,

$$\mathsf{P}\left(\frac{1}{n}\sum_{i=1}^n (u_i - V_i)^2 \le \gamma\right) \le 2^{-\frac{n}{2}\log\left(\frac{\alpha-\beta}{\gamma}\right)}.$$

Finally, we verify Lemma 2-1). For any $\zeta_1 > 0$, according to LLN,

$$\lim_{m\to\infty} \mathsf{P}(|\hat{\rho} - \|\mathbf{w}\|_2| \le \zeta_1) = 1.$$

Note that $\mathbf{W}_0 \triangleq \frac{\hat{\rho}}{\|\mathbf{w}\|_2}\mathbf{w} \in \mathcal{B}_k(\hat{W})$. According to the definition of $\mathcal{Q}(\hat{\rho}, \zeta)$, there must exist $\hat{\mathbf{W}} \in \mathcal{Q}(\hat{\rho}, \zeta)$ such that $\|\hat{\mathbf{W}} - \mathbf{W}_0\|_2 \leq \frac{\zeta}{2}$. Fundamental geometry implies

$$\begin{aligned} \|\hat{\mathbf{W}} - \mathbf{w}\|_2 &\leq \|\hat{\mathbf{W}} - \mathbf{W}_0\|_2 + \|\mathbf{W}_0 - \mathbf{w}\|_2 \\ &\leq \frac{\zeta}{2} + \big|\hat{\rho} - \|\mathbf{w}\|_2\big|. \end{aligned}$$

Hence,

$$\lim_{m\to\infty} \mathsf{P}\left(\|\hat{\mathbf{W}} - \mathbf{w}\|_2 \leq \frac{\zeta}{2} + \zeta_1\right) = 1.$$

Choosing $\zeta_1 \in (0, \zeta/2)$ completes the proof.

2.8.2 Proof of Theorem 2

The main techniques for the proof of Theorem 2 include Fano's inequality and the properties of entropy. It mimics the proof of the converse for the channel coding theorem [32] with proper modification. To establish this theorem, we prove the following equivalent statement:

If there exists a sequence of matrices $\{A^{(m)}\}_{m=k}^{\infty}$, $A^{(m)} \in \mathbb{R}^{n_m \times m}$, and a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{\{1,2,\ldots,m\}}$, such that

$$\frac{1}{n_m m}\|A^{(m)}\|_F^2 \leq \sigma_a^2 \tag{2.44}$$

and

$$\lim_{m\to\infty} \mathsf{P}\{d(A^{(m)}\mathbf{X} + \mathbf{Z}) \neq \mathrm{supp}(\mathbf{X}(\mathbf{w}, \mathbf{S}))\} = 0$$

then

$$\limsup_{m\to\infty} \frac{\log m}{n_m} \leq c(\mathbf{w}). \tag{2.45}$$

In the next, we justify this alternative claim. For any $\mathcal{T} \subseteq [k]$, denote the tuple of random variables $(S_l : l \in \mathcal{T})$ by $S(\mathcal{T})$. From Fano's inequality [32], we have

$$\begin{aligned} H(S(\mathcal{T})|\mathbf{Y}) &\leq H(S_1, ..., S_k|\mathbf{Y}) \\ &\leq \log k! + H(\{S_1, ..., S_k\}|\mathbf{Y}) \\ &\leq \log k! + \overline{P}_e(\mathbf{w}, A^{(m)}) \log \binom{m}{k} + 1. \end{aligned} \tag{2.46}$$

For notation simplicity, let $\overline{P}_e^{(m)} \triangleq \mathsf{P}\{d(A^{(m)}\mathbf{X} + \mathbf{Z}) \neq \text{supp}(\mathbf{X}(\mathbf{w}, \mathbf{S}))$. On the other hand, by a basic permutation argument,

$$\begin{aligned} H(S(\mathcal{T})|S(\mathcal{T}^c)) &= \log \left(\prod_{q=0}^{|\mathcal{T}|-1} (m - (k - |\mathcal{T}|) - q) \right) \\ &= |\mathcal{T}| \log m - n\epsilon_{1,n} \end{aligned} \tag{2.47}$$

where $\mathcal{T}^c \triangleq [k] \backslash \mathcal{T}$ and

$$\epsilon_{1,n} \triangleq \frac{1}{n} \log \left(m^{|\mathcal{T}|} / \prod_{q=0}^{|\mathcal{T}|-1} (m - (k - |\mathcal{T}|) - q) \right)$$

which tends to zero as $n \to \infty$. Hence, combining (2.46) and (2.47), we have

$$
\begin{aligned}
|\mathcal{T}| \log m &= H(S(\mathcal{T})|S(\mathcal{T}^c)) + n\epsilon_{1,n} \\
&= I(S(\mathcal{T}); \mathbf{Y}|S(\mathcal{T}^c)) + H(S(\mathcal{T})|\mathbf{Y}, S(\mathcal{T}^c)) + n\epsilon_{1,n} \\
&\leq I(S(\mathcal{T}); \mathbf{Y}|S(\mathcal{T}^c)) + H(S(\mathcal{T})|\mathbf{Y}) + n\epsilon_{1,n} \qquad (2.48)\\
&\leq I(S(\mathcal{T}); \mathbf{Y}|S(\mathcal{T}^c)) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&= \sum_{i=1}^{n} I(Y_i; S(\mathcal{T})|Y_1^{i-1}, S(\mathcal{T}^c)) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} \\
&\quad + 1 + n\epsilon_{1,n} \qquad (2.49)\\
&\leq \sum_{i=1}^{n} \left(h(Y_i|S(\mathcal{T}^c)) - h(Y_i|S_1, ..., S_k)\right) + \log k! \\
&\quad + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&= \sum_{i=1}^{n} (h(Y_i|S(\mathcal{T}^c)) - h(Z_i)) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} \\
&\quad + 1 + n\epsilon_{1,n} \qquad (2.50)
\end{aligned}
$$

where (2.48) follows the fact that conditioning reduces entropy, (2.49) follows the chain rule of mutual information [32], and (2.50) follows since the measurement matrix is fixed and Z_i is independent of $(S_1, \ldots, S_k)$.

Consider

$$
\begin{aligned}
h(Y_i|S(\mathcal{T}^c)) &= h\left(\sum_{j=1}^{k} w_j a_{i,S_j} + Z_i \middle| S(\mathcal{T}^c)\right) \\
&= h\left(\sum_{j\in\mathcal{T}} w_j a_{i,S_j} + Z_i \middle| S(\mathcal{T}^c)\right) \\
&\le h\left(\sum_{j\in\mathcal{T}} w_j a_{i,S_j} + Z_i\right) \\
&\le \frac{1}{2}\log\left(2\pi e \cdot \mathrm{Var}\left(\sum_{j\in\mathcal{T}} w_j a_{i,S_j} + Z_i\right)\right) \qquad (2.51)
\end{aligned}
$$

where the last inequality follows since the Gaussian random variable maximizes the differential entropy given a variance constraint. To further upper-bound (2.51), note that

$$
\mathrm{Var}\left(\sum_{j\in\mathcal{T}} w_j a_{i,S_j} + Z_i\right) = \mathsf{E}\left[\left(\sum_{j\in\mathcal{T}} w_j a_{i,S_j}\right)^2\right] - \mathsf{E}\left[\sum_{j\in\mathcal{T}} w_j a_{i,S_j}\right]^2 + \sigma_z^2.
$$

Now

$$
\mathsf{E}\left[\sum_{j\in\mathcal{T}} w_j a_{i,S_j}\right] = \sum_{j\in\mathcal{T}} w_j\,\mathsf{E}[a_{i,S_j}] = \sum_{j\in\mathcal{T}} w_j \cdot \frac{1}{m}\sum_{p=1}^{m} a_{i,p}
$$

and

$$
\begin{aligned}
\mathsf{E}\left[\left(\sum_{j\in\mathcal{T}} w_j a_{i,S_j}\right)^2\right] &= \sum_{j\in\mathcal{T}}\sum_{l\in\mathcal{T}} w_j w_l\,\mathsf{E}[a_{i,S_j}a_{i,S_l}] \\
&= \sum_{j\in\mathcal{T}}\sum_{l\in\mathcal{T},l\neq j} w_j w_l\,\mathsf{E}[a_{i,S_j}a_{i,S_l}] + \sum_{j\in\mathcal{T}} w_j^2\,\mathsf{E}[a_{i,S_j}^2] \\
&= \sum_{j\in\mathcal{T}}\sum_{l\in\mathcal{T},l\neq j} \frac{w_j w_l}{m(m-1)} \sum_{p=1}^{m}\sum_{\substack{q=1\\ q\neq p}}^{m} a_{i,p}a_{i,q} \\
&\quad + \sum_{j\in\mathcal{T}} w_j^2 \cdot \frac{1}{m}\sum_{p=1}^{m} a_{i,p}^2 \\
&= \sum_{j\in\mathcal{T}}\sum_{l\in\mathcal{T},l\neq j} \frac{w_j w_l}{m(m-1)} \left(\sum_{p=1}^{m} a_{i,p}\right)^2 \\
&\quad + \frac{1}{m}\left(\sum_{j\in\mathcal{T}} w_j^2 - \tau(m)\right)\sum_{p=1}^{m} a_{i,p}^2
\end{aligned}
$$

where $\tau(m) \triangleq \sum_{j\in\mathcal{T}}\sum_{l\in\mathcal{T},l\neq j}\frac{w_j w_l}{(m-1)} \to 0$ as $m \to \infty$. It can be also easily checked that

$$
\sum_{j\in\mathcal{T}}\sum_{l\in\mathcal{T},l\neq j} \frac{w_j w_l}{m(m-1)} \leq \left(\sum_{j\in\mathcal{T}} \frac{w_j}{m}\right)^2
$$

and thus

$$
\mathrm{Var}\left(\sum_{j\in\mathcal{T}} w_j a_{i,S_j} + Z_i\right) \leq \left(\sum_{j\in\mathcal{T}} w_j^2 - \tau(m)\right)\frac{1}{m}\sum_{p=1}^{m} a_{i,p}^2 + \sigma_z^2.
$$

Returning to (2.102), we have

$$
\begin{aligned}
|\mathcal{T}|\log m &\le \sum_{i=1}^{n}\frac{1}{2}\log\left[2\pi e\left(\left(\sum_{j\in\mathcal{T}}w_j^2-\tau(m)\right)\frac{1}{m}\sum_{p=1}^{m}a_{i,p}^2+\sigma_z^2\right)\right]\\
&\quad-\frac{n}{2}\log(2\pi e\sigma_z^2)+\log k!+\overline{P}_e^{(m)}\log\binom{m}{k}+1+n\epsilon_{1,n}\\
&\le\frac{n}{2}\log\left[2\pi e\left(\left(\sum_{j\in\mathcal{T}}w_j^2-\tau(m)\right)\frac{1}{nm}\sum_{i=1}^{n}\sum_{p=1}^{m}a_{i,p}^2+\sigma_z^2\right)\right]\\
&\quad-\frac{n}{2}\log(2\pi e\sigma_z^2)+\log k!+\overline{P}_e^{(m)}\log\binom{m}{k}+1+n\epsilon_{1,n} \qquad (2.52)\\
&=\frac{n}{2}\log\left(\left(\sum_{j\in\mathcal{T}}w_j^2-\tau(m)\right)\frac{\sigma_a^2}{\sigma_z^2}+1\right)+\log k!\\
&\quad+\overline{P}_e^{(m)}\log\binom{m}{k}+1+n\epsilon_{1,n} \qquad (2.53)
\end{aligned}
$$

where (2.52) is due to Jensen's inequality and (2.53) follows from (2.44). Therefore,

$$
\begin{aligned}
&\limsup_{m\to\infty}\frac{\log m}{n_m}-\frac{\log k!+\overline{P}_e^{(m)}\log\binom{m}{k}+1+n_m\epsilon_{1,n_m}}{|\mathcal{T}|n_m}\\
&\le\frac{1}{2|\mathcal{T}|}\log\left(1+\frac{\sigma_a^2}{\sigma_z^2}\sum_{j\in\mathcal{T}}w_j^2\right) \qquad (2.54)
\end{aligned}
$$

for all $\mathcal{T}\subseteq[k]$. Due to the fact that $\log\binom{m}{k}\le k\log m$, we have

$$
\begin{aligned}
&\limsup_{m\to\infty}\frac{(1-k\overline{P}_e^{(m)}/|\mathcal{T}|)\log m}{n_m}-\frac{\log k!+n_m\epsilon_{1,n_m}+1}{|\mathcal{T}|n_m}\\
&\le\frac{1}{2|\mathcal{T}|}\log\left(1+\frac{\sigma_a^2}{\sigma_z^2}\sum_{j\in\mathcal{T}}w_j^2\right) \qquad (2.55)
\end{aligned}
$$

for all $\mathcal{T} \subseteq [k]$. Since $\lim_{m\to\infty} \overline{P}_e^{(m)} = 0$, we reach the conclusion

$$\limsup_{m\to\infty} \frac{\log m}{n_m} \leq \frac{1}{2|\mathcal{T}|} \log\left(1 + \frac{\sigma_a^2}{\sigma_z^2} \sum_{j\in\mathcal{T}} w_j^2\right)$$

for all $\mathcal{T} \subseteq [k]$, which completes the proof of Theorem 2.

2.8.3 Proof of Theorem 3

We show that

$$\lim_{m\to\infty} \mathsf{P}\{d^{(m)}(A\mathbf{X}(\mathbf{w}^{(m)}, \mathbf{S}) + \mathbf{Z}) \neq \text{supp}(\mathbf{X}(\mathbf{w}^{(m)}, \mathbf{S}))\} = 0$$

provided that the condition

$$\limsup_{m\to\infty} \frac{1}{n_m} \max_{j\in[k_m]} \left[\frac{6k_m \log k_m + 2j \log \frac{me}{j}}{\log\left(\frac{j w_{\text{min}}^2 \sigma_a^2}{\sigma_z^2} + 1\right)} \right] < 1 \tag{2.56}$$

is satisfied. Note that (2.56) implies that $n = \max[\Omega(k \log k), \Omega(\frac{k}{\log k} \log \frac{m}{k})]$, which in turn implies that $k = o(n)$.

We follow the proof of Theorem 1 in Appendix 2.8.1. Recall that in case 2 of the proof of Theorem 1, we first proposed the support recovery rule (2.36). Then, we formed estimates of the nonzero values, and used them to test all possible sets of k indices. The key step was to analyze two types of errors. On the one hand, the true support should satisfy the reconstruction rule (2.36) with high probability. On the other hand, the probability that at least one incorrect support possibility satisfies this rule was controlled to diminish as the problem size increases.

By mainly replicating the steps in Appendix 2.8.1 with necessary accommodations to the new setting with growing number of nonzero entries, we present the proof of Theorem 3 as follows.

1. We first modify the support recovery rule by replacing (2.36) with

$$\frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{\hat{s}_j}\right\|_2^2 \leq (1+\epsilon)\sigma_z^2 + 2\epsilon^2\sigma_a^2. \tag{2.57}$$

2. The cardinality $q(r,\zeta)$ of a minimal $\mathcal{Q}(r,\zeta)$ can be upper-bounded by

$$q(r,\zeta) \leq \left(\frac{\eta_1 k r}{\zeta}\right)^k$$

for some $\eta_1 > 0$. This can be easily shown by first partitioning the k-dimensional hypercube of side $2r$ into identical elementary hypercubes with side not exceeding $\frac{\zeta}{4k}$ and then, for each elementary hypercube that intersects the hypersphere, picking an arbitrary point on the hypersphere within that elementary hypercube. The resulting set of points provides the upper bound above for $q(r,\zeta)$.

3. Define σ_{max}^2 and σ_{min}^2 to be the largest and smallest eigenvalues of the matrix

$$\frac{1}{n\sigma_a^2}[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z}\mathbf{Z}]^\intercal[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z}\mathbf{Z}]$$

respectively. We replace the definition of $\mathcal{E}_{\text{aux}}$ by

$$\begin{aligned}\mathcal{E}_{\text{aux}} \triangleq\ & \{\hat{\rho} - \|\mathbf{w}\|_2 \in (-\epsilon, \epsilon)\} \\ & \cap \left\{\sigma_{\text{max}}^2 \in (1-\epsilon, 1+\epsilon)\right\} \\ & \cap \left\{\sigma_{\text{min}}^2 \in (1-\epsilon, 1+\epsilon)\right\}.\end{aligned}$$

Consider the asymptotic behaviors of the events. First, note that

$$\sqrt{\frac{1}{n\sigma_a^2}\|\mathbf{Y}\|_2^2} = \sqrt{\frac{\|\mathbf{w}\|_2^2\sigma_a^2 + \sigma_z^2}{n\sigma_a^2}} \sqrt{\left\| \frac{\mathbf{Y}}{\sqrt{\|\mathbf{w}\|_2^2\sigma_a^2 + \sigma_z^2}} \right\|_2^2} \tag{2.58}$$

where $\sqrt{\left\| \frac{\mathbf{Y}}{\sqrt{\|\mathbf{w}\|_2^2\sigma_a^2+\sigma_z^2}} \right\|_2^2}$ is χ-distributed with mean $\sqrt{2}\,\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}$ and variance $\left(n - \frac{2\Gamma^2((n+1)/2)}{\Gamma^2(n/2)}\right)$. Then, $\sqrt{\frac{1}{n\sigma_a^2}\|\mathbf{Y}\|_2^2}$ has mean $\sqrt{\frac{\|\mathbf{w}\|_2^2\sigma_a^2+\sigma_z^2}{n\sigma_a^2}}\sqrt{2}\,\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}$ and variance $\frac{\|\mathbf{w}\|_2^2\sigma_a^2+\sigma_z^2}{n\sigma_a^2}\left(n - \frac{2\Gamma^2((n+1)/2)}{\Gamma^2(n/2)}\right)$.

It has been shown [22] that

$$\lim_{x\to\infty} \frac{x\Gamma(x)}{\sqrt{x+1/4}\Gamma(x+1/2)} = 1.$$

Then, as $n \to \infty$, $\sqrt{\frac{1}{n\sigma_a^2}\|\mathbf{Y}\|_2^2}$ has asymptotic mean $\sqrt{\frac{\|\mathbf{w}\|_2^2\sigma_a^2+\sigma_z^2}{\sigma_a^2}}$ and variance $\frac{\|\mathbf{w}\|_2^2\sigma_a^2+\sigma_z^2}{2n\sigma_a^2}$. Since $k = o(n)$, we have $\frac{\|\mathbf{w}\|_2^2\sigma_a^2+\sigma_z^2}{2n\sigma_a^2} \to 0$. Hence, $\lim_{m\to\infty} \mathsf{P}\{\hat{\rho} - \|\mathbf{w}\|_2 \in (-\epsilon, \epsilon)\} = 1$.

Second, σ_{max}^2 and σ_{min}^2 are shown [108] to almost surely converge to $(1+q)^2$ and $(1-q)^2$, respectively, where $q \triangleq \lim_{m\to\infty} \sqrt{(k+1)/n} = 0$. Thus, $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_{\text{aux}}^c) =$

0.

4. Next, we analyze the probability that the true support satisfies the recovery rule. Note that

$$
\begin{aligned}
\frac{1}{n}\left\|\mathbf{Y}-\sum_{j=1}^{k}\hat{W}_j\mathbf{A}_j\right\|_2^2 &= \frac{1}{n}\left\|\sum_{j=1}^{k}w_j\mathbf{A}_j+\mathbf{Z}-\sum_{j=1}^{k}\hat{W}_j\mathbf{A}_j\right\|_2^2 \\
&= \frac{1}{n}\left\|\left[\mathbf{A}_1,\ldots,\mathbf{A}_k,\frac{\sigma_a}{\sigma_z}\mathbf{Z}\right]\begin{bmatrix}\mathbf{w}-\hat{\mathbf{W}}\\ \frac{\sigma_z}{\sigma_a}\end{bmatrix}\right\|_2^2 \\
&\le \sigma_{\text{max}}^2\sigma_a^2\left\|\begin{bmatrix}\mathbf{w}-\hat{\mathbf{W}}\\ \frac{\sigma_z}{\sigma_a}\end{bmatrix}\right\|_2^2 \\
&= \sigma_{\text{max}}^2\sigma_a^2\|\mathbf{w}-\hat{\mathbf{W}}\|_2^2+\sigma_{\text{max}}^2\sigma_z^2. \qquad (2.59)
\end{aligned}
$$

By using the fact that $\sigma_{\text{max}}^2 \to 1$ almost surely as $n \to \infty$ and Lemma 2-1), we have $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_{1,2,\ldots,k}^c) = 0$.

5. Now, suppose we have proceeded to a step similar to (2.40) (that is, to be exact, equipped with the modified rule (2.57) and a proper $\mathcal{E}_{\text{cond}}$). Define the auxiliary vector $\mathbf{w}' \in \mathbb{R}^{k+1}$ as

$$
w'_j = \begin{cases} w_j & \text{if } j \in [k]\backslash\mathcal{T}, \\ w_j - \hat{W}_i & \text{if } j = s'_i \in \mathcal{T}, \\ \frac{\sigma_z}{\sigma_a} & \text{if } j = k+1. \end{cases} \qquad (2.60)
$$

Then,

$$\frac{1}{n}\left\|\sum_{j=1}^{k} w_j \mathbf{A}_j - \sum_{s'_j \in \mathcal{T}} \hat{W}_j \mathbf{A}_{s'_j} + \mathbf{Z}\right\|_2^2 = \frac{1}{n}\left\|\left[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z}\mathbf{Z}\right]\mathbf{w}'\right\|_2^2$$
$$\geq (1-\epsilon)\left\|\mathbf{w}'\right\|_2^2 \sigma_a^2$$
$$\geq (1-\epsilon)\left(\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2\right)\sigma_a^2 + \sigma_z^2\right).$$

From Lemma 1, it follows that (for sufficiently small ϵ)

$$\mathsf{P}\left(\frac{1}{n}\left\|\mathbf{Y} - \sum_{j=1}^{k} \hat{W}_j \mathbf{A}_{s'_j}\right\|_2^2 \leq (1+\epsilon)\sigma_z^2 + 2\epsilon^2\sigma_a^2 \middle| \mathcal{E}_{\text{cond}}\right)$$
$$\leq 2^{-\frac{n}{2}\log\frac{(1-\epsilon)\left(\left(\sum_{j\in[k]\setminus\mathcal{T}} w_j^2\right)\sigma_a^2+\sigma_z^2\right)}{(1+\epsilon)\sigma_z^2+2\epsilon^2\sigma_a^2}}.$$

6. Note that, from [121],

$$\binom{m}{k} \leq \left(\frac{me}{k}\right)^k.$$

Together with the modifications above, we follow the proof steps of Theorem 1 to reach

$$
\begin{aligned}
\mathsf{P}(\mathcal{E}) \leq\ & \mathsf{P}(\mathcal{E}^c_{\text{aux}}) + \mathsf{P}(\mathcal{E}^c_{1,2,\ldots,k}) + k! \cdot q(\|\mathbf{w}\|_2 + \epsilon, \epsilon) \\
& \cdot \sum_{\mathcal{T} \subseteq [k]} \left(\frac{me}{|\mathcal{T}|}\right)^{|\mathcal{T}|} \cdot 2^{-\frac{n}{2} \log \frac{(1-\epsilon)\left(\left(\sum_{j \in \mathcal{T}} w_j^2\right)\sigma_a^2 + \sigma_z^2\right)}{(1+\epsilon)\sigma_z^2 + 2\epsilon^2\sigma_a^2}} \\
\leq\ & \mathsf{P}(\mathcal{E}^c_{\text{aux}}) + \mathsf{P}(\mathcal{E}^c_{1,2,\ldots,k}) + k! \cdot q(\|\mathbf{w}\|_2 + \epsilon, \epsilon) \\
& \cdot \sum_{\mathcal{T} \subseteq [k]} \left(\frac{me}{|\mathcal{T}|}\right)^{|\mathcal{T}|} \cdot 2^{-\frac{n}{2} \log \frac{(1-\epsilon)\left(|\mathcal{T}| w_{\text{min}}^2 \sigma_a^2 + \sigma_z^2\right)}{(1+\epsilon)\sigma_z^2 + 2\epsilon^2\sigma_a^2}} \\
\leq\ & \mathsf{P}(\mathcal{E}^c_{\text{aux}}) + \mathsf{P}(\mathcal{E}^c_{1,2,\ldots,k}) + k! \cdot q(\|\mathbf{w}\|_2 + \epsilon, \epsilon) \\
& \cdot 2^k \cdot \max_{j \in [k]} \left[\left(\frac{me}{j}\right)^j \cdot 2^{-\frac{n}{2} \log \frac{(1-\epsilon)\left(j w_{\text{min}}^2 \sigma_a^2 + \sigma_z^2\right)}{(1+\epsilon)\sigma_z^2 + 2\epsilon^2\sigma_a^2}}\right].
\end{aligned}
\tag{2.61}
$$

Note that

$$
\begin{aligned}
& \log\left(k! \cdot q(\|\mathbf{w}\|_2 + \epsilon, \epsilon) \cdot 2^k \cdot \max_{j \in [k]} \left[\left(\frac{me}{j}\right)^j \cdot 2^{-\frac{n}{2} \log \frac{(1-\epsilon)\left(j w_{\text{min}}^2 \sigma_a^2 + \sigma_z^2\right)}{(1+\epsilon)\sigma_z^2 + 2\epsilon^2\sigma_a^2}}\right]\right) \\
& \leq k \log k + k \log(\eta_1 k^2 w_{\text{max}}/\epsilon) + k \\
& \quad + \max_{j \in [k]} \left[j \log \frac{me}{j} - \frac{n}{2} \log \frac{(1-\epsilon)\left(j w_{\text{min}}^2 \sigma_a^2 + \sigma_z^2\right)}{(1+\epsilon)\sigma_z^2 + 2\epsilon^2\sigma_a^2}\right].
\end{aligned}
\tag{2.62}
$$

It can be readily seen that from the condition (2.56), the upper bound in (2.62) becomes negative and thus $\mathsf{P}(\mathcal{E}) \to 0$ as $m \to \infty$.

2.8.4 Proof of Theorem 4

To establish this theorem, we prove the following equivalent statement:

If there exists a sequence of matrices $\{A^{(m)}\}_{m=k}^{\infty}$, $A^{(m)} \in \mathbb{R}^{n_m \times m}$, and a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{\{1,2,\ldots,m\}}$, such that

$$\frac{1}{n_m m}\|A^{(m)}\|_F^2 \leq \sigma_a^2$$

and

$$\lim_{m\to\infty} \mathsf{P}\{d^{(m)}(A^{(m)}\mathbf{X}(\mathbf{w}^{(m)},\mathbf{S}) + \mathbf{Z}) \neq \text{supp}(\mathbf{X}(\mathbf{w}^{(m)},\mathbf{S}))\} = 0$$

then

$$\limsup_{m\to\infty} \frac{2k_m \log(m/k_m)}{n_m \log\left(\frac{2k_m w_{\text{max}}^2 \sigma_a^2}{\sigma_z^2} + 1\right)} \leq 1.$$

To justify this alternative claim, we follow the steps for the proof of Theorem 2 in Appendix 2.8.2. Necessary modifications and clarifications are presented as follows.

1. Note that

$$\epsilon_{1,n} = \frac{1}{n}\log\left(m^{|\mathcal{T}|}/\prod_{q=0}^{|\mathcal{T}|-1}(m-(k-|\mathcal{T}|)-q)\right) \leq \frac{|\mathcal{T}|}{n}\log\frac{m}{m-k+1}. \quad (2.63)$$

2. For any $\mathcal{T} \subseteq [k]$, we follow (2.53) in Appendix 2.8.2 to reach

$$\begin{aligned} &|\mathcal{T}|\log m - \log k! - \overline{P}_e^{(m)}\log\binom{m}{k} - 1 - n\epsilon_{1,n} \\ &\leq \frac{n}{2}\log\left(\left(\sum_{j\in\mathcal{T}} w_j^2 - \tau(m)\right)\frac{\sigma_a^2}{\sigma_z^2} + 1\right). \end{aligned} \quad (2.64)$$

Note that

$$\sum_{j\in\mathcal{T}} w_j^2 - \tau(m) \le \left|\sum_{j\in\mathcal{T}} w_j^2\right| + \left|\sum_{j\in\mathcal{T}}\sum_{l\in\mathcal{T},l\neq j} \frac{w_j w_l}{(m-1)}\right|$$
$$\le |\mathcal{T}| w_{\text{max}}^2 + \frac{|\mathcal{T}|(|\mathcal{T}|-1) w_{\text{max}}^2}{(m-1)}$$
$$\le 2|\mathcal{T}| w_{\text{max}}^2. \tag{2.65}$$

Then, it follows from (2.63), (2.64), and (2.65) that the inequality

$$|\mathcal{T}|\log m - k\log k - \overline{P}_e^{(m)} k \log m - 1 - n\cdot\frac{|\mathcal{T}|}{n}\log\frac{m}{m-k+1}$$
$$\le \frac{n}{2}\log\left(2|\mathcal{T}| w_{\text{max}}^2 \frac{\sigma_a^2}{\sigma_z^2} + 1\right)$$

must hold for any $\mathcal{T} \subseteq [k]$. By choosing $\mathcal{T} = [k]$, we have

$$\limsup_{m\to\infty} \frac{2k_m\left(\log(m-k_m+1) - \log k_m - \frac{1}{k_m}\right)}{n_m \log\left(\frac{2k_m w_{\text{max}}^2 \sigma_a^2}{\sigma_z^2} + 1\right)} \le 1$$

which equivalently gives

$$\limsup_{m\to\infty} \frac{2k_m\left(\log m - \log k_m\right)}{n_m \log\left(\frac{2k_m w_{\text{max}}^2 \sigma_a^2}{\sigma_z^2} + 1\right)} \le 1.$$

2.8.5 Proof of Theorem 5

For the ease of exposition, we consider two distinct cases on the number of nonzero rows.

Case 1: $k = 1$. In this case, the signal of interest is $X = X(W, S_1)$, where $W = [w_{1,1}, ..., w_{1,l}]$. Fix $\epsilon > 0$. We first form an estimate $\hat{\rho}_i$ of $|w_{1,i}|$ for $i \in [l]$ as

$$\hat{\rho}_i \triangleq \sqrt{\frac{|\frac{1}{n_m}\|\mathbf{Y}_i\|_2^2 - \sigma_z^2|}{\sigma_a^2}}. \tag{2.66}$$

Declare that $\hat{s}_1 \in [m]$ is the estimated index of the nonzero row, i.e., $d^{(m)}(Y) = \{\hat{s}_1\}$, if it is the unique index such that

$$\frac{1}{nl}\left\|Y - \mathbf{A}_{\hat{s}_1}\left[(-1)^{q_1}\hat{\rho}_1, ..., (-1)^{q_l}\hat{\rho}_l\right]\right\|_F^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2 \tag{2.67}$$

for $q_i = 1$ or $q_i = 2$, $i \in [l]$. If there is none or more than one such index, pick an arbitrary index.

We analyze the average probability of error

$$\mathsf{P}(\mathcal{E}) = \mathsf{P}\{d^{(m)}(Y) \neq \text{supp}(X(W, S_1))\}. \tag{2.68}$$

Due to the symmetry in the problem and the measurement matrix generation, we assume without loss of generality $S_1 = 1$, that is,

$$Y = \mathbf{A}_1^{(m)} W + Z \tag{2.69}$$

for some $W = [w_{1,1}, ..., w_{1,l}] \in \mathbb{R}^{1\times l}$. In the following analysis, we drop superscripts and subscripts on m for notational simplicity when no ambiguity arises. Define the events

$$\begin{aligned}\mathcal{E}_s \triangleq \Bigg\{&\forall i \in [l], \exists q_i \in \{1,2\},\\ &\text{such that } \frac{1}{nl}\left\|Y - \mathbf{A}_s[(-1)^{q_i}\hat{\rho}_1, ..., (-1)^{q_l}\hat{\rho}_l]\right\|_F^2 \le \sigma_z^2 + \epsilon^2\sigma_a^2\Bigg\},\ s \in [m].\end{aligned}$$

Then,

$$\mathsf{P}(\mathcal{E}) \le \mathsf{P}\left(\mathcal{E}_1^c \cup \left(\cup_{s=2}^m \mathcal{E}_s\right)\right). \tag{2.70}$$

Let

$$\begin{aligned}\mathcal{E}_{\text{aux}} \triangleq &\left\{\det\left(\frac{1}{n}\left(\mathbf{A}_1 W + Z\right)^\intercal\left(\mathbf{A}_1 W + Z\right)\right) - \det\left(\sigma_a^2 W^\intercal W + \sigma_z^2 I\right) \in (-\epsilon, \epsilon)\right\}\\ &\cap\left(\bigcap_{i=1}^{l}\{\hat{\rho}_i - |w_{1,i}| \in (-\epsilon,\epsilon)\}\right).\end{aligned}$$

Then, by the union of events bound and the fact that $\mathcal{A}^c \cup \mathcal{B} = \mathcal{A}^c \cup (\mathcal{B} \cap \mathcal{A})$,

$$\mathsf{P}(\mathcal{E}) \leq \mathsf{P}(\mathcal{E}_{\text{aux}}^c) + \mathsf{P}(\mathcal{E}_1^c) + \sum_{s=2}^{m} \mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}}). \tag{2.71}$$

We bound each term in (2.71). First, by the weak law of large numbers (LLN), $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_{\text{aux}}^c) = 0$. Next, we consider $\mathsf{P}(\mathcal{E}_1^c)$. It can be readily seen that, with $q_i = (3 + \text{sign}(w_{1,i}))/2$,

$$\lim_{m\to\infty} \mathsf{P}\left(\frac{1}{nl}\|Y - \mathbf{A}_1[(-1)^{q_1}\hat{\rho}_1, ..., (-1)^{q_l}\hat{\rho}_l]\|_F^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2\right) = 1. \tag{2.72}$$

Hence, $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_1^c) = 0$.

Next, we consider the third term in (2.71). We need the following lemma, whose proof is presented at the end of this appendix.

Lemma 3. *Let $B \in \mathbb{R}^{n\times l}$ be a fixed matrix satisfying $(\prod_{i=1}^{l}[\frac{1}{n}B^{\intercal}B]_{i,i})^{\frac{1}{l}} \equiv \alpha > 0$. Let $\mathcal{S} \subseteq [l]$ be a fixed set. Let $D \in \mathbb{R}^{n\times l}$ be a matrix such that, for $j \in \mathcal{S}$, $\mathbf{D}_j \sim \mathcal{N}(\mathbf{0}, \theta_j I)$ with some $\theta_j > 0$; for $j \in [l]\backslash\mathcal{S}$, $\mathbf{D}_j \equiv \mathbf{0}$. All columns of D are independent. Then, for any $\gamma \in (0, \alpha)$,*

$$\mathsf{P}\left(\frac{1}{nl}\|B - D\|_F^2 \leq \gamma\right) \leq 2^{-\frac{n}{2}\log\frac{\alpha^l}{\gamma^l}}. \tag{2.73}$$

We continue the proof of Theorem 5. Consider $\mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}})$ for $s \neq 1$. Note that

$$\mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}}) \leq \mathsf{P}(\mathcal{E}_s | \mathcal{E}_{\text{aux}}) = \int_{Y_1 \in \mathcal{E}_{\text{aux}}} \mathsf{P}(\mathcal{E}_s | \{Y = Y_1\} \cap \mathcal{E}_{\text{aux}}) f(Y_1 | \mathcal{E}_{\text{aux}}) dY_1.$$

Let $[(-1)^{q_1}\hat{\rho}_1, ..., (-1)^{q_l}\hat{\rho}_l] = U\Theta V^\intercal$ denote the singular value decomposition. Since $\mathbf{A}_s$ is independent of Y and $\hat{\rho}_i$ for $s \neq 1$, it follows from Lemma 3 that (by treating $B = YV$ and $D = \mathbf{A}_s U\Theta$), for $q_i = \{1, 2\}, i \in [l]$ and sufficiently small ϵ,

$$\begin{aligned}
&\mathsf{P}\left(\frac{1}{nl}\left\|Y - \mathbf{A}_s[(-1)^{q_1}\hat{\rho}_1, ..., (-1)^{q_l}\hat{\rho}_l]\right\|_F^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2 \middle| \{Y = Y_1\} \cap \mathcal{E}_{\text{aux}}\right) \\
&= \mathsf{P}\left(\frac{1}{nl}\left\|YV - \mathbf{A}_s U\Theta\right\|_F^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2 \middle| \{Y = Y_1\} \cap \mathcal{E}_{\text{aux}}\right) \\
&\leq 2^{-\frac{n}{2}\log\frac{\prod_{i=1}^{l}[\frac{1}{n}V^\intercal Y^\intercal YV]_{i,i}}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}} \\
&\leq 2^{-\frac{n}{2}\log\frac{\det(\frac{1}{n}V^\intercal Y^\intercal YV)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}} \qquad (2.74) \\
&\leq 2^{-\frac{n}{2}\log\frac{\det(\frac{1}{n}Y^\intercal Y)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}} \\
&\leq 2^{-\frac{n}{2}\log\left(\frac{\det\left(\sigma_a^2 W^\intercal W + \sigma_z^2 I\right) - \epsilon}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}\right)}
\end{aligned}$$

where (2.74) follows from the Hadamard's inequality [32]. Thus,

$$\mathsf{P}(\mathcal{E}_s | \{Y = Y_1\} \cap \mathcal{E}_{\text{aux}}) \leq 2^l \cdot 2^{-\frac{n}{2}\log\left(\frac{\det\left(\sigma_a^2 W^\intercal W + \sigma_z^2 I\right) - \epsilon}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}\right)}$$

and hence

$$\sum_{s=2}^{m} \mathsf{P}(\mathcal{E}_s \cap \mathcal{E}_{\text{aux}}) \leq 2^l \cdot m \cdot 2^{-\frac{n}{2}\log\left(\frac{\det\left(\sigma_a^2 W^\intercal W + \sigma_z^2 I\right) - \epsilon}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}\right)}$$

which tends to zero as $m \to \infty$, if

$$\limsup_{m\to\infty} \frac{\log m}{n_m} < \frac{1}{2}\log\left(\frac{\det\left(\sigma_a^2 W^\intercal W + \sigma_z^2 I\right) - \epsilon}{(\sigma_z^2 + \epsilon^2\sigma_a^2)^l}\right). \tag{2.75}$$

Since $\epsilon > 0$ is chosen arbitrarily, we have the desired proof of Theorem 5.

Case 2: $k \geq 2$. In this case, the signal of interest is $X = X(W, \mathbf{S})$. Fix $\epsilon > 0$. First, for $i \in [l]$, we form an estimate of $\|\mathbf{w}_i\|_2$ as

$$\hat{\rho}_i \triangleq \sqrt{\frac{|\frac{1}{n}\|\mathbf{Y}_i\|_2^2 - \sigma_z^2|}{\sigma_a^2}}. \tag{2.76}$$

For $i \in [l]$, given $\hat{\rho}_i$ and ϵ, fix $\mathcal{Q}_i = \mathcal{Q}_i(\hat{\rho}_i, \epsilon)$. Declare $d(Y) = \{\hat{s}_1, ..., \hat{s}_k\} \subseteq [m]$ is the recovered set of indices of nonzero rows of W, if it is the unique set of indices such that

$$\frac{1}{nl}\left\|Y - [\mathbf{A}_{\hat{s}_1}, ..., \mathbf{A}_{\hat{s}_k}]\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 \leq \sigma_z^2 + \epsilon^2\sigma_a^2 \tag{2.77}$$

for some $\hat{\mathbf{W}}_i \in \mathcal{Q}_i$, $i \in [l]$. If there is none or more than one such set, pick an arbitrary set of k indices.

Next, we analyze the average probability of error

$$\mathsf{P}(\mathcal{E}) = \mathsf{P}\{d(Y) \neq X(W, \mathbf{S})\}. \tag{2.78}$$

Without loss of generality, we assume that $S_j = j$ for $j = 1, 2, ..., k$, which gives

$$Y = [\mathbf{A}_1, ..., \mathbf{A}_k]W + Z \tag{2.79}$$

for some W. Define the event

$$\begin{aligned}\mathcal{E}_{s_1,s_2,...,s_k} \triangleq \Bigg\{ &\exists \hat{\mathbf{W}}_i \in \mathcal{Q}_i \text{ and } \{s'_1, s'_2, ..., s'_k\} = \{s_1, s_2, ..., s_k\} \\ &\text{such that } \frac{1}{nl} \left\| Y - [\mathbf{A}_{s'_1}, ..., \mathbf{A}_{s'_k}] \left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right] \right\|_F^2 \leq \sigma_z^2 + \epsilon^2 \sigma_a^2 \Bigg\}.\end{aligned}$$

Define σ^2_{max} and σ^2_{min} to be the largest and smallest eigenvalues of the matrix

$$\frac{1}{n\sigma_a^2}[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z} Z]^\intercal [\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z} Z]$$

respectively. Then

$$\begin{aligned}\mathsf{P}(\mathcal{E}) &= \mathsf{P}\left(\mathcal{E}^c_{1,2,...,k} \cup \left(\bigcup_{s_1 < \cdots < s_k : \{s_1,...,s_k\} \neq [k]} \mathcal{E}_{s_1,s_2,...,s_k} \right)\right) \\ &\leq \mathsf{P}\left(\mathcal{E}^c_{\text{aux}} \cup \mathcal{E}^c_{1,2,...,k} \cup \left(\bigcup_{s_1 < \cdots < s_k : \{s_1,...,s_k\} \neq [k]} (\mathcal{E}_{s_1,s_2,...,s_k} \cap \mathcal{E}_{\text{aux}}) \right)\right) \\ &\leq \mathsf{P}(\mathcal{E}^c_{\text{aux}}) + \mathsf{P}(\mathcal{E}^c_{1,2,...,k}) + \sum_{s_1 < \cdots < s_k : \{s_1,...,s_k\} \neq [k]} \mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} \cap \mathcal{E}_{\text{aux}})\end{aligned} \tag{2.80}$$

where

$$
\begin{aligned}
&\mathcal{E}_{\text{aux}} \triangleq \\
&\left\{\sigma_{\text{max}}^2 \in (1-\epsilon, 1+\epsilon)\right\} \cap \left\{\sigma_{\text{min}}^2 \in (1-\epsilon, 1+\epsilon)\right\} \cap \left(\bigcap_{i=1}^{l} \{\hat{\rho}_i - \|\mathbf{w}_i\|_2 \in (-\epsilon, \epsilon)\} \right).
\end{aligned}
$$

We analyze each term in (2.80). First, note that $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_{\text{aux}}) = 1$ due to LLN and the properties of the extreme eigenvalues of random matrices [108]. Next, consider

$$
\begin{aligned}
&\frac{1}{nl}\left\|Y - [\mathbf{A}_1, ..., \mathbf{A}_k]\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 \\
&= \frac{1}{nl}\left\|[\mathbf{A}_1, ..., \mathbf{A}_k]W + Z - [\mathbf{A}_1, ..., \mathbf{A}_k]\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 \\
&= \frac{1}{nl}\left\|[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z}Z]\begin{bmatrix} W - \left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right] \\ \frac{\sigma_z}{\sigma_a} I_{l\times l}\end{bmatrix}\right\|_F^2 \\
&\le \frac{1}{l}\sigma_{\text{max}}^2\sigma_a^2 \left\|\begin{bmatrix} W - \left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right] \\ \frac{\sigma_z}{\sigma_a} I_{l\times l}\end{bmatrix}\right\|_F^2 \\
&= \sigma_{\text{max}}^2\left(\frac{\sigma_a^2}{l}\left\|W - \left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 + \sigma_z^2\right)
\end{aligned}
\tag{2.81}
$$

By using the fact that $\sigma_{\text{max}}^2 \to 1$ almost surely as $n \to \infty$ [108] and Lemma 2-1), we have $\lim_{m\to\infty} \mathsf{P}(\mathcal{E}_{1,2,...,k}^c) = 0$.

Next, we consider $\mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} \cap \mathcal{E}_{\text{aux}})$ for $\{s_1, s_2, ..., s_k\} \neq [k]$. Note that

$$
\begin{aligned}
&\mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} \cap \mathcal{E}_{\text{aux}}) \\
&\leq \mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} | \mathcal{E}_{\text{aux}}) \\
&= \int \cdots \int_{\{\mathbf{a}_1,...,\mathbf{a}_k,Z_0\}\in\mathcal{E}_{\text{aux}}} \mathsf{P}(\mathcal{E}_{s_1,s_2,...,s_k} | \{\mathbf{A}_1 = \mathbf{a}_1\} \cap \cdots \cap \{\mathbf{A}_k = \mathbf{a}_k\} \cap \{Z = Z_0\} \\
&\qquad\qquad \cap \mathcal{E}_{\text{aux}}) \times f(\mathbf{a}_1, ..., \mathbf{a}_k, Z_0 | \mathcal{E}_{\text{aux}}) d\mathbf{a}_1 \cdots d\mathbf{a}_k dZ_0.
\end{aligned}
\tag{2.82}
$$

For notational simplicity, define $\xi \triangleq \sigma_z^2 + \epsilon^2\sigma_a^2$, $\mathcal{T} \triangleq \{s_1, s_2, ..., s_k\} \cap [k]$, $\mathcal{T}^c \triangleq \{s_1, s_2, ..., s_k\} \backslash \mathcal{T}$, and $\mathcal{E}_{\text{cond}} \triangleq \{\mathbf{A}_1 = \mathbf{a}_1\} \cap \cdots \cap \{\mathbf{A}_k = \mathbf{a}_k\} \cap \{Z = Z_0\} \cap \mathcal{E}_{\text{aux}}$. For any permutation $(s_1', s_2', ..., s_k')$ of $\{s_1, s_2, ..., s_k\}$ and any $\hat{\mathbf{W}}_i \in \mathcal{Q}_i$, $i \in [l]$,

$$
\begin{aligned}
&\mathsf{P}\left(\frac{1}{nl}\left\|Y - [\mathbf{A}_{s_1'}, ..., \mathbf{A}_{s_k'}]\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&= \mathsf{P}\left(\frac{1}{nl}\left\|[\mathbf{A}_1, ..., \mathbf{A}_k]W + Z - [\mathbf{A}_{s_1'}, ..., \mathbf{A}_{s_k'}]\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right)
\end{aligned}
\tag{2.83}
$$

Define the matrix $W' \in \mathbb{R}^{k\times l}$ as

$$
\underline{W'}_j = \begin{cases} \underline{W}_j & \text{if } j \in [k]\backslash\mathcal{T} \\ \underline{W}_j - \underline{\hat{\mathbf{W}}}_i & \text{if } j = s_i' \in \mathcal{T} \end{cases}
\tag{2.84}
$$

where $\underline{\hat{\mathbf{W}}}_i$ denotes the ith row of the matrix $\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]$. Define $\widetilde{W'} \in \mathbb{R}^{k\times l}$ as

$$\underline{\widetilde{W}}'_j = \begin{cases} \underline{\hat{\mathbf{W}}}_j & \text{if } s'_j \notin \mathcal{T} \\ \underline{\mathbf{0}} & \text{if } s'_j \in \mathcal{T} \end{cases} \tag{2.85}$$

where $\underline{\mathbf{0}}$ is a zero row vector of a proper size. Then, continue from (2.83), we have

$$\begin{aligned}
&\mathsf{P}\left(\frac{1}{nl}\left\|Y - [\mathbf{A}_{s'_1}, ..., \mathbf{A}_{s'_k}]\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&= \mathsf{P}\left(\frac{1}{nl}\left\|[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z}Z]\begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a}I \end{bmatrix} - [\mathbf{A}_{s'_1}, ..., \mathbf{A}_{s'_k}]\widetilde{W'}\right\|_F^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&\equiv \mathsf{P}\left(\frac{1}{nl}\left\|[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z}Z]\begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a}I \end{bmatrix} - \widetilde{A}\widetilde{W}'_1\right\|_F^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) && (2.86) \\
&= \mathsf{P}\left(\frac{1}{nl}\left\|[\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z}Z]\begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a}I \end{bmatrix} V - \widetilde{A}U\Theta\right\|_F^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) && (2.87)
\end{aligned}$$

where in (2.86) $\widetilde{W}'_1$ denotes matrix formed by removing the zero rows in $\widetilde{W'}$, and $\widetilde{A}$ denotes the matrix by removing columns of $[\mathbf{A}_{s'_1}, ..., \mathbf{A}_{s'_k}]$ indexed by the indices of the zero rows of $\widetilde{W'}$. By construction, $\widetilde{A}$ is independent of $\mathbf{A}_1, ..., \mathbf{A}_k$. To reach (2.87), let $\widetilde{W}'_1 = U\Theta V^\intercal$ denote the singular value decomposition. The follow lemma, the proof of which is presented at the end of this section, is useful.

Lemma 4. *Let $B \in \mathbb{R}^{p\times q}$, $D \in \mathbb{R}^{q\times r}$. Let σ_b^2 denote the smallest eigenvalue of $B^\intercal B$.*

Then

$$\det((BD)^\intercal BD) \geq (\sigma_b^2)^r \det(D^\intercal D).$$

Let $M \triangleq [\mathbf{A}_1, ..., \mathbf{A}_k, \frac{\sigma_a}{\sigma_z} Z] \begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix} V$. Conditioned on $\mathcal{E}_{\text{cond}}$ and the chosen $\mathcal{Q}_i$ for $i \in [l]$, M is fixed. According to Lemma 4,

$$\det\left(\frac{1}{n} M^\intercal M\right) \geq ((1-\epsilon)\sigma_a^2)^l \det\left(\begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}^\intercal \begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}\right). \tag{2.88}$$

Continue with (2.87). Using Lemma 3 (treating $B = M$ and $D = \widetilde{A} U \Theta$), we have

$$\begin{aligned}
&\mathsf{P}\left(\frac{1}{nl}\left\|Y - [\mathbf{A}_{s_1'}, ..., \mathbf{A}_{s_k'}]\left[\hat{\mathbf{W}}_1, ..., \hat{\mathbf{W}}_l\right]\right\|_F^2 \leq \xi \middle| \mathcal{E}_{\text{cond}}\right) \\
&\leq 2^{-\frac{n}{2}\log \frac{\prod_{i=1}^l [\frac{1}{n} M^\intercal M]_{i,i}}{(\sigma_z^2 + \epsilon^2 \sigma_a^2)^l}} \\
&\leq 2^{-\frac{n}{2}\log \frac{\det(\frac{1}{n} M^\intercal M)}{(\sigma_z^2 + \epsilon^2 \sigma_a^2)^l}} \\
&\leq 2^{-\frac{n}{2}\log \frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}^\intercal \begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}\right)}{(\sigma_z^2 + \epsilon^2 \sigma_a^2)^l}} \\
&\leq 2^{-\frac{n}{2}\log \frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\begin{bmatrix} \underline{W}_{[k]\setminus\mathcal{T}} \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}^\intercal \begin{bmatrix} \underline{W}_{[k]\setminus\mathcal{T}} \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}\right)}{(\sigma_z^2 + \epsilon^2 \sigma_a^2)^l}} &(2.89) \\
&= 2^{-\frac{n}{2}\log \frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\underline{W}_{[k]\setminus\mathcal{T}}^\intercal \underline{W}_{[k]\setminus\mathcal{T}} + \frac{\sigma_z^2}{\sigma_a^2} I\right)}{(\sigma_z^2 + \epsilon^2 \sigma_a^2)^l}} &(2.90)
\end{aligned}$$

where (2.89) uses the fact that

$$\begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}^\intercal \begin{bmatrix} W' \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix} = \begin{bmatrix} \underline{W}_{[k]\setminus\mathcal{T}} \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix}^\intercal \begin{bmatrix} \underline{W}_{[k]\setminus\mathcal{T}} \\ \frac{\sigma_z}{\sigma_a} I \end{bmatrix} + \begin{bmatrix} \underline{W}'_{\mathcal{T}} \\ O \end{bmatrix}^\intercal \begin{bmatrix} \underline{W}'_{\mathcal{T}} \\ O \end{bmatrix}$$

where O denotes the matrix with elements all being zeros, and the fact that [8, Corollary 8.4.15], for positive semidefinite $B, D \in \mathbb{R}^{l\times l}$, $\det(B+D) \geq \det(B)$. By the union of events bound,

$$\begin{aligned}
&\mathsf{P}(\mathcal{E}_{s_1,s_2,\ldots,s_k}|\mathcal{E}_{\text{cond}}) \\
&\leq \sum_{\{s'_1,\ldots,s'_k\}=\{s_1,\ldots,s_k\}} \mathsf{P}\Bigg(\forall i, \exists \hat{\mathbf{W}}_i \in \mathcal{Q}_i \\
&\qquad\qquad \text{such that } \frac{1}{nl}\left\|Y - [\mathbf{A}_{s'_1}, \ldots, \mathbf{A}_{s'_k}]\left[\hat{\mathbf{W}}_1, \ldots, \hat{\mathbf{W}}_l\right]\right\|_F^2 \leq \xi \,\Bigg|\, \mathcal{E}_{\text{cond}}\Bigg) \\
&\leq \sum_{\{s'_1,\ldots,s'_k\}=\{s_1,\ldots,s_k\}} \sum_{\hat{\mathbf{W}}_1\in\mathcal{Q}_1} \\
&\qquad\qquad \cdots \sum_{\hat{\mathbf{W}}_l\in\mathcal{Q}_l} \mathsf{P}\left(\frac{1}{nl}\left\|Y - [\mathbf{A}_{s'_1}, \ldots, \mathbf{A}_{s'_k}]\left[\hat{\mathbf{W}}_1, \ldots, \hat{\mathbf{W}}_l\right]\right\|_F^2 \leq \xi \,\middle|\, \mathcal{E}_{\text{cond}}\right) \\
&\leq k! \cdot \left(\prod_{i=1}^{l} |\mathcal{Q}_i|\right) \cdot 2^{-\frac{n}{2}\log \frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\underline{W}_{[k]\setminus\mathcal{T}}^\intercal \underline{W}_{[k]\setminus\mathcal{T}} + \frac{\sigma_z^2}{\sigma_a^2} I\right)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}}.
\end{aligned}$$

Furthermore, conditioned on $\mathcal{E}_{\text{aux}}$, $\hat{\rho}_i < \|\mathbf{w}_i\|_2 + \epsilon$ for $i \in [l]$ and hence $|\mathcal{Q}_i| \leq$

$q_i(\|\mathbf{w}_i\|_2+\epsilon,\epsilon)$ by Lemma 2-2). Thus,

$$\mathsf{P}(\mathcal{E}_{s_1,s_2,\ldots,s_k}\cap\mathcal{E}_{\text{aux}}) \le k!\cdot\left(\prod_{i=1}^{l} q_i(\|\mathbf{w}_i\|_2+\epsilon,\epsilon)\right)\cdot 2^{-\frac{n}{2}\log\frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\underline{W}_{[k]\setminus\mathcal{T}}^{\intercal}\underline{W}_{[k]\setminus\mathcal{T}}+\frac{\sigma_z^2}{\sigma_a^2}I\right)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}}. \tag{2.91}$$

Note that the probability upper-bound (2.91) depends on $s_1,\ldots,s_k$ only through $\mathcal{T}$. Grouping the $\binom{m-k}{k-|\mathcal{T}|}$ events $\{\mathcal{E}_{s_1,s_2,\ldots,s_k}\cap\mathcal{E}_{\text{aux}}\}$ with the same $\mathcal{T}$,

$$\begin{aligned}
\mathsf{P}(\mathcal{E}) &\le \mathsf{P}(\mathcal{E}_{\text{aux}}^c)+\mathsf{P}(\mathcal{E}_{1,2,\ldots,k}^c)+\sum_{\mathcal{T}\subset[k]}\binom{m-k}{k-|\mathcal{T}|}\cdot k!\cdot\left(\prod_{i=1}^{l} q_i(\|\mathbf{w}_i\|_2+\epsilon,\epsilon)\right)\\
&\qquad\cdot 2^{-\frac{n}{2}\log\frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\underline{W}_{[k]\setminus\mathcal{T}}^{\intercal}\underline{W}_{[k]\setminus\mathcal{T}}+\frac{\sigma_z^2}{\sigma_a^2}I\right)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}}\\
&\le \mathsf{P}(\mathcal{E}_{\text{aux}}^c)+\mathsf{P}(\mathcal{E}_{1,2,\ldots,k}^c)+k!\cdot\left(\prod_{i=1}^{l} q_i(\|\mathbf{w}_i\|_2+\epsilon,\epsilon)\right)\\
&\qquad\cdot\sum_{\mathcal{T}\subset[k]} 2^{(k-|\mathcal{T}|)\log m}\cdot 2^{-\frac{n}{2}\log\frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\underline{W}_{[k]\setminus\mathcal{T}}^{\intercal}\underline{W}_{[k]\setminus\mathcal{T}}+\frac{\sigma_z^2}{\sigma_a^2}I\right)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}}\\
&= \mathsf{P}(\mathcal{E}_{\text{aux}}^c)+\mathsf{P}(\mathcal{E}_{1,2,\ldots,k}^c)+k!\cdot\left(\prod_{i=1}^{l} q_i(\|\mathbf{w}_i\|_2+\epsilon,\epsilon)\right)\\
&\qquad\cdot\sum_{\mathcal{T}\subseteq[k]} 2^{|\mathcal{T}|\log m}\cdot 2^{-\frac{n}{2}\log\frac{\left((1-\epsilon)\sigma_a^2\right)^l \det\left(\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}+\frac{\sigma_z^2}{\sigma_a^2}I\right)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l}}
\end{aligned}$$

which tends to zero as $m\to\infty$, if

$$\limsup_{m\to\infty}\frac{\log m}{n_m} < \frac{1}{2|\mathcal{T}|}\log\frac{((1-\epsilon)\sigma_a^2)^l\det\left(\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}+\frac{\sigma_z^2}{\sigma_a^2}I\right)}{(\sigma_z^2+\epsilon^2\sigma_a^2)^l} \tag{2.92}$$

for all $\mathcal{T}\subseteq[k]$. Since $\epsilon>0$ is arbitrarily chosen, the proof of Theorem 5 is complete.

Next, we prove Lemma 3. For $j \in \mathcal{S}$, $(b_{i,j} - D_{i,j})^2/\theta_j$ is a noncentral χ^2 random variable. Its moment generating function is [74] (for $t < 1/2$)

$$\mathsf{E}[e^{t(b_{i,j}-D_{i,j})^2/\theta_j}] = \frac{e^{\frac{tb_{i,j}^2/\theta_j}{1-2t}}}{(1-2t)^{\frac{1}{2}}}.$$

By changing variable $\frac{\theta_j t}{nl} \to t$, we have

$$\mathsf{E}[e^{\frac{t(b_{i,j}-D_{i,j})^2}{nl}}] = \frac{e^{\frac{\frac{t}{nl}b_{i,j}^2}{1-\frac{2\theta_j t}{nl}}}}{(1-\frac{2\theta_j t}{nl})^{\frac{1}{2}}}.$$

For $j \in [l]\backslash\mathcal{S}$ with $\mathbf{D}_j \equiv \mathbf{0}$, we additionally define $\theta_j = 0$. In this case,

$$\mathsf{E}[e^{\frac{t(b_{i,j}-D_{i,j})^2}{nl}}] = \mathsf{E}[e^{\frac{tb_{i,j}^2}{nl}}] = e^{\frac{t}{nl}b_{i,j}^2} = \frac{e^{\frac{\frac{t}{nl}b_{i,j}^2}{1-\frac{2\theta_j t}{nl}}}}{(1-\frac{2\theta_j t}{nl})^{\frac{1}{2}}}.$$

Define

$$S_n \triangleq \frac{1}{nl}\|B - D\|_F^2 = \frac{1}{l}\sum_{j=1}^{l}\frac{1}{n}\|\mathbf{b}_j - \mathbf{D}_j\|_2^2.$$

Then, we have

$$
\begin{aligned}
\mathsf{E}[e^{tS_n}] &= \mathsf{E}[e^{\frac{t}{nl}\|B-D\|_F^2}] \\
&= \mathsf{E}[e^{\frac{t}{l}\sum_{j=1}^{l}\frac{1}{n}\|\mathbf{b}_j-\mathbf{D}_j\|_2^2}] \\
&= \prod_{j=1}^{l}\mathsf{E}[e^{\frac{t}{nl}\|\mathbf{b}_j-\mathbf{D}_j\|_2^2}] \\
&= \prod_{j=1}^{l}\frac{e^{\frac{\frac{t}{nl}\|\mathbf{b}_j\|_2^2}{1-\frac{2\theta_j t}{nl}}}}{(1-\frac{2\theta_j t}{nl})^{\frac{n}{2}}}
\end{aligned}
$$

The Chernoff bound indicates that

$$
\begin{aligned}
&\mathsf{P}(S_n \le \gamma) \\
&\le \min_{s>0} e^{s\gamma}\, \mathsf{E}[e^{-sS_n}] \\
&= \min_{s>0} e^{s\gamma} \prod_{j=1}^{l} \frac{e^{\frac{-\frac{s}{nl}\|\mathbf{b}_j\|_2^2}{1+\frac{2\theta_j s}{nl}}}}{(1+\frac{2\theta_j s}{nl})^{\frac{n}{2}}} \\
&= \min_{p<0} e^{-p\gamma} \prod_{j=1}^{l} \frac{e^{\frac{\frac{p}{nl}\|\mathbf{b}_j\|_2^2}{1-\frac{2\theta_j p}{nl}}}}{(1-\frac{2\theta_j p}{nl})^{\frac{n}{2}}} \\
&= \exp\left\{\min_{p<0}\left\{-p\gamma + \sum_{j=1}^{l}\left[\frac{\frac{p}{nl}\|\mathbf{b}_j\|_2^2}{1-\frac{2\theta_j p}{nl}} - \frac{n}{2}\log\left(1-\frac{2\theta_j p}{nl}\right)\right]\right\}\right\} \\
&= \exp\left\{\min_{p<0}\left\{-lp\gamma + \sum_{j=1}^{l}\left[\frac{\frac{p}{n}\|\mathbf{b}_j\|_2^2}{1-\frac{2\theta_j p}{n}} - \frac{n}{2}\log\left(1-\frac{2\theta_j p}{n}\right)\right]\right\}\right\} \\
&= \exp\left\{\min_{p<0}\left\{-lp\gamma - \sum_{j=1}^{l}\underbrace{\frac{\frac{-p}{n}\|\mathbf{b}_j\|_2^2}{1-\frac{2\theta_j p}{n}}}_{>0} - \frac{n}{2}\log\prod_{j=1}^{l}\left(1-\frac{2\theta_j p}{n}\right)\right\}\right\} \\
&\le \exp\left\{\min_{p<0}\left\{-lp\gamma - l\left(\prod_{j=1}^{l}\frac{\frac{-p}{n}\|\mathbf{b}_j\|_2^2}{1-\frac{2\theta_j p}{n}}\right)^{\frac{1}{l}} - \frac{n}{2}\log\prod_{j=1}^{l}\left(1-\frac{2\theta_j p}{n}\right)\right\}\right\} \qquad (2.93) \\
&= \exp\left\{\min_{p<0}\left\{-lp\gamma - l\frac{\left(\prod_{j=1}^{l}\frac{-p}{n}\|\mathbf{b}_j\|_2^2\right)^{\frac{1}{l}}}{\left(\prod_{j=1}^{l}(1-\frac{2\theta_j p}{n})\right)^{\frac{1}{l}}} - \frac{nl}{2}\log\left(\left(\prod_{j=1}^{l}\left(1-\frac{2\theta_j p}{n}\right)\right)^{\frac{1}{l}}\right)\right\}\right\} \\
&= \exp\left\{\min_{p<0}\left\{\underbrace{-lp\gamma + lp\frac{\left(\prod_{j=1}^{l}\frac{1}{n}\|\mathbf{b}_j\|_2^2\right)^{\frac{1}{l}}}{\left(\prod_{j=1}^{l}(1-\frac{2\theta_j p}{n})\right)^{\frac{1}{l}}} - \frac{nl}{2}\log\left(\left(\prod_{j=1}^{l}\left(1-\frac{2\theta_j p}{n}\right)\right)^{\frac{1}{l}}\right)}_{\triangleq f(p)}\right\}\right\} \\
&= \exp\left\{\min_{p<0} f(p)\right\}. \qquad (2.94)
\end{aligned}
$$

where (2.93) follows from the fact that the arithmetic mean is no smaller than the geometric mean. On the other hand, define the function

$$g(p,\theta) = -lp\gamma + lp\frac{\left(\prod_{j=1}^{l}\frac{1}{n}\|\mathbf{b}_j\|_2^2\right)^{\frac{1}{l}}}{1-\frac{2\theta p}{n}} - \frac{nl}{2}\log\left(1-\frac{2\theta p}{n}\right). \tag{2.95}$$

Recall that $\left(\prod_{j=1}^{l}\frac{1}{n}\|\mathbf{b}_j\|_2^2\right)^{\frac{1}{l}} = \alpha$. It can be readily seen that, for a fixed $p < 0$,

$$f(p) \le \max_{\theta\ge 0} g(p,\theta) \tag{2.96}$$

which is because there exists $\theta \ge 0$ such that $1-\frac{2\theta p}{n} = \left(\prod_{j=1}^{l}\left(1-\frac{2\theta_j p}{n}\right)\right)^{\frac{1}{l}}$. Thus,

$$\min_{p<0} f(p) \le \min_{p<0}\left(\max_{\theta\ge 0} g(p,\theta)\right).$$

Our goal is to show

$$\min_{p<0}\left(\max_{\theta\ge 0} g(p,\theta)\right) = -\frac{nl}{2}\log\frac{\alpha}{\gamma}$$

which will lead to $\mathsf{P}(S_n \le \gamma) \le \exp(-\frac{nl}{2}\log\frac{\alpha}{\gamma})$ as desired. To this end, we first consider, for a fixed p,

$$\frac{\partial g(p,\theta)}{\partial\theta} = \frac{pl(\frac{2p\alpha}{n}-\frac{2p\theta}{n}+1)}{(1-\frac{2\theta p}{n})^2}.$$

By setting $\frac{\partial g(p,\theta)}{\partial\theta} = 0$, we have the only stationary point $\theta^* = \alpha + \frac{n}{2p}$. Examine the second derivative

$$\left.\frac{\partial^2 g(p,\theta)}{\partial\theta^2}\right|_{\theta=\theta^*} = \left.\frac{(1-\frac{2\theta p}{n})(-\frac{2p^2 l}{n})(\frac{2p\theta}{n}-\frac{4p\alpha}{n}-1)}{(1-\frac{2\theta p}{n})^4}\right|_{\theta=\theta^*} = \frac{(-\frac{2p^2 l}{n})}{(\frac{2\alpha p}{n})^2} < 0.$$

Due to the constraint $\theta \geq 0$, we have

$$\max_{\theta\geq 0} g(p,\theta) = \begin{cases} -pl\gamma - \frac{nl}{2} - \frac{nl}{2}\log(-\frac{2p\alpha}{n}) & \text{if } p \leq -\frac{n}{2\alpha}; \\ pl(\alpha-\gamma) & \text{if } -\frac{n}{2\alpha} \leq p < 0. \end{cases}$$

Next, we calculate $\min_{p<0}\left(\max_{\theta\geq 0} g(p,\theta)\right)$. First,

$$\min_{-\frac{n}{2\alpha}\leq p<0}\left(\max_{\theta\geq 0} g(p,\theta)\right) = \min_{-\frac{n}{2\alpha}\leq p<0} pl(\alpha-\gamma) = -\frac{nl}{2}\left(1-\frac{\gamma}{\alpha}\right).$$

Then, to figure out $\min_{p\leq -\frac{n}{2\alpha}}\left(\max_{\theta\geq 0} g(p,\theta)\right)$, we compute

$$\frac{\partial \max_{\theta>0} g(p,\theta)}{\partial p} = \frac{\partial\left(-pl\gamma - \frac{nl}{2} - \frac{nl}{2}\log(-\frac{2p\alpha}{n})\right)}{\partial p} = -l\gamma - \frac{nl}{2p} =_{\text{set}} 0$$

which gives the stationary point $p^* = -\frac{n}{2\gamma}$. Check for the second derivative,

$$\left.\frac{\partial^2 \max_{\theta\geq 0} g(p,\theta)}{\partial p^2}\right|_{p=p^*} = \frac{nl}{2(p^*)^2} > 0.$$

Therefore, $p^* = -\frac{n}{2\gamma}(\leq -\frac{n}{2\alpha})$ is the minimizer. As a result,

$$\min_{p\leq -\frac{n}{2\alpha}}\left(\max_{\theta\geq 0} g(p,\theta)\right) = -pl\gamma - \frac{nl}{2} - \frac{nl}{2}\log(-\frac{2p\alpha}{n})\bigg|_{p=p^*} = -\frac{nl}{2}\log\frac{\alpha}{\gamma}.$$

Overall,

$$\min_{p<0}\left(\max_{\theta\geq 0} g(p,\theta)\right) = \min\left(-\frac{nl}{2}\left(1-\frac{\gamma}{\alpha}\right), -\frac{nl}{2}\log\frac{\alpha}{\gamma}\right).$$

Using the fact that $0 \leq 1 - \frac{1}{x} \leq \log x$ for $x > 1$, we finally have

$$\min_{p<0}\left(\max_{\theta\geq 0} g(p,\theta)\right) = -\frac{nl}{2}\log\frac{\alpha}{\gamma}.$$

Therefore,

$$\begin{aligned}
\mathsf{P}(S_n \leq \gamma) &\leq \exp\left\{\min_{p<0} f(p)\right\} \\
&\leq \min_{p<0}\left(\max_{\theta\geq 0} g(p,\theta)\right) \\
&= 2^{-\frac{n}{2}\log\frac{\prod_{j=1}^{l}\frac{1}{n}\|\mathbf{b}_j\|_2^2}{\gamma^l}} \\
&= 2^{-\frac{n}{2}\log\frac{\prod_{j=1}^{l}[\frac{1}{n}B^\intercal B]_{j,j}}{\gamma^l}}.
\end{aligned}$$

The remaining task is to prove Lemma 4. Let $\sigma_{b,1}^2 \geq \cdots \geq \sigma_{b,q}^2$ be the q eigenvalues of $B^\intercal B$, where $\sigma_{b,q}^2 = \sigma_b^2$. The eigen-decomposition states that there exists a unitary matrix $J \in \mathbb{R}^{q\times q}$, such that $B^\intercal B = JGGJ^\intercal$, where $G \in \mathbb{R}^{q\times q}$ is a diagonal matrix with the ith diagonal element being $\sigma_{b,i}$. Thus, $D^\intercal B^\intercal BD = D^\intercal JGGJ^\intercal D = FT$, where $F = D^\intercal JG$ and $T = F^\intercal$. Note that

$$
\begin{aligned}
&\det((BD)^\intercal BD) \\
&= \det(FT) \\
&= \sum_{1\le j_1<\cdots<j_r\le q} \det\begin{bmatrix} f_{1,j_1} & \cdots & f_{1,j_r} \\ \vdots & & \vdots \\ f_{r,j_1} & \cdots & f_{r,j_r} \end{bmatrix} \det\begin{bmatrix} t_{j_1,1} & \cdots & t_{j_1,r} \\ \vdots & & \vdots \\ t_{k_r,1} & \cdots & t_{j_r,r} \end{bmatrix} \qquad (2.97) \\
&= \sum_{1\le j_1<\cdots<j_r\le q} \left(\det\begin{bmatrix} f_{1,j_1} & \cdots & f_{1,j_r} \\ \vdots & & \vdots \\ f_{r,j_1} & \cdots & f_{r,j_r} \end{bmatrix} \right)^2 \\
&= \sum_{1\le j_1<\cdots<j_r\le q} \left(\det\left\{ \begin{bmatrix} [D^\intercal J]_{1,j_1} & \cdots & [D^\intercal J]_{1,j_r} \\ \vdots & & \vdots \\ [D^\intercal J]_{r,j_1} & \cdots & [D^\intercal J]_{r,j_r} \end{bmatrix} \operatorname{diag}(\sigma_{b,j_1}, ..., \sigma_{b,j_r}) \right\} \right)^2 \\
&\ge (\sigma_b^2)^r \sum_{1\le j_1<\cdots<j_r\le q} \left(\det\begin{bmatrix} [D^\intercal J]_{1,j_1} & \cdots & [D^\intercal J]_{1,j_r} \\ \vdots & & \vdots \\ [D^\intercal J]_{r,j_1} & \cdots & [D^\intercal J]_{r,j_r} \end{bmatrix} \right)^2 \\
&= (\sigma_b^2)^r \det(D^\intercal J^\intercal JD) \\
&= (\sigma_b^2)^r \det(D^\intercal D)
\end{aligned}
$$

where (2.97) is due to the Binet-Cauchy formula [42].

2.8.6 Proof of Theorem 6

To establish this theorem, we prove the following equivalent statement:

If there exist a sequence of matrices $\{A^{(m)}\}_{m=k}^{\infty}$, $A^{(m)} \in \mathbb{R}^{n_m \times m}$, and a sequence of support recovery maps $\{d^{(m)}\}_{m=k}^{\infty}$, $d^{(m)} : \mathbb{R}^{n_m} \mapsto 2^{\{1,2,\ldots,m\}}$, such that

$$\frac{1}{n_m m}\|A^{(m)}\|_F^2 \leq \sigma_a^2$$

and

$$\lim_{m\to\infty} \mathsf{P}\{d^{(m)}(A^{(m)}X + Z) \neq \text{supp}(X(W, \mathbf{S}))\} = 0$$

then

$$\limsup_{m\to\infty} \frac{\log m}{n_m} \leq c(W).$$

For any $\mathcal{T} \subseteq [k]$, denote the tuple of random variables $(S_l : l \in \mathcal{T})$ by $S(\mathcal{T})$. For notation simplicity, let $\overline{P}_e^{(m)} \triangleq \mathsf{P}\{d^{(m)}(A^{(m)}X + Z) \neq \text{supp}(X(W, \mathbf{S}))\}$. From Fano's inequality [32], we have

$$\begin{aligned} H(S(\mathcal{T})|Y) &\leq H(S_1, ..., S_k|Y) \\ &\leq \log k! + H(\{S_1, ..., S_k\}|Y) \\ &\leq \log k! + \overline{P}_e \log \binom{m}{k} + 1. \end{aligned} \tag{2.98}$$

On the other hand,

$$
\begin{aligned}
H(S(\mathcal{T})|S(\mathcal{T}^c)) &= \log\left(\prod_{q=0}^{|\mathcal{T}|-1}(m-(k-|\mathcal{T}|)-q)\right) \\
&= |\mathcal{T}|\log m - n\epsilon_{1,n}
\end{aligned}
\tag{2.99}
$$

where $\mathcal{T}^c \triangleq [k]\backslash\mathcal{T}$ and

$$
\epsilon_{1,n} \triangleq \frac{1}{n}\log\left(m^{|\mathcal{T}|}/\prod_{q=0}^{|\mathcal{T}|-1}(m-(k-|\mathcal{T}|)-q)\right)
$$

which tends to zero as $n \to \infty$. Hence, combining (2.98) and (2.99), we have

$$
\begin{aligned}
&|\mathcal{T}| \log m \\
&= H(S(\mathcal{T})|S(\mathcal{T}^c)) + n\epsilon_{1,n} \\
&= I(S(\mathcal{T}); Y|S(\mathcal{T}^c)) + H(S(\mathcal{T})|Y, S(\mathcal{T}^c)) + n\epsilon_{1,n} \\
&\leq I(S(\mathcal{T}); Y|S(\mathcal{T}^c)) + H(S(\mathcal{T})|Y) + n\epsilon_{1,n} \qquad (2.100) \\
&\leq I(S(\mathcal{T}); Y|S(\mathcal{T}^c)) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&= \sum_{i=1}^{n} I(\underline{\mathbf{Y}}_i; S(\mathcal{T})|\underline{Y}_{[i-1]}, S(\mathcal{T}^c)) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&= \sum_{i=1}^{n} \big(h(\underline{\mathbf{Y}}_i|\underline{Y}_{[i-1]}, S(\mathcal{T}^c)) - h(\underline{\mathbf{Y}}_i|\underline{Y}_{[i-1]}, S([k]))\big) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} \\
&\qquad + 1 + n\epsilon_{1,n} \\
&\leq \sum_{i=1}^{n} (h(\underline{\mathbf{Y}}_i|S(\mathcal{T}^c)) - h(\underline{\mathbf{Y}}_i|S_1, ..., S_k)) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&\qquad (2.101) \\
&= \sum_{i=1}^{n} (h(\underline{\mathbf{Y}}_i|S(\mathcal{T}^c)) - h(\underline{\mathbf{Z}}_i)) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \qquad (2.102)
\end{aligned}
$$

where $\underline{Y}_{[i-1]}$ denotes the set $\{\underline{\mathbf{Y}}_1, ..., \underline{\mathbf{Y}}_{i-1}\}$. To explain some intermediate steps, (2.100) follows from the fact that conditioning reduces entropy, (2.101) holds because $\underline{\mathbf{Y}}_i$ is independent of $\underline{Y}_{[i-1]}$ when conditioned on $S([k])$, and (2.102) follows since the measurement matrix is fixed and $\underline{\mathbf{Z}}_i$ is independent of $(S_1, \ldots, S_k)$.

Consider

$$
\begin{aligned}
&h(\underline{\mathbf{Y}}_i|S(\mathcal{T}^c)) \\
&= h\left(A_{i,S([k])}W + \underline{\mathbf{Z}}_i \middle| S(\mathcal{T}^c)\right) \\
&= h\left(A_{i,S(\mathcal{T})}\underline{W}_{\mathcal{T}} + \underline{\mathbf{Z}}_i \middle| S(\mathcal{T}^c)\right) \\
&\leq h\left(A_{i,S(\mathcal{T})}\underline{W}_{\mathcal{T}} + \underline{\mathbf{Z}}_i\right) \\
&\leq \frac{1}{2}\log\left((2\pi e)^l \cdot \det\left(\mathsf{E}[(A_{i,S(\mathcal{T})}\underline{W}_{\mathcal{T}} + \underline{\mathbf{Z}}_i)^\intercal(A_{i,S(\mathcal{T})}\underline{W}_{\mathcal{T}} + \underline{\mathbf{Z}}_i)]\right.\right. \\
&\qquad\qquad\qquad \left.\left. - \mathsf{E}[A_{i,S(\mathcal{T})}\underline{W}_{\mathcal{T}} + \underline{\mathbf{Z}}_i]^\intercal \, \mathsf{E}[A_{i,S(\mathcal{T})}\underline{W}_{\mathcal{T}} + \underline{\mathbf{Z}}_i]\right)\right) \qquad (2.103) \\
&\leq \frac{1}{2}\log\left((2\pi e)^l \cdot \det\left(\underline{W}_{\mathcal{T}}^\intercal\left(\mathsf{E}[A_{i,S(\mathcal{T})}^\intercal A_{i,S(\mathcal{T})}] - \mathsf{E}[A_{i,S(\mathcal{T})}]^\intercal \, \mathsf{E}[A_{i,S(\mathcal{T})}]\right)\underline{W}_{\mathcal{T}} + \sigma_z^2 I\right)\right) \\
&\qquad (2.104)
\end{aligned}
$$

where (2.103) follows from the fact that with the same covariance the Gaussian random vector maximizes the entropy [32], and the randomness in $A_{i,S(\mathcal{T})}$ is due to the randomness of the index set $S(\mathcal{T})$. Note that

$$
\mathsf{E}[A_{i,S(\mathcal{T})}] = \frac{1}{m}\sum_{p=1}^{m} a_{i,p}\mathbf{1}^\intercal. \qquad (2.105)
$$

Meanwhile

$$
\mathsf{E}[A_{i,S(\mathcal{T})}^\intercal A_{i,S(\mathcal{T})}] = \frac{1}{m}\sum_{p=1}^{m} a_{i,p}^2 I + \frac{1}{m(m-1)}\sum_{p=1}^{m}\sum_{\substack{q=1 \\ q\neq p}}^{m} a_{i,p}a_{i,q}(\mathbf{1}\cdot\mathbf{1}^\intercal - I). \qquad (2.106)
$$

Thus

$$
\begin{aligned}
&\mathsf{E}[A_{i,S(\mathcal{T})}^\intercal A_{i,S(\mathcal{T})}] - \mathsf{E}[A_{i,S(\mathcal{T})}]^\intercal \,\mathsf{E}[A_{i,S(\mathcal{T})}] \\
&= \frac{1}{m}\sum_{p=1}^{m} a_{i,p}^2 I + \frac{1}{m(m-1)} \sum_{p=1}^{m} \sum_{\substack{q=1 \\ q\neq p}}^{m} a_{i,p} a_{i,q} (\mathbf{1}\cdot\mathbf{1}^\intercal - I) - \frac{1}{m^2}\left(\sum_{p=1}^{m} a_{i,p}\right)^2 \mathbf{1}\cdot\mathbf{1}^\intercal \\
&= \frac{1}{m}\sum_{p=1}^{m} a_{i,p}^2 I + \frac{1}{m(m-1)} \left(\left(\sum_{p=1}^{m} a_{i,p}\right)^2 - \sum_{p=1}^{m} a_{i,p}^2\right)(\mathbf{1}\cdot\mathbf{1}^\intercal - I) \\
&\qquad - \frac{1}{m^2}\left(\sum_{p=1}^{m} a_{i,p}\right)^2 \mathbf{1}\cdot\mathbf{1}^\intercal \\
&= \frac{1}{m}\sum_{p=1}^{m} a_{i,p}^2 \left(I - \frac{1}{m-1}(\mathbf{1}\cdot\mathbf{1}^\intercal - I)\right) \\
&\qquad + \left(\sum_{p=1}^{m} a_{i,p}\right)^2 \left(\frac{1}{m(m-1)}(\mathbf{1}\cdot\mathbf{1}^\intercal - I) - \frac{1}{m^2}\mathbf{1}\cdot\mathbf{1}^\intercal\right).
\end{aligned}
\tag{2.107}
$$

Note that $\frac{1}{m(m-1)}(\mathbf{1}\cdot\mathbf{1}^\intercal - I) - \frac{1}{m^2}\mathbf{1}\cdot\mathbf{1}^\intercal = \frac{1}{m^2(m-1)}\mathbf{1}\cdot\mathbf{1}^\intercal - \frac{1}{m(m-1)} I$ is negative semidefinite for sufficiently large m, and so is $\underline{W}_{\mathcal{T}}^\intercal \left(\frac{1}{m^2(m-1)}\mathbf{1}\cdot\mathbf{1}^\intercal - \frac{1}{m(m-1)} I\right) \underline{W}_{\mathcal{T}}$. Hence

$$
\begin{aligned}
&\det\left(\underline{W}_{\mathcal{T}}^\intercal \left(\mathsf{E}[A_{i,S(\mathcal{T})}^\intercal A_{i,S(\mathcal{T})}] - \mathsf{E}[A_{i,S(\mathcal{T})}]^\intercal \,\mathsf{E}[A_{i,S(\mathcal{T})}]\right) \underline{W}_{\mathcal{T}} + \sigma_z^2 I\right) \\
&\leq \det\left(\frac{1}{m}\sum_{p=1}^{m} a_{i,p}^2 \underline{W}_{\mathcal{T}}^\intercal \left(I - \frac{1}{m-1}(\mathbf{1}\cdot\mathbf{1}^\intercal - I)\right) \underline{W}_{\mathcal{T}} + \sigma_z^2 I\right)
\end{aligned}
$$

as a result of [8, Corollary 8.4.15]. Therefore

$$
\begin{aligned}
&|\mathcal{T}| \log m \\
&\leq \sum_{i=1}^{n} \left[\frac{1}{2} \log \left((2\pi e)^l \cdot \det \left(\frac{1}{m} \sum_{p=1}^{m} a_{i,p}^2 \underline{W}_{\mathcal{T}}^{\intercal} \left(I - \frac{1}{m-1} (\mathbf{1} \cdot \mathbf{1}^{\intercal} - I) \right) \underline{W}_{\mathcal{T}} + \sigma_z^2 I \right) \right) \right. \\
&\qquad \left. - \frac{1}{2} \log \left((2\pi e \sigma_z^2)^l \right) \right] + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&= \sum_{i=1}^{n} \frac{1}{2} \log \det \left(\frac{1}{m\sigma_z^2} \sum_{p=1}^{m} a_{i,p}^2 \underline{W}_{\mathcal{T}}^{\intercal} \left(I - \frac{1}{m-1} (\mathbf{1} \cdot \mathbf{1}^{\intercal} - I) \right) \underline{W}_{\mathcal{T}} + I \right) \\
&\qquad + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&\leq \frac{n}{2} \log \det \left(\frac{1}{nm\sigma_z^2} \sum_{i=1}^{n} \sum_{p=1}^{m} a_{i,p}^2 \underline{W}_{\mathcal{T}}^{\intercal} \left(I - \frac{1}{m-1} (\mathbf{1} \cdot \mathbf{1}^{\intercal} - I) \right) \underline{W}_{\mathcal{T}} + I \right) \\
&\qquad + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} + 1 + n\epsilon_{1,n} \\
&\leq \frac{n}{2} \log \det \left(\frac{\sigma_a^2}{\sigma_z^2} \underline{W}_{\mathcal{T}}^{\intercal} \left(I - \frac{1}{m-1} (\mathbf{1} \cdot \mathbf{1}^{\intercal} - I) \right) \underline{W}_{\mathcal{T}} + I \right) + \log k! + \overline{P}_e^{(m)} \log \binom{m}{k} \\
&\qquad + 1 + n\epsilon_{1,n} \\
&\leq \frac{n}{2} \log \det \left(\frac{\sigma_a^2}{\sigma_z^2} \underline{W}_{\mathcal{T}}^{\intercal} \left(I - \frac{1}{m-1} (\mathbf{1} \cdot \mathbf{1}^{\intercal} - I) \right) \underline{W}_{\mathcal{T}} + I \right) + \log k! + \overline{P}_e^{(m)} k \log m \\
&\qquad + 1 + n\epsilon_{1,n}.
\end{aligned}
\tag{2.108}
$$

Then, we have

$$\begin{aligned}
&\limsup_{m\to\infty} \frac{(1-k\overline{P}_e^{(m)}/|\mathcal{T}|)\log m}{n_m} - \frac{\log k! + n_m\epsilon_{1,n}+1}{|\mathcal{T}|n_m} \\
&\le \limsup_{m\to\infty} \frac{1}{2|\mathcal{T}|}\log\det\left(\frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\left(I-\frac{1}{m-1}(\mathbf{1}\cdot\mathbf{1}^{\intercal}-I)\right)\underline{W}_{\mathcal{T}}+I\right) \\
&= \frac{1}{2|\mathcal{T}|}\log\det\left(\frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}+I\right)
\end{aligned} \tag{2.109}$$

for all $\mathcal{T}\subseteq[k]$. Since $\lim_{m\to\infty}\overline{P}_e^{(m)}=0$, we reach the conclusion

$$\limsup_{m\to\infty}\frac{\log m}{n_m}\le\frac{1}{2|\mathcal{T}|}\log\det\left(\frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}^{\intercal}\underline{W}_{\mathcal{T}}+I\right) \tag{2.110}$$

for all $\mathcal{T}\subseteq[k]$. This completes the proof of Theorem 6.

2.8.7 Proof of Corollary 4

To justify this corollary, we need to show

$$\min_{\mathcal{T}\subseteq[k]}\left[\frac{1}{2|\mathcal{T}|}\log\det\left(I+\frac{\sigma_a^2}{\sigma_z^2}\underline{W}_{\mathcal{T}}{}^{\intercal}\underline{W}_{\mathcal{T}}\right)\right]=2\cdot\frac{1}{2k}\log\left(1+k\cdot\frac{\sigma_a^2}{\sigma_z^2}\right).$$

To begin with, recall that k is even, and $\mathbf{w}_2$ is defined in (2.24). For a given $\mathcal{T}\subseteq[k]$, let $\mathcal{T}_1=\mathcal{T}\cap[\frac{k}{2}]$, $\mathcal{T}_2=\mathcal{T}\backslash\mathcal{T}_1$, $t=|\mathcal{T}|$, $t_1=|\mathcal{T}_1|$, and $t_2=|\mathcal{T}_2|$. One can obtain

$$\underline{W}_{\mathcal{T}}{}^{\intercal}\underline{W}_{\mathcal{T}}=\begin{bmatrix} t & t_1-t_2 \\ t_1-t_2 & t\end{bmatrix}.$$

Let $\alpha \triangleq \frac{\sigma_a^2}{\sigma_z^2}$ for notational simplicity. Thus

$$\frac{1}{2|\mathcal{T}|} \log\det\left(I + \alpha \underline{W}_{\mathcal{T}}{}^{\intercal} \underline{W}_{\mathcal{T}}\right) = \frac{1}{2t} \log\det \begin{bmatrix} 1+\alpha t & \alpha(t_1 - t_2) \\ \alpha(t_1 - t_2) & 1+\alpha t \end{bmatrix}$$
$$= \frac{1}{2t} \log\left(1 + 2\alpha t + 4\alpha^2 t_1 t_2\right) \tag{2.111}$$

where we use the fact that $t = t_1 + t_2$. Note that, for a given $t \in [k]$,

$$\min_{\mathcal{T}:\mathcal{T}\subseteq[k],|\mathcal{T}|=t\leq\frac{k}{2}} \frac{1}{2t} \log\left(1 + 2\alpha t + 4\alpha^2 t_1 t_2\right) = \frac{1}{2t} \log(1 + 2\alpha t) \tag{2.112}$$

and

$$\min_{\mathcal{T}:\mathcal{T}\subseteq[k],|\mathcal{T}|=t>\frac{k}{2}} \frac{1}{2t} \log\left(1 + 2\alpha t + 4\alpha^2 t_1 t_2\right) = \frac{1}{2t} \log\left(1 + 2\alpha t + 4\alpha^2 \frac{k}{2}\left(t - \frac{k}{2}\right)\right) \tag{2.113}$$

where we use the implicit constraints $t_1, t_2 \leq \frac{k}{2}$. Then, the problem becomes evaluating

$$\min_{t:t\in[k]} f(t), \text{ where } f(t) = \begin{cases} \frac{1}{2t}\log(1+2\alpha t) & \text{if } 0 < t \leq \frac{k}{2}, \\ \frac{1}{2t}\log\left(1 + 2\alpha t + 4\alpha^2\frac{k}{2}\left(t - \frac{k}{2}\right)\right) & \text{if } \frac{k}{2} + 1 \leq t \leq k. \end{cases} \tag{2.114}$$

First, it can be readily seen that $\min_{t:t\in[\frac{k}{2}]} f(t) = \frac{1}{k}\log(1+\alpha k)$. Next, we consider the function

$$g(t) \triangleq \frac{\log(\beta_1 + \beta_2 t)}{2t}$$

where $\beta_1 \triangleq 1 - \alpha^2 k^2$ and $\beta_2 \triangleq 2\alpha(1 + \alpha k)$ for $t \in [\frac{k}{2}, k]$. Note that[8]

$$\frac{\partial g(t)}{\partial t} = \frac{1 - \frac{\beta_1}{\beta_1 + \beta_2 t} - \log(\beta_1 + \beta_2 t)}{t^2}. \tag{2.115}$$

To obtain stationary points, we solve

$$1 - \frac{\beta_1}{\beta_1 + \beta_2 t} + \log \frac{1}{\beta_1 + \beta_2 t} = 0,\ t \neq 0 \tag{2.116}$$

which is equivalent to

$$1 + v(t) = \beta_1 e^{v(t)},\ t \neq 0 \tag{2.117}$$

where $v(t) \triangleq \log \frac{1}{\beta_1 + \beta_2 t}$. Note that $\beta_1 < 1$. We will consider three different cases. The first case is $0 < \beta_1 < 1$. By comparing the curves of $1 + v$ and $\beta_1 e^v$ as functions of v, we see that there are two solutions with opposite signs, namely $v_1 < 0$ and $v_2 > 0$, to (2.117). Note that

$$g\left(\frac{k}{2}\right) = g(k) = \frac{1}{k} \log(1 + \alpha k).$$

[8] For the purpose of analysis, the base of logarithm is not important, as long as all of them are consistent. Here, we choose natural logarithm to simplify the calculation.

Meanwhile, $v(t)$ is monotonically decreasing on $[\frac{k}{2}, k]$, and

$$v(k) = \log \frac{1}{(1+\alpha k)^2} < v\left(\frac{k}{2}\right) = \log \frac{1}{1+\alpha k} < 0.$$

Therefore, it is evident that $v(k) < v_1 < v\left(\frac{k}{2}\right) < v_2$. Further, it can be readily seen that

$$\left.\frac{\partial g(t)}{\partial t}\right|_{t=\frac{k}{2}} = \left.\frac{1+v(t)-\beta_1 e^{v(t)}}{t^2}\right|_{t=\frac{k}{2}} > 0$$
$$\left.\frac{\partial g(t)}{\partial t}\right|_{t=k} = \left.\frac{1+v(t)-\beta_1 e^{v(t)}}{t^2}\right|_{t=k} < 0.$$

In summary, $g(t)$ is increasing at $t = \frac{k}{2}$ and decreasing at $t = k$, it takes the same value at these two points, and there exists only one stationary point in between. These observations lead to the conclusion that $\min_{t:t\in[k]\setminus[\frac{k}{2}]} f(t) = f(k) = \frac{1}{k}\log(1+\alpha k)$.

To analyze the cases for $\beta_1 = 0$ and $\beta_1 < 0$, we only need to note that there is only one solution v_1 to (2.117). Thus, similar argument applies to these two cases.

Chapter 3

Algorithm Design and Analysis Using Multiuser Information Theoretic Perspectives

The previous chapter discussed the usage of multiuser information theory in deriving the performance limits of an optimal approach for support recovery of sparse signals. The results therein indicate the best possible performance that any algorithm can achieve. In this chapter, we turn our attention to practical algorithms. We demonstrate that multiuser information theoretic tools can be also applied to motivate and analyze practical algorithms for sparse signal recovery. To this end, we first introduce the MultiPass algorithmic framework which is inspired by the group detector in multiuser detection. Under this framework, the MultiPass Lasso and Reweighted MultiPass Lasso algorithms are studied in detail. Next, we establish a connection between orthogonal matching pursuit (OMP) and successive interference cancellation. An intuitive criterion for the success of OMP is discussed, and experiments are performed to support the utility of the analysis.

3.1 Introduction

Here we again consider the estimation of a sparse signal through its linear measurements in the presence of measurement noise, namely based on the model

$$\mathbf{Y} = A\mathbf{x} + \mathbf{N}$$

where $\mathbf{x} \in \mathbb{R}^m$ is the signal of interest, $A \in \mathbb{R}^{n \times m}$ is the measurement matrix, $\mathbf{N} \in \mathbb{R}^n$ is the measurement noise, and $\mathbf{Y} \in \mathbb{R}^n$ is the noisy measurement. We assume $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. Denote by k the number of nonzero entries in $\mathbf{x}$, i.e., $k = |\text{supp}(\mathbf{x})|$, where $\text{supp}(\mathbf{x})$ denotes the support of $\mathbf{x}$. Then, $\mathbf{x}$ is said to be sparse when $k \ll m$. Given the measurement $\mathbf{Y}$ and the measurement matrix A, the goal is to reconstruct the sparse signal $\mathbf{x}$.

Among existing algorithms for reconstructing sparse solutions, one can broadly classify them into two categories according to their underlying principles. The first class of algorithms employ a greedy search approach and the locations of the nonzero entries in $\mathbf{x}$ are sequentially determined via a number of iterations. At each iteration, the algorithm finds the columns of A that best correlate with the current residual signal and then removes their contributions from the current residual signal to form the new residual signal for the next iteration. The algorithm terminates employing a stopping criterion such as the number of iterations or the strength of the residual signal, among others. Algorithms related to this principle include matching pursuit [82], orthogonal matching pursuit [94], stagewise orthogonal matching pursuit [37], and subspace pursuit [34]. The second class of algorithms involve solving an optimization problem with a carefully chosen cost function to which the minimizers are considered reasonable estimates of the sparse signals of interest. As opposed to the class of sequential selection algorithms above, an algorithm of the second class jointly estimates all the nonzero entries. Basis pursuit (BP) [24], FOCUSS [60], Lasso [113], and reweighted ℓ_1 minimization [20] are examples of this latter type of joint recovery algorithms. Especially, with important relevance to our work, the Lasso algorithm solves for

$$\mathbf{X}_{\text{Lasso}} = \arg\min_{\tilde{\mathbf{x}} \in \mathbb{R}^m} \frac{1}{2n} \|\mathbf{Y} - A\tilde{\mathbf{x}}\|_2^2 + \lambda \|\tilde{\mathbf{x}}\|_1 \tag{3.1}$$

where $\lambda \geq 0$ is the regularization parameter. Note that (3.1) is convex in nature and can be solved by many existing convex optimization routines [48, 63, 73].

3.2 The MultiPass Algorithmic Framework and the MultiPass Lasso Algorithms

3.2.1 Motivation of the MultiPass Algorithmic Framework

According to the theoretical analysis for basis pursuit and Lasso [14, 81], they are capable of dealing well with signals with similar nonzero magnitudes but they do not exploit the variation in the dynamic range of the nonzero magnitudes. Sequential selection methods perform notably better when such variation in nonzero entries exists, especially when the nonzero magnitudes form an exponentially decaying profile [52], but they perform poorly when the magnitudes are similar. Note that practical signals usually exhibit neither an ideal flat profile nor an ideal exponentially decaying profile. Instead, we model the nonzero entries as clusters, where each cluster comprises of a group of nonzero entries with comparable magnitudes, and magnitudes across different clusters are disparate. This character of practical signals poses a new challenge to the algorithm design for sparse signal recovery.

According to the information theoretic perspective introduced in Section 2.3, a nonzero entry can be viewed as a sender and the measurement matrix as the codebook. A sequential selection method for sparse signal recovery can be viewed as a successive interference cancellation scheme for multiple user detection, and methods such as Lasso as joint detection schemes [118]. Inspired by the group detectors [65, 77, 119] in multiuser communication, we suggest that different clusters be identified in a sequential manner whereas the nonzero entries within a cluster be detected jointly. This gives rise to the MultiPass algorithmic framework, which tries to make the best of both sequential selection methods and joint recovery methods. Under this framework, a joint recovery method is applied sequentially over multiple iterations (i.e., passes), and at each iteration this joint recovery method aims to detect a subset of the support on which the nonzero entries have similar magnitudes.

To derive a concrete algorithm under the MultiPass algorithmic framework, one needs to choose a proper joint recovery method for the iterations. Next, let us study a specific instantiation of this algorithmic framework in detail.

3.2.2 The MultiPass Lasso Algorithm (MPL)

We propose the MultiPass Lasso algorithm, which employs the Lasso algorithm for each iteration. The reason for this choice is that we are inspired by recent theoretical analysis of Lasso unveiled in [44, 123]. On one hand, with a larger regularization parameter λ, in general the solution vector $\mathbf{X}_{\text{Lasso}}$ tends to be sparser. Meanwhile, a larger λ increases the probability of the support of the reconstructed signal being a subset of the true support. Based on these facts, the MultiPass Lasso algorithm suitably reflects the design idea that at each iteration a partial support can be correctly recovered with high probability so that the union of these sequentially recovered partial supports has a good chance of being the true support.

The MultiPass Lasso algorithm is described as follows.

Step 1: Let $\mathbf{Y}^{(0)} = \mathbf{Y}$, $\mathcal{S}^{(0)} = \emptyset$, $A^{(0)} = A$, and $l = 1$.

Step 2: At the lth iteration. Choose a regularization parameter $\lambda^{(l)}$ within the range $[0, \frac{1}{n}\|A^{(l-1)\intercal}\mathbf{Y}^{(l-1)}\|_\infty)$. Solve for

$$\mathbf{X}^{(l)} = \arg\min_{\tilde{\mathbf{x}}} \frac{1}{2n}\|\mathbf{Y}^{(l-1)} - A^{(l-1)}\tilde{\mathbf{x}}\|_2^2 + \lambda^{(l)}\|\tilde{\mathbf{x}}\|_1.$$

Let

$$\mathcal{S}^{(l)} = \mathcal{S}^{(l-1)} \cup \text{supp}(\mathbf{X}^{(l)}),\ \mathbf{Y}^{(l)} = P^{\perp}_{\mathcal{S}^{(l)}}\mathbf{Y}^{(l-1)},\ A^{(l)} = P^{\perp}_{\mathcal{S}^{(l)}}A^{(l-1)}$$

where $P^{\perp}_{\mathcal{S}} \triangleq I - A_{\mathcal{S}}(A^{\intercal}_{\mathcal{S}}A_{\mathcal{S}})^{-1}A^{\intercal}_{\mathcal{S}}$.

Step 3: If the predefined termination condition is satisfied, output the final estimate of the sparse signal as $\mathbf{X}_{\text{MPL}} = A^{\dagger}_{\mathcal{S}^{(l)}}\mathbf{Y}$. Otherwise, set $l \to l + 1$ and go to Step 2.

3.2.3 Observations on MPL

Selection of Regularization Parameter $\lambda^{(l)}$

At the lth iteration, MPL solves for a Lasso solution based on the current residual signal $\mathbf{Y}^{(l-1)}$ and the current measurement matrix $A^{(l-1)}$. According to the discussion in [90], generally, $\mathbf{X}^{(l)} \neq \mathbf{0}$ if and only if $\lambda^{(l)} \in [0, \frac{1}{n}\|A^{(l-1)\intercal}\mathbf{Y}^{(l-1)}\|_\infty)$. Meanwhile, based on the analysis in [44], as $\lambda^{(l)} \to 0_+$, in general more nonzero entries will be included in $\mathbf{X}^{(l)}$. One feasible way in implementation is to use the form $\lambda^{(l)} = \frac{\gamma}{n}\|A^{(l)\intercal}\mathbf{Y}^{(l)}\|_\infty$ with some fixed $\gamma \in (0, 1)$ for all iterations. Our experiences indicate that $\gamma \in (0.4, 0.9)$ typically gives good results. With larger γ, the solution in each iteration tends to be sparser, and generally more iterations will be carried out.

Termination Condition

We propose two possible criteria for algorithm termination, which are inspired by matching pursuit algorithms [31, 82, 94]. In the first criterion, we choose some $\delta > 0$ such that the algorithm stops if $\frac{1}{n}\|\mathbf{Y}^{(l)}\|_2^2 \le \delta$. When the measurement noise is absent, i.e., $\mathbf{N} = \mathbf{0}$, δ can be chosen as a small positive quantity related to machine precision. When the measurement is noisy, δ should be chosen based on the noise variance σ^2. As the second termination criterion, a threshold $s_{\text{max}} \in \mathbb{N}$ is set, and the algorithm terminates if $|\mathcal{S}^{(l)}| \ge s_{\text{max}}$.

Update of Measurement Matrix $A^{(l)}$

The update for $A^{(l)}$ in Step 2 is also employed in Order Recursive Matching Pursuit [31] for removing the contribution from previously selected columns. Note that the columns of $A^{(l)}$ indexed by $\mathcal{S}^{(l)}$ are actually zero vectors due to the orthogonal projection $P^{\perp}_{\mathcal{S}^{(l)}}$. For purpose of implementation, one can use the submatrix of $A^{(l)}$ by removing all the zero columns and properly re-indexing the remaining columns.

3.2.4 Preliminary Analysis of MPL

Recent theoretical work on Lasso [123] characterizes the probability lower bound for the event that the support of the recovered signal being a subset of the true support of the sparse signal. We can leverage the analytical techniques therein with necessary modifications to accommodate the dependencies among iterations to demonstrate the performance improvement enabled by MPL.

For ease of exposition, let us consider a simple but representative scenario, where the nonzero entries of a sparse signal can only take one of two possible values. Formally, let $\mathcal{S}_h, \mathcal{S}_l \subset [m]$ satisfying $\mathcal{S}_h \cap \mathcal{S}_l = \emptyset, |\mathcal{S}_h| = k_h, |\mathcal{S}_l| = k_l$. For $0 < x_{\text{low}} \leq x_{\text{high}}$, the signal $\mathbf{x}$ is given as $x_i = x_{\text{high}}$ if $i \in \mathcal{S}_h$; $x_i = x_{\text{low}}$ if $i \in \mathcal{S}_l$; and $x_i = 0$ otherwise. Suppose the elements of the measurement matrix A is independently generated according to $\mathcal{N}(0, 1)$. Let

$$\lambda_1 \triangleq \frac{x_{\text{high}}}{8\sqrt{\frac{k}{n}} + \sqrt{\frac{128k\log(m-k)}{n}} + \sqrt{18} + 1}$$
$$\lambda_2 \triangleq \frac{x_{\text{low}}}{\frac{n}{n-k}\left(8\left(\frac{k_l}{n-k_h}\right)^{1/4} + \sqrt{128} + 1\right) + \sqrt{18}}.$$

Fix some $\delta > \sigma^2$. Let $\rho > 0$ be an arbitrarily small constant. For tractable analysis, we only employ the first termination criterion discussed above.

For the first iteration of MPL, we choose the regularization parameter as $\lambda^{(1)} = \min(\lambda_1, \frac{1}{n}\|A^\intercal \mathbf{Y}\|_\infty) - \rho$. If the algorithm proceeds to the second iteration, we choose $\lambda^{(2)} = \min(\lambda_2, \frac{1}{n}\|A^{(1)\intercal}\mathbf{Y}^{(1)}\|_\infty) - \rho$. For any possible further iteration l, we pick an arbitrary $\lambda^{(l)} \in (0, \frac{1}{n}\|A^{(l-1)\intercal}\mathbf{Y}^{(l-1)}\|_\infty)$. Then, we can actually compute a lower bound for $\mathsf{P}(\text{supp}(\mathbf{X}_{\text{MPL}}) = \text{supp}(\mathbf{x}))$, the derivation of which is presented in Section 3.6.2. Due to the complex nature of this lower bound, we visualize it as well as the probability lower bound for support recovery by Lasso [123] in different ways in Figure 3.1 to compare the performance guarantees offered by different algorithms.

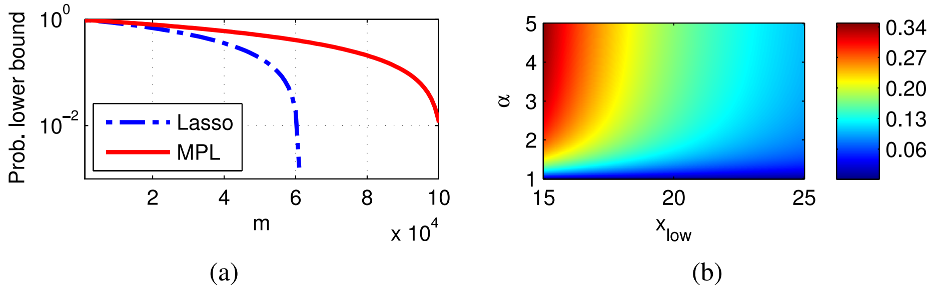


Figure 3.1: Comparisons of performance lower bounds for Lasso and MPL. Parameters: (a) $n = 800, \sigma^2 = 1, \delta = 3, x_{\text{high}} = 20, k_h = 10, x_{\text{low}} = 10, k_l = 5$. (b) $n = 800, m = 10^5, x_{\text{high}} = \alpha x_{\text{low}}, \sigma^2 = 1, \delta = 3, k_h = 10, k_l = 5$.

First, we examine the impact of m by the simulation in Figure 3.1(a). Note that, for large m (i.e., more possible locations to monitor), MPL provides better performance guarantee than Lasso while holding other parameters fixed. Next, we study the impact of the dynamic range of nonzero magnitudes on the reconstruction performance in Figure 3.1(b). Note that each point (x_{low}, α) corresponds to the nonzero signal value pair $(x_{\text{low}}, x_{\text{high}})$ where $x_{\text{high}} = \alpha x_{\text{low}}$. The color of a point indicates the difference between the probability lower bound for MPL and that of Lasso. As we can see, for a given the noise level, when x_{low} is relatively small, a nontrivial distance between x_{high} and x_{low} (i.e., a large dynamic range) enables MPL to enjoy better performance guarantee.

3.2.5 The Reweighted MultiPass Lasso Algorithm (RMPL)

The MPL algorithm enables the opportunity of substituting the Lasso routine in existing algorithms for obtaining their alternative MPL versions. Inspired by reweighted ℓ_1 minimization [20], we develop the reweighted MultiPass Lasso algorithm as follows.

Step 1: Set $q = 1$, $w_i^{(1)} = 1$ for $i \in [m]$. Choose $\epsilon > 0$, $q_{\text{max}} \in \mathbb{N}$.

Step 2: At the qth iteration. Run MultiPass Lasso based on the modified measurement matrix $A \cdot \text{diag}((w_i^{(q)})^{-1})$ and the measurement $\mathbf{Y}$, and obtain the output as $\mathbf{Z}^{(q)}$. For $i \in [m]$, compute

$$x_i^{(q)} = (w_i^{(q)})^{-1} z_i^{(q)}, \quad w_i^{(q+1)} = (|x_i^{(q)}| + \epsilon)^{-1}.$$

Step 3: If $\mathbf{X}^{(q)} = \mathbf{X}^{(q-1)}$ or $q = q_{\text{max}}$, the algorithm terminates. Otherwise, set $q \to q+1$ and go to Step 2.

3.3 Experimental Study

We perform experiments to empirically study the performance of the proposed MultiPass Lasso algorithms. The common experimental setup is as follows. We set $n = 100$ and $m = 256$. The measurement matrix A has elements i.i.d. according to $\mathcal{N}(0, 1)$. The number of nonzero entries k increases from 9 to 57 with a step size 6, and the nonzero entries are independently drawn from $\mathcal{N}(0, 1)$. Note that the Gaussian distribution for the nonzero entries is commonly employed in the literature and it leads to variety in magnitude. Both termination criteria in Section 3.2.3 are employed. We claim a success in sparse signal recovery if $\|\hat{\mathbf{X}} - \mathbf{x}\|_2 / \|\mathbf{x}\|_2 \leq \tau$, where $\hat{\mathbf{X}}$ denotes the estimated sparse signal by some algorithm, and τ is a pre-defined constant. Each performance curve in this section is averaged over 200 random trials.

3.3.1 MPL: Selection of Regularization Parameter $\lambda^{(l)}$

First, we study the selection of the regularization parameter $\lambda^{(l)}$ in each internal iteration of MPL. As earlier mentioned, the form $\lambda^{(l)} = \frac{\gamma}{n}\|A^{(l)\mathsf{T}}\mathbf{Y}^{(l)}\|_\infty$ with a fixed $\gamma \in (0,1)$ for all iterations is employed. We set $\sigma = 0$, $\delta = 10^{-14}$, $s_{\max} = 0.7n$, $\tau = 10^{-3}$, and a zero entry is determined using the threshold 10^{-7}. The MPL algorithm is implemented using FPC [63]. Figure 3.2 illustrates the behavior of MPL with different choices of γ.

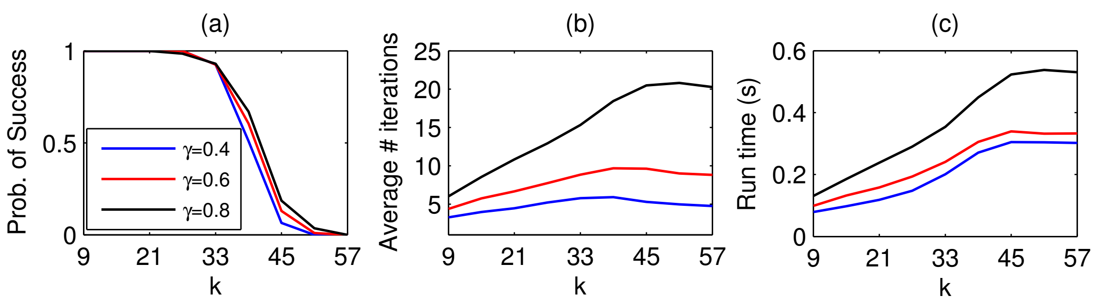
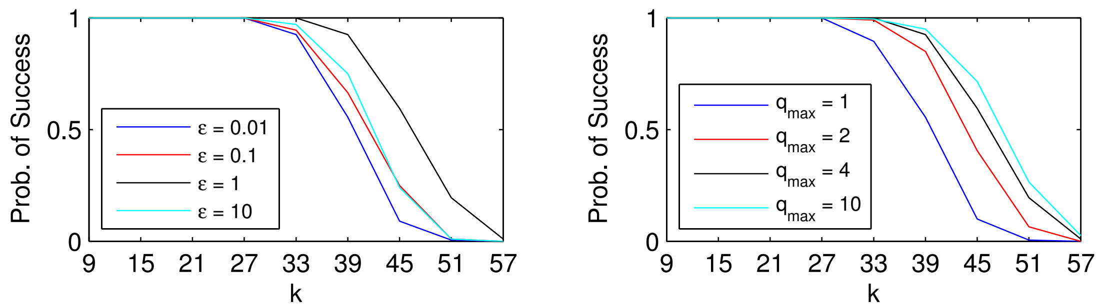


Figure 3.2: The role of γ in $\lambda^{(l)}$ for MPL.

Note that as γ increases, $\lambda^{(l)}$ increases, and a partial support of the sparse signal can be detected with higher probability [123]. In Figure 3.2(a), for the cases with more nonzero entries, i.e., larger k, using larger λ gives slight performance improvement. Meanwhile, as $\lambda^{(l)}$ increases, in general fewer nonzero entries appear in the solution vector $\mathbf{X}^{(l)}$ [44]. Hence, the average number of iterations needed for each experiment increases, leading to higher computational cost. Figure 3.2(b), (c) agree with this analysis.

3.3.2 RMPL: Selection of ϵ and $q_{\max}$

We explore the selection of parameters for RMPL. First, we set $q_{\max} = 4$, and focus on the effect of ϵ. For the component MPL, we choose $\gamma = 0.45$, $s_{\max} = 0.95n$. Other parameters are the same as in Section 3.3.1. Figure 3.3(a) shows the result.



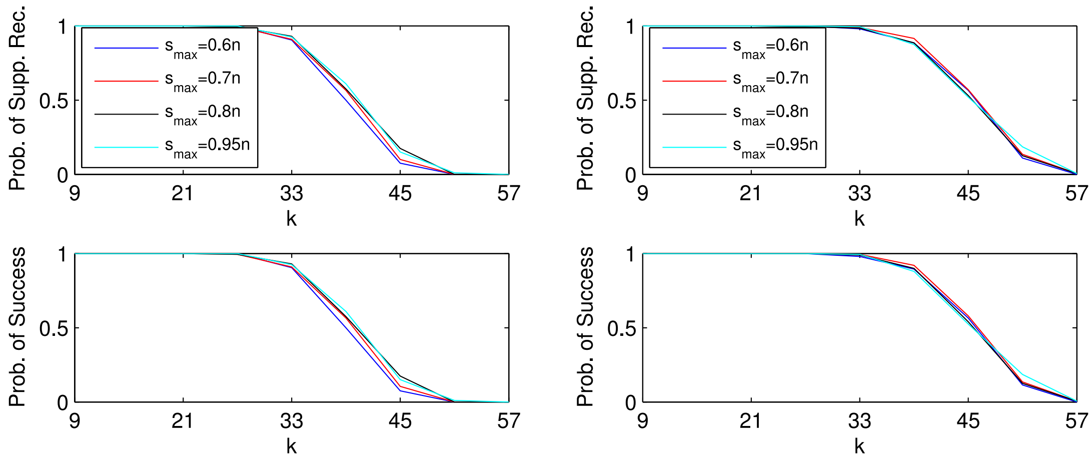
(a) The role of ϵ (b) The role of $q_{\max}$

Figure 3.3: Parameter selection for RMPL.

From Figure 3.3(a), we see that RMPL with $\epsilon = 1$ achieves the best performance. Next, let us fix $\epsilon = 1$, and study the impact of $q_{\max}$ on the performance of the algorithm. This is illustrated in Figure 3.3(b). Clearly, allowing more reweighted iterations helps improve the performance at the expense of computational cost.

3.3.3 MPL and RMPL: Selection of $s_{\max}$

Next, we study the role of the parameter $s_{\max}$ for MPL and RMPL. We choose $s_{\max} = 0.6n, 0.7n, 0.8n, 0.95n$, respectively. The experiment setup is the same as in Sections 3.3.1 and 3.3.2, except we fix $\gamma = 0.6$ for MPL, and $\epsilon = 1$ and $q_{\max} = 4$ for RMPL. The performance metrics are the success rate and the support recovery rate (using 10^{-7} to determine zero entry for all final solutions). Figure 3.4 summarizes the results.



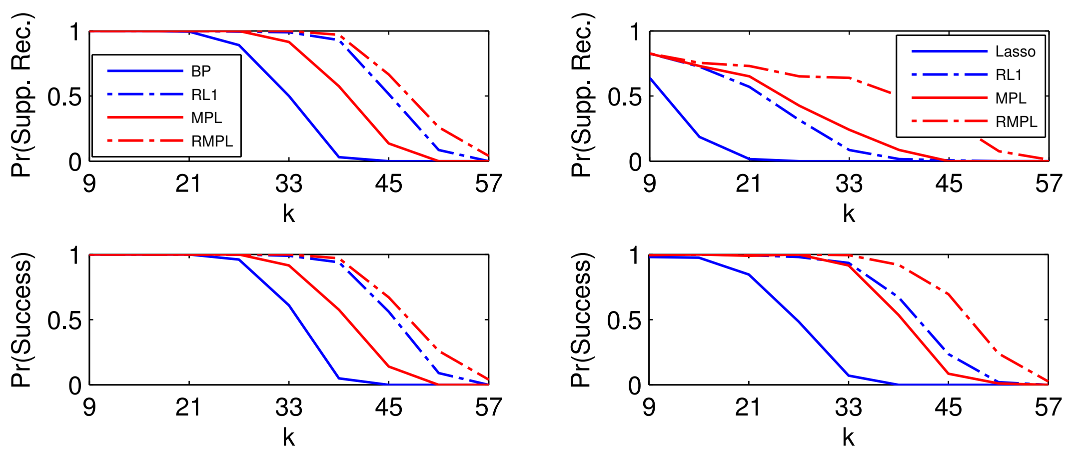
(a) MPL (b) RMPL

Figure 3.4: The role of $s_{\max}$ for MPL and RMPL. In each case, the upper plot shows support recovery rate, whereas the lower plot shows success rate.

From Figure 3.4(a), we see that the increase of $s_{\max}$ can slightly improve the performance of MPL for the cases with more nonzero entries. This is reasonable in the following sense. A large $s_{\max}$ may allow more indices to be selected, and it is more likely that $\mathrm{supp}(\mathbf{x}) \subseteq \mathcal{S}^{(l)}$ (assuming l iterations in total). Then, in the noiseless setting, the final least squares estimation in Step 3 of MPL may set the coefficients very close to zero on the indices outside the true support. These indices, though selected in $\mathcal{S}^{(l)}$, will be judged as corresponding to zero entries via thresholding. Next, based on Figure 3.4(b), we can see that the performance of RMPL with different $s_{\max}$ are very similar, leading to the observation that the multiple runs of MPL in RMPL has made the selection of $s_{\max}$ less important. In summary, this set of experiments indicates that the performances of MPL and RMPL, in terms of support recovery and estimation accuracy, are relatively insensitive to the selection of $s_{\max}$.

3.3.4 Comparison with BP, Lasso, and Reweighted ℓ_1 Minimization

First, we consider the noiseless scenario, i.e., $\sigma = 0$. We compare the performance between BP and MPL, and between reweighted ℓ_1 minimization (RL1) and RMPL. In this case, BP and RL1 are implemented via ℓ_1-MAGIC [17]. For MPL, $\gamma = 0.6$. For both RL1 and RMPL, $\epsilon = 1$ and $q_{\text{max}} = 10$. Other parameters are chosen the same way as in Sections 3.3.1 and 3.3.2. Figure 3.5(a) summarizes the results. It can be seen that MPL and RMPL outperform BP and RL1, respectively, in terms of both rate of success and rate of support recovery (using 10^{-5} to determine zero entry for all final solutions).



(a) Noiseless case (b) Noisy case

Figure 3.5: Sparse signal recovery in different settings. In each case, the upper plot shows support recovery rate, whereas the lower plot shows success rate.

Next, we consider the noisy setting with $\sigma = 0.01$. We compare among Lasso, RL1, MPL, and RMPL. All algorithms are implemented using the same Lasso routine [63]. For Lasso, we choose $\lambda = \sigma\sqrt{(2\log m)/n}$ as suggested in [24, 123]. This choice of λ is also employed by RL1. For MPL and RMPL, $\delta = 2.25\sigma^2$, and we use 10^{-4} as the threshold for determining zero entries. Meanwhile, $\tau = 10^{-2}$. Other parameters remain the same as in the noiseless case above. Figure 3.5(b) shows the result. We can see that MPL outperforms Lasso, and RMPL gives better performance than RL1 in both criteria (using 10^{-3} for determining zero entry in all final solutions). This set of experiments supports the design goal of MPL from which better support recovery is expected. The effectiveness of the proposed MPL and RMPL in various settings are also illustrated.

3.3.5 Computational Efficiency of MPL and RMPL

We examine the computational efficiency of MPL and RMPL with comparison to Lasso and RL1, respectively, in the noisy setting. The parameters are chosen the same as in the noisy case in Section 3.3.4, except that for RL1 and RMPL, we set $q_{\max} = 4$.

Three different Lasso solvers are employed, namely the Fixed-Point Continuation method (FPC) [63], the Truncated Newton Interior-Point method with Preconditioned Conjugate Gradients (L1LS) [73], and the Basic Gradient Projection method (GPSR) [48]. With each of the three Lasso solvers, we implement Lasso, RL1, MPL, and RMPL with the same initialization, termination condition, and precision control to ensure fair comparison. The computer in use runs MATLAB R2010a in Windows XP environment. Figure 3.6 summarizes the results.

According to Figure 3.6, the empirical probabilities of success for each algorithm using different Lasso solvers are almost identical. Most implementations of MPL perform faster than Lasso for a large range of k. Especially, MPL with GPSR implementation is faster than Lasso for the whole range of k tested here, and it is faster than

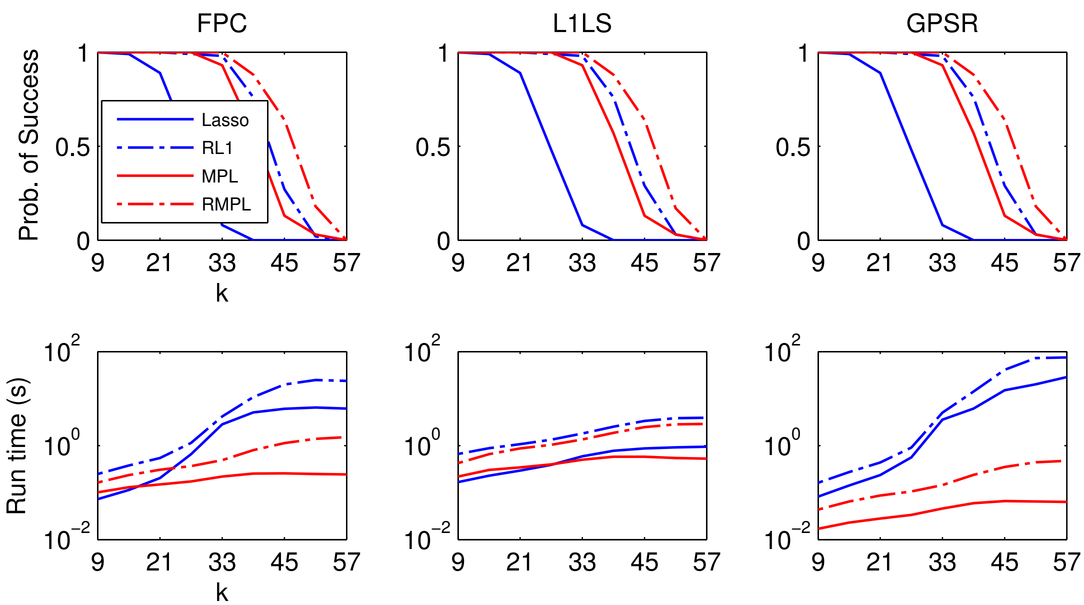


Figure 3.6: Comparison of algorithm efficiency. Each column presents the performance and the corresponding run time of algorithms implemented via the same Lasso solver specified in the title. The upper row indicates the empirical probability of success, and the lower row illustrates the average time consumed per experiment.

other implementations of Lasso as well. Meanwhile, with each optimization solver, RMPL achieves better performance and lower computational cost than RL1 built upon the same solver. The GPSR version of RMPL achieves the lowest computational cost among the tested cases. Overall, MPL and RMPL achieve both better performance and lower computational cost than their Lasso counterparts, respectively, suggesting their potential as effective and efficient algorithmic choices for sparse signal recovery.

3.4 Performance Limit of Orthogonal Matching Pursuit

In this section, we demonstrate that the connection between sparse signal recovery and multiuser communication can be also applied to shed light on the performance of practical algorithms. Specifically, we discuss the performance limits of the orthogonal matching pursuit algorithm. The analysis in this section is intuitive in nature, which means that the statements will not be justified by rigorous proof. Experiments will be presented as evidence to support the utility of the analysis. To begin with, we briefly review the orthogonal matching pursuit algorithm.

3.4.1 Orthogonal Matching Pursuit (OMP)

Orthogonal matching pursuit [94] is a greedy algorithm that utilizes sequential forward selection to determine the support of the sparse signal as follows.

Step 1: Let $\mathbf{r}_0 = \mathbf{Y}$, $\mathcal{S}_0 = \emptyset$, and $i = 1$.

Step 2: At iteration i, compute

$$p_i = \arg\max_{j \in [m]} \left| \mathbf{A}_j^\intercal \mathbf{r}_{i-1} \right| .$$

Let $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{p_i\}$, and

$$\mathbf{r}_i = P_{\mathcal{S}_i}^{\perp} \mathbf{r}_{i-1}$$

where $P_{\mathcal{S}_i}$ denotes the orthogonal projection to the subspace spanned by the columns of A indexed by $\mathcal{S}_i$, and $P_{\mathcal{S}_i}^{\perp}$ denotes the orthogonal complement projection of $P_{\mathcal{S}_i}$, i.e., $P_{\mathcal{S}_i}^{\perp} = I - P_{\mathcal{S}_i}$.

Step 3: Check if a predefined termination condition is satisfied. If so, output $\mathcal{S}_i$ as the estimated support, and $\hat{\mathbf{X}}_{\text{OMP}} = (A_{\mathcal{S}_i}^\intercal A_{\mathcal{S}_i})^{-1} A_{\mathcal{S}_i}^\intercal \mathbf{Y}$ as the values on the support $\mathcal{S}_i$.

Otherwise, let $i \to i + 1$, and go to Step 2.

3.4.2 Connection to Successive Interference Cancellation

Successive interference cancellation (SIC) is an effective approach for decoding in multiuser communication [118]. The basic characters of an SIC scheme are summarized as follows.

1. Decode one sender by treating all other undecoded senders as interferences.

2. Remove the contribution of the previously decoded senders from the received signal.

3. Iterate over the two steps above until certain stopping criterion is satisfied.

Based on the connection between sparse signal recovery and multiuser communication discussed in Section 2.3, we realize that OMP can be viewed as an SIC decoding scheme. It first chooses a sender whose codeword has the maximum correlation with the residual signal, and then removes its contribution by projecting the residual signal onto the orthogonal complement of the space spanned by previously decoded codewords. This connection provides us with the opportunity to obtain insights into the performance limit of OMP.

3.4.3 Intuitive Necessary Condition for OMP to Succeed

To shed light on the performance limit of OMP, we employ the criterion for an SIC scheme to successfully decode all senders. For ease of exposition, we first discuss a simple case where only two senders access a MAC, i.e., $k = 2$, whose capacity region is depicted in Figure 3.7 (assuming that each sender obeys the power constraint σ_a^2 and the Gaussian channel noise has variance σ_z^2).

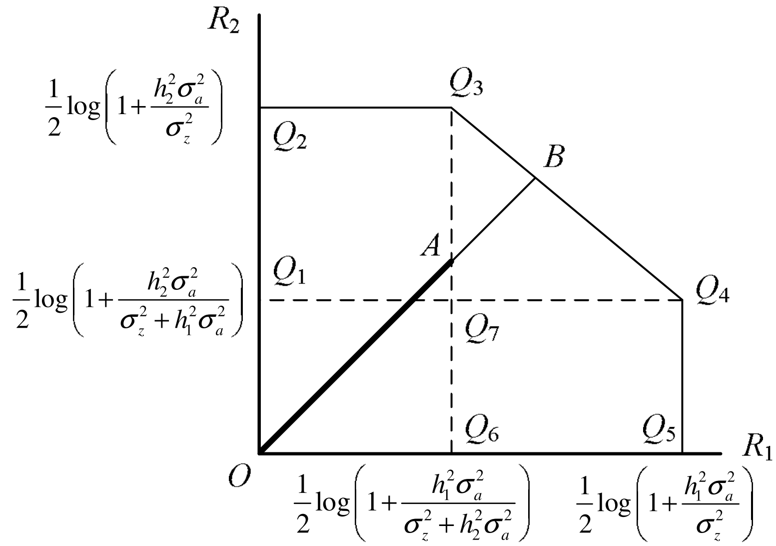


Figure 3.7: Capacity region of a 2-sender Gaussian MAC.

The capacity region is given by the pentagon $OQ_2Q_3Q_4Q_5$. Note that in rectangular region $OQ_2Q_3Q_6$, where

$$R_1 \leq \frac{1}{2}\log\left(1+\frac{h_1^2\sigma_a^2}{\sigma_z^2+h_2^2\sigma_a^2}\right) \tag{3.2}$$

$$R_2 \leq \frac{1}{2}\log\left(1+\frac{h_2^2\sigma_a^2}{\sigma_z^2}\right) \tag{3.3}$$

sender 1 can be decoded first by treating sender 2 as interference, which in effect leads to a raised noise level according to (3.2). After removing the decoded sender 1's information from the received signal, (3.3) indicates that sender 2 can be successfully decoded as if it is the only sender accessing a Gaussian channel. Rate pairs in rectangular region

$OQ_1Q_4Q_5$ can be decoded similarly but in the reverse order. The intersection $OQ_1Q_7Q_6$ can be decoded in an arbitrary order. However, no successive interference cancellation scheme can successfully decode both senders when the channel operates in triangular region $Q_7Q_3Q_4$, which instead should be decoded by joint decoding methods.

In this example, an ideal SIC scheme should be able to successfully decode both senders if and only if there exists a sequence $\{i_{(1)}, i_{(2)}\}$, $i_{(1)}, i_{(2)} \in \{1, 2\}$, such that

$$R_{i_{(j)}} \leq \frac{1}{2} \log \left(1 + \frac{h_{i_{(j)}}^2 \sigma_a^2}{\sum\limits_{u=j+1}^{2} h_{i_{(u)}}^2 \sigma_a^2 + \sigma_z^2} \right), j \in \{1, 2\}. \tag{3.4}$$

To generalize (3.4) for a k-sender MAC, it can be shown that [118] an ideal SIC scheme can successfully decode k senders if and only if there exists a sequence $\{i_{(1)}, i_{(2)}, ..., i_{(k)}\}$ such that

$$R_{i_{(j)}} \leq \frac{1}{2} \log \left(1 + \frac{h_{i_{(j)}}^2 \sigma_a^2}{\sigma_z^2 + \sum\limits_{u=j+1}^{k} h_{i_{(u)}}^2 \sigma_a^2} \right), j \in [k]. \tag{3.5}$$

Consequently, the connection between multiuser communication and sparse signal recovery discussed in Section 2.3 indicates the following intuitive necessary condition for OMP. Note that we consider a sequence of support recovery problems where the sparse signals are generated according to the same nonzero signal vector $\mathbf{w}$.

Let the elements of A be i.i.d. $\mathcal{N} \sim (0, \sigma_a^2)$. If OMP can recover the support of the sparse signals with probability converging to one as $m \to \infty$, then there exists a sequence $\{w'_{(1)}, \cdots, w'_{(k)}\}$, where $\{w'_{(1)}, \cdots, w'_{(k)}\} = \{w_1, ..., w_k\}$, satisfying

$$\frac{\log m}{n} \leq \frac{1}{2} \log \left(1 + \frac{{w'_{(i)}}^2}{\mathsf{SNR}^{-1} + \sum_{u=i+1}^{k} {w'_{(u)}}^2} \right), i \in [k] \tag{3.6}$$

where $\mathsf{SNR} \triangleq \sigma_a^2/\sigma^2$ for notation simplicity.

Remark. This condition can be conjectured as necessary for support recovery via OMP due to the facts that (i) OMP may not be an ideal SIC method, and (ii) the unknown channel gain problem discussed in Section 2.3.3 may exert a negative impact on the achievable performance.

3.4.4 Observations on the Performance Limit of OMP

Role of the Distribution of Nonzero Magnitudes

Based on (3.6), we realize that the magnitude distribution of nonzero entries may play an important role in support recovery via OMP. The following two propositions, which extend the discussion in [131] to the noisy setting, offer some insights.

Proposition 1: If all nonzero entries have equal magnitude, i.e., $|w_i| = c, i \in [k]$, for some $c > 0$, then the ability of OMP to recover the indices of all nonzero entries is comparable to recovering any one of them at the first iteration.

Proposition 2: If the magnitudes of the nonzero entries form an exponentially decaying profile, i.e., $|w_i| = \alpha^{i-1}$ with some constant $0 < \alpha < 1$, then for any $j \in [k-1]$, there exists α' such that when $\alpha < \alpha'$, OMP cannot recover the indices corresponding to the j smallest nonzero entries in magnitude.

Large Measurement Matrix

We note that (3.6) can also shed light on the size of the measurement matrix A that OMP can work with. Especially, asymptotically successful support recovery is still possible even if m grows at an exponential order of n, i.e., $m = \Theta(\beta^{nR})$ for some $R > 0$, assuming all logarithms are base-β. This phenomenon echoes with the discoveries obtained via other analysis technique in [116], which stated the possibility of signal reconstruction using measurement matrices with similar dimensions.

Suboptimality of OMP

The common codebook problem discussed in Section 2.3.3 requires all senders to work at equal rate. For the case $k = 2$, this corresponds to $\overline{OB}$ in Figure 3.7. Because of its successive nature, OMP may at best work within $\overline{OA}$ (thick) only. In contrast, joint recovery methods can, in theory, at best work within $\overline{OB}$. This observation echoes the discussions in [18, 38], which observed that joint recovery methods (specifically, ℓ_1 norm minimization) outperformed OMP in challenging settings.

3.4.5 Experiments

We present experiments to offer insights into the performance of OMP. All measurement matrices have elements independently generated according to the same Gaussian distribution, and then every column is normalized to have unit norm. The results are averaged over a certain number of random independent trials.

We should point out that, from an information theoretic viewpoint, the analysis above actually focuses on the limit case where the problem size grows to infinity, i.e., $m \to \infty$. In contrary, computer simulations can only deal with problems up to certain size. Hence, using a simulation study to echo the theoretical analysis seems inadequate. However, as we shall see, when the problem size is relatively large, the trend of the performance of OMP, observed over a finite range of problem size, does not contradict the theoretic analysis.

Relation between SIC Criterion and Performance of OMP

We examine whether the criterion (3.6) exhibits practical guidance on the support recovery using OMP. Define the events

$\mathcal{E}$ = {There exists a sequence of $\{w'_{(1)}, \cdots, w'_{(k)}\}$ such that (3.6) is satisfied.}

$\mathcal{C}$ = {OMP correctly recovers the support of the sparse signal.}

$\mathcal{S}$ = {The indices, ordered according to the recovery stages of OMP, form a sequence satisfying (3.6).}

Figure 3.8(a)-(c) are produced under the setup that $m = 10^4$, $k = 7$, and n varies from 40 to 100. Figure 3.8(a) shows the empirical $\mathsf{P}(\mathcal{C}|\mathcal{E})$. Note that, in this case, the sparse signals are generated such that (3.6) can be satisfied. The curves indicates that as n increases, the empirical probability of successful support recovery increases. Figure 3.8(b) shows the empirical $\mathsf{P}(\mathcal{C}|\mathcal{E}^c)$. In this case, the sparse signals are generated such that (3.6) cannot be satisfied. We note that for the range of n considered in this experi-

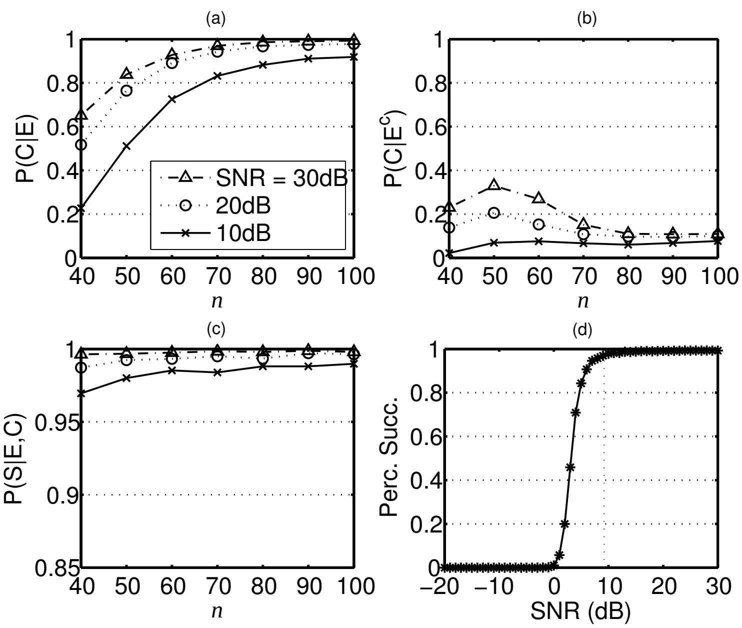


Figure 3.8: Performance of OMP in different scenarios. In (a) (b) (c): $m = 10^4, k = 7$. In (d): $n = 100, m = 1000, k = 7, w_i = 0.4, i \in [k]$.

ment, the success rate of support recovery are at very low level for all choices of SNR. Next, Figure 3.8(c) illustrates the empirical $\mathsf{P}(\mathcal{S}|\mathcal{E}, \mathcal{C})$. As we can see from the result, if OMP correctly recovers the support, it most likely proceeds according to a sequence satisfying (3.6) when such a sequence exists. Finally, Figure 3.8(d) simulates a smooth transition from a problem setup for which (3.6) cannot be satisfied to a problem setup for which (3.6) can be satisfied, which is implemented by increasing SNR while keeping other parameters fixed. The goal is to study the performance evolution of OMP as condition improves. The vertical dashed line shows the boundary ($\mathsf{SNR} = 9.2$dB) between the two types of problem setups. As we can see, at low SNR, OMP is very likely to fail. At high SNR, OMP succeeds in support recovery with very high probability. Before crossing the boundary, OMP starts to be able to recover the support with increasing probability. The actual crossing occurs at a problem setup with overwhelming probability of success.

Support Recovery with Large Measurement Matrices

To echo the discussion in Section 3.4.4 on the size of measurement matrix A that OMP can work with, we perform the following experiment. By keeping the ratio $(\log m)/n$ fixed while increasing m and n, we examine the empirical $\mathsf{P}(\mathcal{C}|\mathcal{E})$. Figure 3.9 summarizes the result.

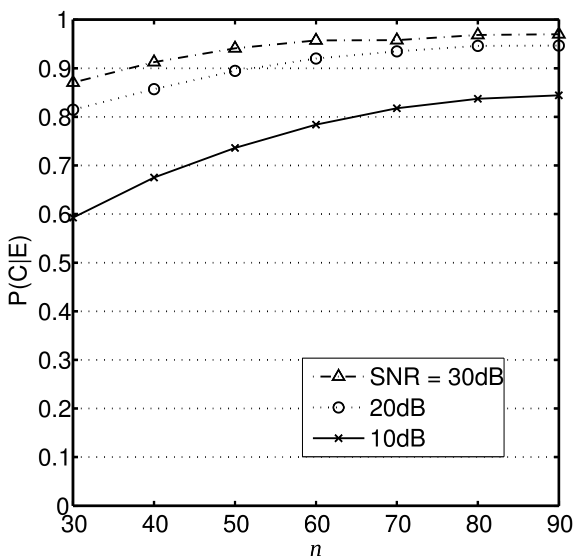


Figure 3.9: The performance of support recovery of OMP with large measurement matrices. ($m = \lfloor 2^{0.2n} \rfloor, k = 7$)

Note that, in this case, all sparse signals meet (3.6). From Figure 3.9, OMP seems capable of correctly recovering the support as m and n increase while maintaining the exponential relationship. Smaller error probability is obtained for larger m and n, which implies that OMP can yield performance improvement as the problem size grows while an exponential relationship between m and n is maintained.

In summary, the intuitive analysis on the performance limit of OMP is supported by the experiments. It is also worthwhile to realize that similar analysis can also be applied to the matching pursuit algorithm and its other variants.

3.5 Acknowledgements

Chapter 3 is, in part, a reprint of the paper “MultiPass Lasso Algorithms for Sparse Signal Recovery,” Y. Jin and B. D. Rao, *published in the proceedings of the International Symposium on Information Theory*, 2011, and the paper “Performance Limit of Matching Pursuit Algorithms,” Y. Jin and B. D. Rao, *published in the proceedings of the International Symposium on Information Theory*, July 2008. In both cases, I was the primary author, and B. D. Rao supervised the research.

3.6 Appendices

3.6.1 Probability Lower Bound for Lasso

This section presents the probability lower bound for Lasso on support recovery, which is employed to produce the figures in Section 3.2.4. This lower bound for Lasso is based on Theorem 3 in [123], with slight modification to accommodate flexible settings, to offer a probabilistic characterization of the support recovery performance. Define the auxiliary function for notation simplicity, for $\eta \in (0, \frac{1}{2})$, $\lambda > 0$,

$$g(\eta, \lambda) \triangleq \left(1 + \max(\eta, 8\sqrt{k/n})\right) \left(k/n + \sigma^2(n-k)/\lambda^2 n^2\right).$$

Theorem 8 (Modified based on Theorem 3 of [123])**.** *Suppose the elements of the measurement matrix A is independently generated according to $\mathcal{N}(0,1)$. Choose $\lambda > 0$. Let*

$$\mathcal{E}_1 \triangleq \{\mathit{supp}(\mathbf{X}_{\mathit{Lasso}}) \subseteq \mathit{supp}(\mathbf{x})\}$$
$$\mathcal{E}_2 \triangleq \{\|\mathbf{X}_{\mathit{Lasso}} - \mathbf{x}\|_\infty \leq \lambda(8\sqrt{k/n} + \sqrt{128k\log(m-k)/n} + \sqrt{18} + 1)\}.$$

Then, for $\eta \in (0, \frac{1}{2})$,

$$\begin{aligned} \mathsf{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 &- 2(m-k)\exp\left(-\frac{1}{2g(\eta,\lambda)}\right) \\ &- 2\exp\left(-\frac{3(n-k)\eta^2}{16}\right) - 4\exp\left(-\frac{k}{2}\right) \\ &- 2k\exp\left(-\log(m-k)\right) - 2k\exp\left(-\frac{n\lambda^2}{\sigma^2}\right) \\ &- 2\exp\left(-\frac{n}{2}\right). \end{aligned} \tag{3.7}$$

Proof Sketch. The proof follows Section V of [123] with mainly the following modification. We choose $t = \sqrt{18}\lambda$ (as opposed to $t = 20\sqrt{\sigma^2 \log k/n}$ in the reference) to reach a bound similar to (42) of [123]. Thus, (3.7) can be reached via straightforward derivation parallel to [123] with minor clarifications on some terms. $\square$

Remark 1. This theorem provides a lower bound, for a given choice of λ, on the probability that the reconstructed support being contained in the true support and the elementwise maximum distance of the reconstructed signal to the true signal being upper bounded. For the scenario defined in Section 3.2.4, successful support recovery is implied by $\mathcal{E}_1 \cap \mathcal{E}_2$ if we require

$$\lambda(8\sqrt{k/n} + \sqrt{128k\log(m-k)/n} + \sqrt{18} + 1) < x_{\text{low}}.$$

This relation naturally gives the selection of λ for Lasso to maximize this probability lower bound for support recovery.

Remark 2. Note that Theorem 3 in [123] mainly focuses on the relations among model parameters for successful signal reconstruction in the asymptotic sense. It states that $\mathsf{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - c_1 \exp(-c_2 \min(k, \log(m-k)))$ for some constants $c_1, c_2 > 0$ with certain family of λ and implicitly under the scaling $n = \Omega(k\log(m-k))$. In contrast, our modification (3.7) emphasizes the roles of n, m, k in finite settings. Meanwhile, it can afford a more flexible upper bound for $\|\mathbf{X}_{\text{Lasso}} - \mathbf{x}\|_\infty$, in contrast to the original bound (33) of [123] which may not be smaller than $20\sqrt{\sigma^2 \log k/n}$. Further, we explicitly work out all the constants to facilitate further comparisons.

The internal free parameter is chosen as $\eta = 0.49$ for plotting the figures in Section 3.2.4.

3.6.2 Derivation of Probability Lower Bound for MPL

In this section, we derive the probability lower bound for successful support recovery using MultiPass Lasso in the scenario defined in Section 3.2.4. This lower bound is then employed to produce the figures in that section.

Define $\mathcal{S} \triangleq \text{supp}(\mathbf{x}) = \mathcal{S}_h \cup \mathcal{S}_l$, and the events

$$\begin{aligned}
\mathcal{G} &\triangleq \left\{\text{supp}(\mathbf{X}^{(1)}) \subseteq \mathcal{S}\right\} \cap \left\{\mathcal{S}_h \subseteq \text{supp}(\mathbf{X}^{(1)})\right\} \\
&\quad \cap \left\{\lambda_1 < \frac{1}{n}\|A^{\intercal}\mathbf{Y}\|_\infty\right\} \cap \left\{\frac{1}{n}\|\mathbf{N}\|_2^2 \le \delta\right\} \\
\mathcal{T} &\triangleq \left\{\text{supp}(\mathbf{X}^{(2)}) \subseteq (\mathcal{S}\backslash\text{supp}(\mathbf{X}^{(1)}))\right\} \\
&\quad \cap \left\{(\mathcal{S}_l\backslash\text{supp}(\mathbf{X}^{(1)})) \subseteq \text{supp}(\mathbf{X}^{(2)})\right\} \\
&\quad \cap \left\{\lambda_2 < \frac{1}{n}\|A^{(1)\intercal}\mathbf{Y}^{(1)}\|_\infty\right\} \cap \left\{\frac{1}{n}\|\mathbf{Y}^{(1)}\|_2^2 > \delta\right\}.
\end{aligned}$$

According to the total probability rule,

$$
\begin{aligned}
&\mathsf{P}(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}) \\
&\geq \mathsf{P}\left(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}\big|\mathcal{G}\right) \mathsf{P}\left(\mathcal{G}\right) \\
&\geq \big(\mathsf{P}(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}, \mathrm{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}|\mathcal{G}) \\
&\quad + \mathsf{P}(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}, \mathrm{supp}(\mathbf{X}^{(1)}) = \mathcal{S}|\mathcal{G})\big)\, \mathsf{P}(\mathcal{G}) \\
&= \big(\mathsf{P}(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}, \mathrm{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}|\mathcal{G}) \\
&\quad + \mathsf{P}(\mathrm{supp}(\mathbf{X}^{(1)}) = \mathcal{S}|\mathcal{G})\big)\, \mathsf{P}(\mathcal{G}) && (3.8) \\
&\geq \big(\mathsf{P}(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}, \mathrm{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}, \mathcal{T}|\mathcal{G}) \\
&\quad + \mathsf{P}(\mathrm{supp}(\mathbf{X}^{(1)}) = \mathcal{S}|\mathcal{G})\big)\, \mathsf{P}(\mathcal{G}) \\
&\geq \big(\mathsf{P}\left(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}|\mathcal{T}, \mathrm{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}, \mathcal{G}\right) \\
&\quad \cdot \mathsf{P}(\mathcal{T}, \mathrm{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}\big|\mathcal{G}) + \mathsf{P}(\mathrm{supp}(\mathbf{X}^{(1)}) = \mathcal{S}|\mathcal{G})\big)\, \mathsf{P}(\mathcal{G}) \\
&= \big(\mathsf{P}(\mathcal{T}, \mathrm{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}\big|\mathcal{G}) + \mathsf{P}(\mathrm{supp}(\mathbf{X}^{(1)}) = \mathcal{S}|\mathcal{G})\big)\, \mathsf{P}(\mathcal{G}) && (3.9)
\end{aligned}
$$

where (3.8) follows from

$$\mathsf{P}(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}|\mathrm{supp}(\mathbf{X}^{(1)}) = \mathcal{S}, \mathcal{G}) = 1$$

and (3.9) follows from

$$\mathsf{P}(\mathrm{supp}(\mathbf{X}_{\mathrm{MPL}}) = \mathcal{S}|\mathcal{T}, \mathrm{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}, \mathcal{G}) = 1.$$

Let $\mathbf{1}$ denote a vector of all ones with a proper length. Define the following auxiliary random variable and events, for any fixed set $\mathcal{V} \subset \mathcal{S}$ satisfying $\mathcal{S}_h \subseteq \mathcal{V}$,

$$
\begin{aligned}
\mathbf{Q}_{\mathcal{V}} &\triangleq P_{\mathcal{V}}^{\perp}(x_{\text{low}} A_{\mathcal{S}\setminus\mathcal{V}} \mathbf{1} + \mathbf{N}) \\
\mathcal{F}_{\mathcal{V}} &\triangleq \left\{ \lambda_2 < \frac{1}{n} \|(P_{\mathcal{V}}^{\perp} A)^{\mathsf{T}} \mathbf{Q}_{\mathcal{V}}\|_{\infty} \right\} \\
\mathcal{H}_{\mathcal{V}} &\triangleq \left\{ \frac{1}{n} \|\mathbf{Q}_{\mathcal{V}}\|_2^2 > \delta \right\} \\
\mathcal{J}_{\mathcal{V}} &\triangleq \left\{ \text{supp}(\mathbf{X}^{(1)}) = \mathcal{V} \right\}.
\end{aligned}
$$

Thus, using the fact that $\mathsf{P}(\mathcal{A} \cap \mathcal{B}) \geq \mathsf{P}(\mathcal{A}) - \mathsf{P}(\mathcal{B}^c)$, one can have

$$
\begin{aligned}
&\mathsf{P}(\mathcal{T}, \text{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S} \big| \mathcal{G}) \\
&= \mathsf{P}(\mathcal{T}, \text{supp}(\mathbf{X}^{(1)}) \subset \mathcal{S}, \mathcal{G}) / \mathsf{P}(\mathcal{G}) \\
&= \sum_{\mathcal{V}:\mathcal{V}\subset\mathcal{S},\mathcal{S}_h\subseteq\mathcal{V}} \frac{\mathsf{P}(\mathcal{T}, \mathcal{J}_{\mathcal{V}}, \mathcal{G})}{\mathsf{P}(\mathcal{G})} \\
&= \sum_{\mathcal{V}:\mathcal{V}\subset\mathcal{S},\mathcal{S}_h\subseteq\mathcal{V}} \frac{\mathsf{P}(\text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})) \subseteq \mathcal{S}\setminus\mathcal{V}, \mathcal{S}_l\setminus\mathcal{V} \subseteq \text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})), \mathcal{F}_{\mathcal{V}}, \mathcal{H}_{\mathcal{V}}, \mathcal{J}_{\mathcal{V}}, \mathcal{G})}{\mathsf{P}(\mathcal{G})} \\
&\geq \sum_{\mathcal{V}:\mathcal{V}\subset\mathcal{S},\mathcal{S}_h\subseteq\mathcal{V}} \\
&\qquad \frac{\mathsf{P}(\mathcal{J}_{\mathcal{V}}, \mathcal{G})) - \mathsf{P}((\text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})) \subseteq \mathcal{S}\setminus\mathcal{V}, \mathcal{S}_l\setminus\mathcal{V} \subseteq \text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})), \mathcal{F}_{\mathcal{V}}, \mathcal{H}_{\mathcal{V}})^c)}{\mathsf{P}(\mathcal{G})} \\
&= \sum_{\mathcal{V}:\mathcal{V}\subset\mathcal{S},\mathcal{S}_h\subseteq\mathcal{V}} \\
&\qquad \mathsf{P}(\mathcal{J}_{\mathcal{V}} | \mathcal{G}) - \frac{\mathsf{P}((\text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})) \subseteq \mathcal{S}\setminus\mathcal{V}, \mathcal{S}_l\setminus\mathcal{V} \subseteq \text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})), \mathcal{F}_{\mathcal{V}}, \mathcal{H}_{\mathcal{V}})^c)}{\mathsf{P}(\mathcal{G})}
\end{aligned}
\tag{3.10}
$$

where $\text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}}))$ indicates that the reconstruction is based on $\mathbf{Q}_{\mathcal{V}}$, which is a possible form of $\mathbf{Y}^{(1)}$. Due to the facts that $\sum_{\mathcal{V}:\mathcal{V}\subset\mathcal{S},\mathcal{S}_h\subseteq\mathcal{V}} \mathsf{P}(\mathcal{J}_{\mathcal{V}}|\mathcal{G}) + \mathsf{P}(\text{supp}(\mathbf{X}^{(1)}) = \mathcal{S}|\mathcal{G}) = 1$, and that $\mathsf{P}((\mathcal{A}\cap\mathcal{B})^c) \leq \mathsf{P}(\mathcal{A}^c) + \mathsf{P}(\mathcal{B}^c)$, we have

$$\begin{aligned}
&\mathsf{P}(\text{supp}(\mathbf{X}_{\text{MPL}}) = \mathcal{S}) \\
&\geq \\
&\left[1 - \sum_{\mathcal{V}:\mathcal{V}\subset\mathcal{S},\mathcal{S}_h\subseteq\mathcal{V}} \frac{\mathsf{P}\left((\text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})) \subseteq \mathcal{S}\backslash\mathcal{V}, \mathcal{S}_l\backslash\mathcal{V} \subseteq \text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})), \mathcal{F}_{\mathcal{V}}, \mathcal{H}_{\mathcal{V}})^c\right)}{\mathsf{P}(\mathcal{G})}\right] \\
&\quad \cdot \mathsf{P}(\mathcal{G}) \\
&\geq \mathsf{P}(\mathcal{G}) - \sum_{\mathcal{V}:\mathcal{V}\subset\mathcal{S},\mathcal{S}_h\subseteq\mathcal{V}} \left[\mathsf{P}\left((\text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})) \subseteq \mathcal{S}\backslash\mathcal{V}, \mathcal{S}_l\backslash\mathcal{V} \subseteq \text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})))^c\right)\right. \\
&\quad \left. + \mathsf{P}(\mathcal{F}_{\mathcal{V}}^c) + \mathsf{P}(\mathcal{H}_{\mathcal{V}}^c)\right] \qquad (3.11)
\end{aligned}$$

It remains to work out proper bounds for the terms in (3.11), respectively. First, we consider $\mathsf{P}(\mathcal{G})$, which concerns the behavior in the first iteration of the MultiPass Lasso algorithm. Note that

$$\begin{aligned}
\mathsf{P}(\mathcal{G}) \geq 1 &- \mathsf{P}\left(\|\mathbf{N}\|_2^2/n \ > \delta\right) - \mathsf{P}\left(\lambda_1 \geq \|A^{\intercal}\mathbf{Y}\|_\infty/n\right) \\
&- \mathsf{P}((\text{supp}(\mathbf{X}^{(1)}) \subseteq \mathcal{S}, \mathcal{S}_h \subseteq \text{supp}(\mathbf{X}^{(1)}))^c). \qquad (3.12)
\end{aligned}$$

First, using the Chernoff bound for χ^2 random variables as in Lemma 1, we have, for

$\delta > \sigma^2$,

$$\mathsf{P}\left(\|\mathbf{N}\|_2^2/n > \delta\right) \leq \exp\left(-n \cdot h(\sigma^2, \delta)\right). \tag{3.13}$$

where $h(\alpha, \beta) \triangleq \frac{\beta}{2\alpha} - \frac{1}{2} + \frac{1}{2}\log\frac{\alpha}{\beta}$.

Next, we choose an arbitrary $i \in \mathcal{S}_h$. Then, for any η_1, η_2 such that $(1 + \sqrt{18})^{-1} < \eta_1 < 1, \eta_2 > 1$,

$$\begin{aligned}\mathsf{P}\left(\lambda_1 \geq \|A^\intercal\mathbf{Y}\|_\infty/n\right) \leq \mathsf{P}&\left(\lambda_1 \geq \|A^\intercal\mathbf{Y}\|_\infty/n \middle| \eta_1 < \|\mathbf{A}_i\|_2^2/n < \eta_2\right)\\ &+ \mathsf{P}\left(\|\mathbf{A}_i\|_2^2/n \leq \eta_1\right) + \mathsf{P}\left(\|\mathbf{A}_i\|_2^2/n \geq \eta_2\right).\end{aligned}$$

where $\mathbf{A}_i$ denotes the ith column of A. It can be readily seen that

$$\begin{aligned}&\mathsf{P}(\|\mathbf{A}_i\|_2^2/n \leq \eta_1) \leq \exp\left(-n \cdot h(1, \eta_1)\right)\\ &\mathsf{P}(\|\mathbf{A}_i\|_2^2/n \geq \eta_2) \leq \exp\left(-n \cdot h(1, \eta_2)\right).\end{aligned}$$

Due to the fact that

$$\frac{1}{n}\|A^\intercal\mathbf{Y}\|_\infty \geq \left|\frac{x_{\text{high}}}{n}\|\mathbf{A}_i\|_2^2 + \frac{1}{n}\mathbf{A}_i^\intercal(x_{\text{high}}A_{\mathcal{S}_h\setminus\{i\}}\mathbf{1} + x_{\text{low}}A_{\mathcal{S}_l}\mathbf{1} + \mathbf{N})\right|$$

and the independence among $\mathbf{A}_i$, $A_{\mathcal{S}\setminus\{i\}}$ and $\mathbf{N}$, we employ the tail bound [123, Appendix A] to obtain

$$\begin{aligned}
&\mathsf{P}\left(\lambda_1 \geq \frac{1}{n}\|A^\intercal \mathbf{Y}\|_\infty \middle| \mathbf{A}_i, \eta_1 < \|\mathbf{A}_i\|_2^2/n < \eta_2\right) \\
&\leq \exp\left(-\frac{n(\lambda_1 - x_{\text{high}}\eta_1)^2}{2(x_{\text{high}}^2(k_h-1) + x_{\text{low}}^2 k_l + \sigma^2)\eta_2}\right).
\end{aligned}$$

Putting pieces together, one can have

$$\begin{aligned}
\mathsf{P}\left(\lambda_1 \geq \frac{1}{n}\|A^\intercal \mathbf{Y}\|_\infty\right) &\leq \exp\left(-\frac{n(\lambda_1 - x_{\text{high}}\eta_1)^2}{2(x_{\text{high}}^2(k_h-1) + x_{\text{low}}^2 k_l + \sigma^2)\eta_2}\right) \\
&\quad + \exp\left(-n \cdot h(1,\eta_1)\right) + \exp\left(-n \cdot h(1,\eta_2)\right). \qquad (3.14)
\end{aligned}$$

Next, note that, for any μ satisfying $0 < \mu < x_{\text{high}}$, the event $\{\text{supp}(\mathbf{X}^{(1)}) \subseteq \mathcal{S}\} \cap \{\|\mathbf{X}^{(1)} - \mathbf{x}\|_\infty \leq \mu\}$ implies the event $\{\text{supp}(\mathbf{X}^{(1)}) \subseteq \mathcal{S}\} \cap \{\mathcal{S}_h \subseteq \text{supp}(\mathbf{X}^{(1)})\}$. This observation leads to

$$\begin{aligned}
&\mathsf{P}((\text{supp}(\mathbf{X}^{(1)}) \subseteq \mathcal{S}, \mathcal{S}_h \subseteq \text{supp}(\mathbf{X}^{(1)}))^c) \\
&\leq \mathsf{P}((\text{supp}(\mathbf{X}^{(1)}) \subseteq \mathcal{S}, \|\mathbf{X}^{(1)} - \mathbf{x}\|_\infty < x_{\text{high}})^c) \\
&\leq \mathsf{P}((\text{supp}(\mathbf{X}^{(1)}) \nsubseteq \mathcal{S}) + \mathsf{P}(\|\mathbf{X}^{(1)} - \mathbf{x}\|_\infty \geq x_{\text{high}}). \qquad (3.15)
\end{aligned}$$

We recognize that the probability bound above belongs to the scenario considered in Theorem 1 with the choice $\lambda = \lambda_1 - \rho$. Therefore, $\mathsf{P}(\text{supp}(\mathbf{X}^{(1)}) \subseteq \mathcal{S}, \mathcal{S}_h \subseteq \text{supp}(\mathbf{X}^{(1)}))$ can be lower bounded by the right hand side of (3.7) with the choice $\lambda = \lambda_1 - \rho$. Thus far, we obtained a lower bound for $\mathsf{P}(\mathcal{G})$.

Next, we consider the remaining terms in the brackets of (3.11). From an analytical perspective, these terms seem similar to the terms in (3.12), except that now the effect of the orthogonal projection $P_{\mathcal{V}}^{\perp}$ after the first iteration must be considered accordingly. Here, we demonstrate a useful technique to address this issue by working out one remaining term in (3.11) as an example. Let us focus on $\mathsf{P}(\mathcal{H}_{\mathcal{V}}^{c})$. Let $|\mathcal{V}| = v$. First, we condition the analysis on $A_{\mathcal{V}}$, which implies the conditioning on $P_{\mathcal{V}}^{\perp}$. It is easy to verify that $P_{\mathcal{V}}^{\perp} P_{\mathcal{V}}^{\perp} = P_{\mathcal{V}}^{\perp}$ and $P_{\mathcal{V}}^{\perp\intercal} = P_{\mathcal{V}}^{\perp}$. One can decompose $P_{\mathcal{V}}^{\perp} = U^{\intercal} U$ where $U \in \mathbb{R}^{(n-v)\times n}$ and $UU^{\intercal} = I$. With these at hand,

$$\frac{1}{n}\|\mathbf{Q}_{\mathcal{V}}\|_2^2 = \frac{1}{n}\|P_{\mathcal{V}}^{\perp}(x_{\text{low}} A_{\mathcal{S}\setminus\mathcal{V}}\mathbf{1} + \mathbf{N})\|_2^2 = \frac{1}{n}\|U(x_{\text{low}} A_{\mathcal{S}\setminus\mathcal{V}}\mathbf{1} + \mathbf{N})\|_2^2 \tag{3.16}$$

where $U(x_{\text{low}} A_{\mathcal{S}\setminus\mathcal{V}}\mathbf{1}+\mathbf{N}) \in \mathbb{R}^{n-v}$ and $U(x_{\text{low}} A_{\mathcal{S}\setminus\mathcal{V}}\mathbf{1}+\mathbf{N}) \sim \mathcal{N}(\mathbf{0}, ((k-v)x_{\text{low}}^2+\sigma^2)I)$. Hence,

$$\begin{aligned}
\mathsf{P}(\mathcal{H}_{\mathcal{V}}^{c}|A_{\mathcal{V}}) &= \mathsf{P}\left(\frac{1}{n}\|\mathbf{Q}_{\mathcal{V}}\|_2^2 \le \delta \Big| A_{\mathcal{V}}\right) \\
&= \mathsf{P}\left(\frac{1}{n-v}\|U(x_{\text{low}} A_{\mathcal{S}\setminus\mathcal{V}}\mathbf{1} + \mathbf{N})\|_2^2 \le \frac{n}{n-v}\delta \Big| A_{\mathcal{V}}\right) \\
&\le \exp\big(-(n-v)\cdot h((k-v)x_{\text{low}}^2 + \sigma^2, n\delta/(n-v))\big).
\end{aligned} \tag{3.17}$$

Note that (3.17) is independent of $A_{\mathcal{V}}$ (or, $P_{\mathcal{V}}^{\perp}$). Therefore, $\mathsf{P}(\mathcal{H}_{\mathcal{V}}^{c})$ can be also upper bounded by (3.17).

Incorporating the techniques above for dealing with $P_{\mathcal{V}}^{\perp}$ into similar steps we performed for lower bounding $\mathsf{P}(\mathcal{G})$, one should be able to work out the following results via straightforward derivations,

$$\begin{aligned}\mathsf{P}(\mathcal{F}_{\mathcal{V}}^c) \leq \exp\left(-\frac{n(\lambda_2 - x_{\text{low}}\eta_1)^2}{2(x_{\text{low}}^2(k-v-1)+\sigma^2)\eta_2}\right) \\ + \exp\Big(-(n-v)\cdot h\Big(1, n\eta_1/(n-v)\Big)\Big) \\ + \exp\big(-(n-v)\cdot h(1, n\eta_2/(n-v))\big).\end{aligned} \tag{3.18}$$

Next, slightly different approach should be taken to analyze $\mathsf{P}((\text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})) \subseteq \mathcal{S}\backslash\mathcal{V}, \mathcal{S}_l\backslash\mathcal{V} \subseteq \text{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})))^c)$. The reason is that the number of remaining nonzero entries to recover ranges from 1 to k_l, whereas the probability lower bound (3.7) scales as $1-\exp(-\min(k, \log(m-k)))$. We need to modify the derivation in order to provide meaningful results for the worst case $v = k-1$ in the summation in (3.11). Therefore, we choose $t = (k/n)^{1/4}$ when applying Lemma 9 of [123] (as opposed to $t = \sqrt{k/n}$). Further, in Appendix G of [123], we instead work with the probability $\mathsf{P}(|||D|||_2 \geq 8(k/n)^{1/4}) \leq 2\exp(-\sqrt{nk}/2)$, later condition on $\{|||D|||_2 < 8(k/n)^{1/4}\}$, and continue the proof therein with the choice $t = \sqrt{128}$ to reach a bound similar to the last inequality of that section. It can be verified that $(\lambda_2-\rho)(\frac{n}{n-v}(8(\frac{k-v}{n-v})^{1/4}+\sqrt{128}+1)+\sqrt{18}) < x_{\text{low}}$ for $k_h \leq v < k$. With these modifications, one should be able to obtain that, for

$\eta \in (0, \frac{1}{2})$,

$$
\begin{aligned}
&\mathsf{P}\left((\mathrm{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})) \subseteq \mathcal{S}\backslash\mathcal{V}, \mathcal{S}_l\backslash\mathcal{V} \subseteq \mathrm{supp}(\mathbf{X}^{(2)}(\mathbf{Q}_{\mathcal{V}})))^c\right) \\
&\leq 2(m-k)\exp\Big(-\frac{1}{2g_1(v,\eta,\lambda_2-\rho)}\Big) \\
&\quad + 2\exp\Big(-\frac{3(n-k)\eta^2}{16}\Big) + 4\exp\Big(-\frac{\sqrt{(n-v)(k-v)}}{2}\Big) \\
&\quad + 2(k-v)\exp\Big(-\sqrt{\frac{n-v}{k-v}}\Big) \\
&\quad + 2(k-v)\exp\Big(-\frac{(n-v)(\lambda_2-\rho)^2}{\sigma^2}\Big) \\
&\quad + 2\exp\Big(-\frac{n-v}{2}\Big)
\end{aligned} \tag{3.19}
$$

where $g_1(v,\eta,\lambda) \triangleq (1+\max(\eta, 8(\frac{k-v}{n-v})^{1/4}))(\frac{k-v}{n-v} + \frac{\sigma^2(n-k)}{\lambda^2 n^2})$.

Finally, $\mathsf{P}(\mathrm{supp}(\mathbf{X}_{\text{MPL}}) = \mathcal{S})$ can be lower bounded by substituting (3.12) [with (3.13), (3.14), and (3.15) by replacing $\lambda = \lambda_1 - \rho$ in (3.7)], (3.17), (3.18), (3.19) into (3.11). For figures shown in Section 3.2.4, we choose the internal free parameters as $\eta_1 = 0.8, \eta_2 = 1.7, \eta = 0.49$.

Chapter 4

Robust Linear Regression by Exploiting the Connection to Sparse Signal Recovery

Starting in this chapter, we focus on novel applications of sparse signal recovery. The key approach is to identify a sparse component and then reformulate the problem in such a way that the usefulness of the techniques for sparse signal recovery become evident. Specifically, this chapter addresses the problem of robust regression. As we shall see, the outliers in the observations can be viewed as the sparse component, leading to the opportunity of using Bayesian techniques for sparse signal recovery to design novel algorithms for robust regression.

4.1 Introduction

Consider the problem of linear regression with the model

$$y_i = \mathbf{a}^\intercal \mathbf{x}_i + e_i,\ i \in [M] \tag{4.1}$$

where $\mathbf{x}_i \in \mathbb{R}^L$ is usually termed as the explanatory variable, y_i is the response variable, $\mathbf{a} \in \mathbb{R}^L$ is the regression coefficients, L is the model order, and e_i is the measurement noise in the ith response. Note that model (4.1) can be compactly represented by

$$\mathbf{y} = X\mathbf{a} + \mathbf{e} \tag{4.2}$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M]^\intercal$, $\mathbf{e} = [e_1, e_2, ..., e_M]^\intercal \in \mathbb{R}^M$, and $\mathbf{y} = [y_1, y_2, ..., y_M]^\intercal \in \mathbb{R}^M$. We assume X has full column rank. The goal is to determine the regression coefficients $\mathbf{a}$ and this is often achieved by using a suitable optimization criterion. This problem has many applications in science and engineering. An important factor that makes this problem interesting and challenging is that the response variable $\mathbf{y}$ may usually contain outliers. The popular ordinary Least Squares (LS) is sensitive to outliers and hence robust regression methods are of interest. Numerous approaches for robust regression have been developed [64,67,69] with the goal of extracting the model parameters reliably in the presence of outliers.

4.1.1 Background

As a popular technique, ordinary LS estimation determines the model parameters $\mathbf{a}$ by minimizing $\sum_{i=1}^{M} \tilde{e}_i^2$, where $\tilde{e}_i = y_i - \mathbf{a}^\intercal \mathbf{x}_i$ is the fitting error. This criterion is sensitive to outliers and hence not robust. Many existing methods for robust regression follow the idea that one should de-emphasize the impact of data with large deviation in order to obtain robustness. For example, the method of least absolute value (LAV) [43] is a well-known representative of this kind. This method minimizes $\sum_{i=1}^{M} |\tilde{e}_i|$, which can be equivalently viewed as imposing a Laplacian distribution on the measurement noise e_i. Alternatively, the family of M-estimates [69] consider flexible weighting schemes on the fitting error $\tilde{e}_i$. The weighting functions used in M-estimates aim to de-emphasize samples with large deviation, and they can also be related to certain probability densities imposed on the measurement noise. By assuming the measurement noise e_i is drawn from the Student's t-distribution [75], the impact of extreme errors is also effectively downscaled. Further, the model of Gaussian mixtures also has been employed in robust regression wherein samples of the measurement noise e_i are assumed to be i.i.d. and drawn from a mixture of two Gaussians with one accounting for regular noise and the other for outliers [56].

In addition, robust procedures that aim to explicitly remove the impact of extreme errors have also been developed. For instance, the method of Least Trimmed Squares (LTS) [103] employs the optimization criterion that minimizes only a portion of the squared fitting errors with smallest magnitudes. The essence of this method can be viewed as roughly detecting outliers and removing their impact at the data fitting stage. This idea can be generalized to various outlier diagnosis techniques [103]. For an extensive survey of previous work on robust regression and outlier detection, interested readers are referred to [67, 69, 103] and the references therein.

It is interesting to note that the probability distribution underlying various robust regression methods indeed have counterparts in the context of the sparse signal recovery. The heavy-tailed outlier-tolerating priors imposed on the measurement noise correspond to the sparsity-inducing distributions in sparse signal recovery. The Laplacian distribution in the LAV method and its use in the corresponding ℓ_1-norm minimization based sparse signal recovery algorithms serves as an excellent example of this kind. As another example, the LTS method exhibits very similar ingredient to the thresholding method that is used for finding sparse solutions. Our work examines this connection more deeply. Intuitively, this connection is made possible by the fact that outliers are events that occur infrequently, and thus sparse. Next, we start by proposing a two-component model for the additive noise and reformulate the regression problem such that the usefulness of sparse recovery methods is evident.

4.2 The Two-Component Model of Measurement Noise

We leverage the fact that outliers occur infrequently and hence are sparse. Unfortunately, the linear model (4.1) leaves us little opportunity to take advantage of this observation, since a single measurement noise term e_i deals with both the impact of outlier and regular noise. To explicitly make use of the sparsity of outliers, we suggest an alternative model by splitting e_i into two independent additive components, namely w_i and ϵ_i, as follows,

$$y_i = \mathbf{a}^\intercal \mathbf{x}_i + w_i + \epsilon_i,\ i \in [M]. \tag{4.3}$$

The interpretations of $\mathbf{x}_i, y_i$ and $\mathbf{a}$ are carried over from (4.1). If a response y_i is not an outlier, then the corresponding w_i is assumed to be zero. If y_i is an outlier, then w_i can be viewed as the anomalous error in y_i such that $(y_i - w_i)$ appears to be a response contaminated only by regular noise. The term ϵ_i, on the other hand, contains the regular measurement noise in response y_i, and it is modeled as i.i.d. zero mean Gaussian noise, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Compactly, model (4.3) can be represented by

$$\mathbf{y} = X\mathbf{a} + \mathbf{w} + \boldsymbol{\epsilon} = [X, I] \begin{bmatrix} \mathbf{a} \\ \mathbf{w} \end{bmatrix} + \boldsymbol{\epsilon}, \tag{4.4}$$

where, in addition to (4.2), $\mathbf{w} = [w_1, w_2, ..., w_M]^\intercal$, and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, ..., \epsilon_M]^\intercal$. By definition, $\mathbf{w}$ is a sparse vector, which means the number of nonzero entries of $\mathbf{w}$ is (much) smaller than the length of $\mathbf{w}$. As we shall see, this model (4.4) enables the opportunity to adapt sparse signal recovery methods to robust regression.

4.3 Sparse Signal Recovery Algorithms for Robust Linear Regression

Since $\mathbf{w}$ is a sparse vector, one can utilize ideas from sparse signal recovery to develop robust linear regression methods. In this work, we consider Bayesian methods: MAP techniques and empirical Bayesian methods. For MAP methods, we assume a super-Gaussian prior for $\mathbf{w}$ to encourage sparsity and are discussed next.

4.3.1 Maximum a Posteriori (MAP) Based Robust Regression

To simultaneously estimate the regression coefficients $\mathbf{a}$ and the outliers $\mathbf{w}$, we propose to solve the following optimization problem,

$$\hat{\mathbf{a}}, \hat{\mathbf{w}} = \arg\min_{\mathbf{a},\mathbf{w}} \; \|\mathbf{y} - X\mathbf{a} - \mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_p^p, \tag{4.5}$$

where $\|\mathbf{w}\|_p = \left(\sum_{i=1}^{M} |w_i|^p\right)^{\frac{1}{p}}$, $0 < p \leq 1$, and λ is a regularization parameter. This approach can be viewed as an MAP estimation with a super-Gaussian prior distribution $P(w_i) \propto \exp\{-\lambda|w_i|^p\}$. It encourages sparse $\mathbf{w}$ to be recovered. Another closely related algorithm could be immediately obtained as follows,

$$\hat{\mathbf{a}}, \hat{\mathbf{w}} = \arg\min_{\mathbf{a},\mathbf{w}} \; \|\mathbf{w}\|_p, \;\; \text{s.t.} \;\; \|\mathbf{y} - \mathrm{X}\mathbf{a} - \mathbf{w}\|_2 \leq \zeta, \tag{4.6}$$

where ζ is a regularization parameter.

Note that for $p = 1$ these two algorithms are variants of the sparse signal recovery methods, namely Lasso [113] (or basis pursuit denoising [24]) and ℓ_1-regularization problem [18]. By estimating $\mathbf{w}$, these algorithms determine how each observation is contaminated. In contrast, the LAV method, which can be obtained by letting $\zeta \to 0$ in (4.6), assumes a Laplacian prior on the total noise and minimizes the sum of ℓ_1-norm of the fitting errors. As a result, it is not able to clarify the underlying mechanism of noise contamination.

To solve the above optimization problems, (4.5) and (4.6) will become convex optimization problems when $p = 1$ and (4.5) will be considered in the simulation study. Motivated by the analysis in [24, Section 5.2] and our experience, we choose the regularization parameter $\lambda = \frac{\tilde{\sigma}\sqrt{2\log M}}{3}$, where $\tilde{\sigma}$ is a proper estimation of scale.[1] Procedures can be developed for other choices of p [60].

4.3.2 Empirical Bayesian Inference Based Robust Regression

This method adopts the empirical Bayesian approach for robust regression. In particular, we utilize the sparse Bayesian learning methodology developed in [114, 133]. To this end, it is assumed that w_i is a random variable with prior distribution $w_i \sim \mathcal{N}(0, \gamma_i)$, where γ_i is the hyperparameter that controls the variance of each w_i and has to be learnt. If $\gamma_i = 0$, it means the corresponding w_i will be zero, resulting in no anomalous error being added into observation y_i. If $\gamma_i > 0$, an anomalous noise whose

[1] In the experiments presented in Section 4.4, LAV is employed to obtain $\tilde{\sigma}$. Other robust techniques for scale estimation could also be used.

magnitude depends on γ_i will contaminate y_i, and it results in an outlier in the measurement.

To estimate the regression coefficients, we jointly find

$$\hat{\mathbf{a}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}^2 \quad = \quad \arg\max_{\mathbf{a},\boldsymbol{\gamma},\sigma^2} \mathsf{P}(\mathbf{y}|X, \mathbf{a}, \boldsymbol{\gamma}, \sigma^2), \tag{4.7}$$

where $\boldsymbol{\gamma} \triangleq \{\gamma_1, \gamma_2, ..., \gamma_M\}$. Then $\mathbf{w}$ can be estimated by the posterior mean, i.e.

$$\hat{\mathbf{w}} = \mathsf{E}[\mathbf{w}|X, \mathbf{y}, \hat{\mathbf{a}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}^2]. \tag{4.8}$$

Note that the essence of this method is that the robust regression problem is cast into the framework of sparse Bayesian learning (SBL) with appropriate modifications, and this is made possible by our proposed two-component noise modeling technique. The algorithm development, analysis and experimental study of the original SBL for sparse signal recovery have been extensively discussed in [114, 132, 133]. Interested readers are referred to these references for more detail. We summarize the extended SBL based robust regression algorithm, which can be derived using the expectation-maximization (EM) approach, as follows.[2]

Step 1: Initialize $\mathbf{a}_{(0)}$, $\sigma^2_{(0)}$ and $\gamma_{i_{(0)}}$ for $i = 1, 2, ..., M$. Let

$$\Gamma_{(k)} \triangleq \mathrm{diag}(\gamma_{1_{(k)}}, ..., \gamma_{M_{(k)}}).$$

[2] Inspired by the SBL implementations by authors of [114, 133], we suggest that the hyperparameters $\gamma_{i_{(k)}}$ that are smaller than a predefined threshold be pruned from future iterations.

Step 2: At iteration k, compute

$$\hat{\mathbf{w}}_{(k)} = (I + \sigma^2_{(k-1)}\Gamma^{-1}_{(k-1)})^{-1}(\mathbf{y} - X\mathbf{a}_{(k-1)}) \tag{4.9}$$

$$\widehat{\mathbf{W}}_{(k)} = \hat{\mathbf{w}}_{(k)}\hat{\mathbf{w}}^{\intercal}_{(k)} + (\sigma^{-2}_{(k-1)}I + \Gamma^{-1}_{(k-1)})^{-1} \tag{4.10}$$

$$\gamma_{i_{(k)}} = [\widehat{\mathbf{W}}_{(k)}]_{i,i} \tag{4.11}$$

$$\sigma^2_{(k)} = \frac{1}{M}\|\mathbf{y} - X\mathbf{a}_{(k-1)}\|^2_2 + \frac{1}{M}\mathrm{tr}(\widehat{\mathbf{W}}_{(k)}) - \frac{2}{M}(\mathbf{y} - X\mathbf{a}_{(k-1)})^{\intercal}\hat{\mathbf{w}}_{(k)} \tag{4.12}$$

$$\mathbf{a}_{(k)} = (X^{\intercal}X)^{-1}X^{\intercal}(\mathbf{y} - \hat{\mathbf{w}}_{(k)}). \tag{4.13}$$

Step 3: Check for convergence. If convergence criterion is not satisfied, go to Step 2. If it has converged, output $\mathbf{a}_{(k)}$ as the regression coefficients.

To provide an interpretation of this algorithm, at each iteration it first estimates the posterior mean $\hat{\mathbf{w}}_{(k)}$ obtaining the current estimate of the outlier components. Then, it performs an ordinary LS estimation on the corrected data, i.e. $(\mathbf{y} - \hat{\mathbf{w}}_{(k)})$. It is also worthwhile to note that this algorithm can be generalized to general robust regression problems, where the model will be $\mathbf{y} = f(X, \mathbf{a}) + \mathbf{w} + \boldsymbol{\epsilon}$ and f is a general functional relationship assumed on the data. One can similarly derive the updating rule for $\mathbf{a}_{(k)}$ as

$$\text{Find } \mathbf{a}_{(k)} \text{ s.t. } \big(\mathbf{y} - f(X, \mathbf{a}) - \hat{\mathbf{w}}_{(k)}\big)^{\intercal}\frac{\partial f(X, \mathbf{a})}{\partial \mathbf{a}} = 0 \tag{4.14}$$

and use the general function $f(X, \mathbf{a})$ in the algorithm as needed.

4.4 Experiments

4.4.1 Simulated Data Sets

To study the statistical behavior of the proposed algorithms, we consider the multiple linear regression problem as follows,

$$y_i = \sum_{k=1}^{5} a_k x_{k,i} + e_i,\ i \in [M] \tag{4.15}$$

where $\mathbf{a} = [1, 2, -1.5, -3, 2.5]^\intercal$ and $M = 100$. The explanatory variables are independently generated according to $x_{1,i} \sim \mathrm{U}(1, 31), x_{2,i} \sim \mathrm{U}(-200, -150), x_{3,i} \sim \text{Laplacian}(1, 10), x_{4,i} \sim \mathcal{N}(10, 5^2), x_{5,i} \sim \text{Poisson}(10)$.

Consider the following two cases. (i) Symmetric outlier distribution. Let us assume $e_i \sim (1-\delta)\mathcal{N}(0, 0.1^2) + \delta\xi$, where $\xi \sim \mathcal{N}(b, \eta^2)$, b takes value equally likely on $\{-20, 20\}$, $\eta \sim \mathrm{U}(0, 10)$, and δ controls the percentage of outlier contamination. (ii) Asymmetric outlier distribution. We assume $e_i \sim (1-\delta)\mathcal{N}(0, 0.1^2) + \delta\xi$, where $\xi \sim \mathcal{N}(-20, \eta^2)$, and $\eta \sim \mathrm{U}(0, 10)$. For each case, two different levels of outlier contamination will be considered, namely $\delta = 5\%$ and $\delta = 30\%$.

The MAP based robust regression algorithm (4.5) in Section 4.3.1 (denoted by **Alg1**) and the empirical Bayesian inference based algorithm in Section 4.3.2 (denoted by **Alg2**) are used to compute the regression coefficients. Additionally, the following algorithms will also be employed to compute the regression coefficients: (i) Least Absolute Value (**LAV**); (ii) Least Trimmed Squares (**LTS**), where 75% of the squared errors are kept; (iii) M-estimate (**M**), where Huber's function is employed; (iv) Gaussian Mixture model for noise (**GM**), (v) Student's t-distribution for noise (**Stu-t**), i.e. $P(e_i|\nu,\theta) \propto [1+\theta e_i^2/\nu]^{-(\nu+1)/2}$ with ν is fixed to be 3.

For each algorithm 5000 random data sets are processed. The performances are compared in terms of the empirical bias and the empirical variance of the estimate of each regression coefficient. The results are shown in Figures 4.1 and 4.2. The subplot (a) and (b) in each figure correspond to the empirical bias and the empirical variance, respectively, for 5% outlier contamination. Subplot (c) and (d) are for 30% outlier contamination.

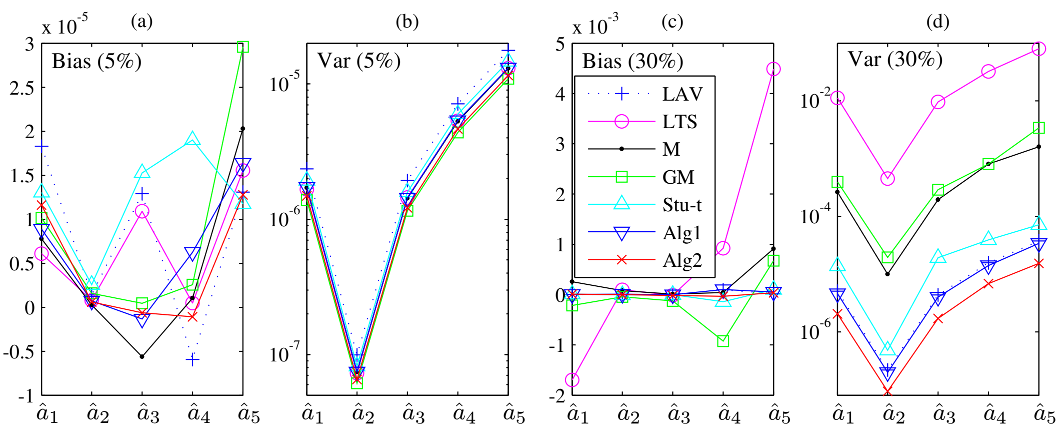


Figure 4.1: Empirical bias and variance (Symmetric outlier case).

To analyze the results, first, our proposed algorithms, especially Alg2, show consistent performance with lower bias and lower variance in most cases. Second, our methods tend to serve as feasible algorithmic choices for a large range of percentage of outlier contamination, as it works well with small (5%) and large (30%) portions of outliers. Third, the empirical Bayesian algorithm (Alg2) actually outperforms the MAP type algorithm (4.5) (Alg1) in most cases. This observation echoes the fact that

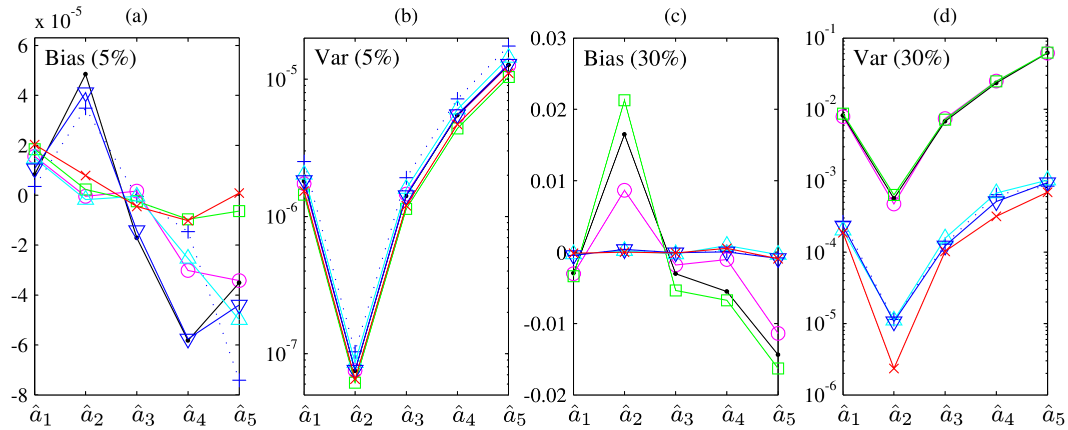


Figure 4.2: Empirical bias and variance (Asymmetric outlier case. Legends are the same with Figure 4.1.)

empirical Bayesian inference could perform better since the posterior mean of the hyperparameter is more representative for the posterior probability mass [78].

4.4.2 Brownlee's Stackloss Data Set

This data set, which has been extensively studied in statistical literature (cf. [103, pp.76]), contains 21 four-dimensional observations regarding the operation of a plant for the oxidation of ammonia to nitric acid. We select several algorithms for linear regression. Following the methodology in [103, Chapter 3], the index plots associated with different algorithms are shown in Figure 4.3.

As in Figure 4.3, our algorithms exhibit very similar results to the LAV method for this data set. Observations 1, 3, 4, and 21 can be identified as outliers, which is consistent with existing analyses on this data set [75,103]. Specifically, for each explanatory variable (e.v.), the estimated coefficient ($\hat{a}_i$), its standard error (std) and t-value (t-v) are tabulated for LS and Alg1 in Table 4.1.

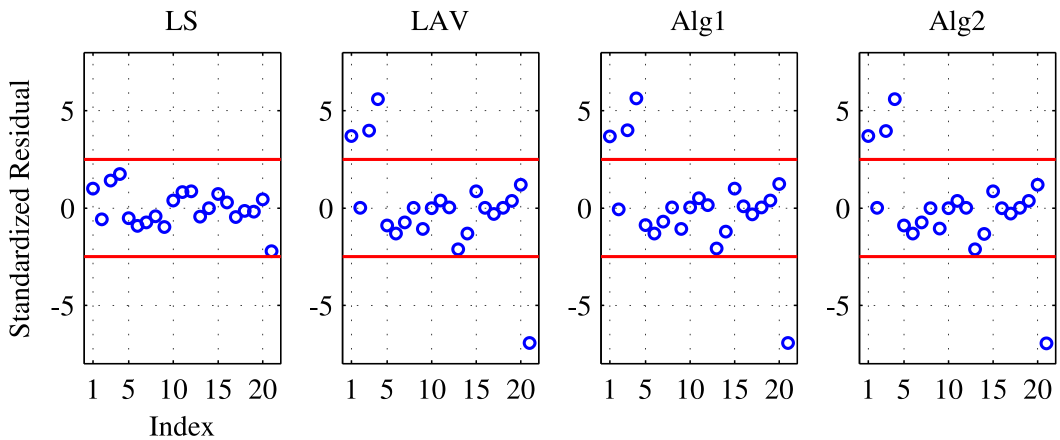


Figure 4.3: Index plots for different regression algorithms. The interval $[-2.5, 2.5]$ is marked by red lines for inspecting outliers.

Table 4.1: Regression results for LS and Alg1.

	LS			Alg1		
e.v.	$\hat{a}_i$	std	t-v	$\hat{a}_i$	std	t-v
rate	0.716	0.135	5.307	0.834	0.073	11.46
temp.	1.295	0.368	3.520	0.596	0.180	3.330
acid.	-0.152	0.156	-0.973	-0.071	0.067	-1.062
const.	-39.92	11.90	-3.356	-39.47	5.105	-7.732

The LS estimation identifies that regression coefficient for acid concentration (acid.) is not significantly different from zero at 5% level. This is confirmed by Alg1 (and also by LAV and Alg2).[3] By proper treatment of outliers, the significance of the rate and the constant term (const.) are enhanced by Alg1 (also LAV and Alg2), and narrower confidence intervals can be constructed. Combining the analysis based on the index plots and the test statistics, we can conclude that the robust regression methods could be more trustworthy in revealing the underlying pattern in the data set.

[3]For LS, use $t_{17,0.05} = 2.11$. For LAV, Alg1/2, use $t_{13,0.05} = 2.16$.

4.4.3 Bupa Liver Data Set

This data set [12] mainly contains blood test results regarding liver function for 345 patients. We focus on the task of using the AST and γGT levels to linearly predict the ALT level. The data is processed by log-transformation [11]. Results are shown in the Table 4.2.

Table 4.2: Regression results for Bupa liver data set.

	LS			LTS		
e.v.	$\hat{a}_i$	std	t-v	$\hat{a}_i$	std	t-v
AST	0.693	0.066	10.57	0.788	0.061	12.87
γGT	0.204	0.030	6.871	0.207	0.026	7.794
const.	0.425	0.176	2.408	0.145	0.166	0.870

	Alg1			Alg2		
e.v.	$\hat{a}_i$	std	t-v	$\hat{a}_i$	std	t-v
AST	0.736	0.062	11.92	0.756	0.062	12.24
γGT	0.202	0.027	7.557	0.208	0.027	7.796
const.	0.306	0.168	1.821	0.220	0.168	1.308

The LS estimation indicates that all variables and the constant term (intercept) are significant at 5% level. However, the robust methods we tested here inform us that the constant term is actually not significant.[4] Having this observation, we re-perform the linear regression without constant term. To examine the difference, we perform the F-test with null hypothesis H_0 : the constant term is equal to zero. The results are summarized in the Table 4.3.

Table 4.3: F-values of different algorithms.

	LS	LTS	Alg1	Alg2
F-value	5.801	1.995	2.708	1.959

We can see that except LS, the F-values for all the robust regression methods

[4]Specifically, $df_{\text{LTS}} = 331$, $df_{\text{Alg1}} = df_{\text{Alg2}} = 332$. Hence, $t_{\infty,0.05} = 1.96$, $F_{1,\infty,0.05} = 3.84$ are used.

indicate that the constant term is not significant at 5% level, which confirms our earlier observation. To be complete, the simpler linear models learned by different algorithms are given in Table 4.4.

Table 4.4: Final regression results.

	LTS		Alg1		Alg2	
e.v.	$\hat{a}_i$	t-v	$\hat{a}_i$	t-v	$\hat{a}_i$	t-v
AST	0.835	29.28	0.830	28.85	0.834	28.98
γGT	0.205	7.752	0.204	7.630	0.202	7.555

Based on experiments on these real data sets, we conclude that the proposed algorithms exhibit consistent results compared to existing robust regression algorithms and demonstrate their usefulness for robust regression.

4.5 Acknowledgements

Chapter 4 is, in part, a reprint of the material published as "Algorithms for Robust Linear Regression by Exploiting the Connection to Sparse Signal Recovery," Y. Jin, B. D. Rao, *International Conference on Acoustics, Speech, and Signal Processing*, March 2010. I was the primary author, and B. D. Rao supervised the research.

Chapter 5

LMS Type Adaptive Filtering Algorithms that Incorporate Sparsity

Thus far, the algorithms we considered are based on the batch estimation techniques, which are performed on blocks of measurements. There is a need for simpler, online learning algorithms that can deal with non-stationary environment. This gives rise to the motivation of adaptive algorithms. In this chapter, we consider the development of adaptive algorithms that can exploit the sparsity structure of the system. We show how to systematically transform the techniques for sparse signal recovery to facilitate adaptive filtering applications.

5.1 Introduction

Adaptive filters have many potential applications and have been a topic of much study and interest [68, 83, 106]. We develop and study adaptive filtering algorithms that attempt to impose and take advantage of the sparsity structure in the filter coefficients. At the center of the problem is the estimation of the underlying sparse predictor. This problem arises in a wide spectrum of applications such as echo cancellation [41, 85], channel equalization [97], noise cancellation [127], wireless communication [29], speech processing [27, 104], radar signal processing [109], and image processing [45, 58].

The recent progress in the area of sparse signal recovery makes available a plethora of approaches for potentially tackling the problem of adaptive filters that promote sparsity. The problem of sparse signal recovery aims at recovering a sparse signal based on only a small number of linear measurements. Computationally efficient algorithms have been proposed to find or approximate the sparse solution in various settings, which are introduced in Section 1.1.1. To leverage these techniques for sparse predictor estimation, for example, Cotter and Rao [29] employed the matching pursuit algorithm to estimate the sparse impulse response in the scenario of wireless communication. Bajwa, Haupt, Raz, and Nowak [5] proposed to use the Dantzig selector [15] to estimate a sparse communication channel.

The development of adaptive algorithms that exploit sparsity is a natural complement to the above mentioned batch estimation techniques. Adaptive filters provide the opportunity of adaptively adjusting the predictor coefficients based on the current input signal and the previous predictor states [106]. Compared to filters with fixed coefficients, adaptive filters have the ability to learn the signal statistics and to keep track of potential variations. In contrast to batch processing methods where space is needed to store enough amount of samples and processing power is demanded for expensive operations such as matrix inversion, adaptive filters implement less expensive recursive updates that drastically reduce the storage and computation burdens. As examples, the least-mean-squares (LMS) algorithm [126] and the normalized LMS (NLMS) algorithm [86] are well known adaptive filters employed in many practical applications.

Designing adaptive filters that can exploit the sparsity structure in the underlying predictor to achieve better performance has been an area of recent interest [3, 7, 10, 25, 33, 35, 41, 54, 61, 84, 101]. Of particular interest is the proportionate NLMS (PNLMS) algorithm introduced by Duttweiler [41] to deal with the sparse channel estimation that arises in echo cancellers. Although this algorithm was not formally derived by minimizing an underlying objective function, it was well motivated with theoretic analysis and simulations to support its effectiveness. To address the issue of slow convergence of PNLMS at the late stages of adaptation, various methods were suggested for improvement, such as PNLMS++ [54], improved PNLMS (IPNLMS) [7], and improved IPNLMS (IIPNLMS) [33]. Experimental results showed that certain drawback of PNLMS can be alleviated via these variants of PNLMS. Motivated by the techniques in sparse signal recovery, Chen, Gu, and Hero derived the zero attracting LMS (ZA-LMS) and the reweighted zero attracting LMS (RZA-LMS) [25, 26], of which the optimization criteria are motivated by some popular diversity measures that are minimized by the batch algorithms to find sparse solutions. Using the approximations of ℓ_0 norm as a diversity measure, Gu, Jin, and Mei proposed the ℓ_0-LMS and ℓ_0-NLMS algorithms that are capable of estimating sparse predictors [61].

The goal of this work is to further develop the field of adaptive filtering algorithms that incorporate sparsity considerations by developing alternate LMS type adaptive filtering algorithms that can benefit from the sparsity structure underlying the predictor. Motivated by the techniques for sparse signal recovery, we develop a general mathematical framework for deriving a class of adaptive algorithms that are suitable for sparse predictors. This framework can potentially benefit from many existing techniques for sparse signal recovery by systematically transforming them to develop the corresponding adaptive algorithms for sparse predictors. The key characteristic of this framework is that the optimization procedure employs the affine scaling method, and the algorithms perform the optimization in the affine scaling domain. This is different from the derivations of many existing adaptive algorithms in which the optimization is carried out in the original coefficient domain. The affine scaling method belongs to the family of interior-point methods which have gained considerable popularity for solving convex and nonlinear optimization problems. It involves a centering operation which has the advantage of transforming the original optimization problem into an equivalent one in which the current point is favorably positioned for a constrained form of the steepest-descent method [87]. Therefore, the proposed framework leads to algorithms that can support larger learning steps along the search direction, and the overall adaptation can be sped up. To demonstrate the potential of this algorithmic framework, we study in detail the pALMS and the pANLMS adaptive filters, which are instantiations of this framework and closely related to FOCUSS [60] in sparse signal recovery. The goal of this work is very similar to that in [84] wherein also a general class of sparsity promoting adaptive filter algorithms are developed. The approach employed there is based on the natural gradient and quite different from the interior-point method motivation and approach in this work. More discussion on [84] will be presented in later sections.

To avoid confusion, it is worth noting that the affine scaling method employed in our approach fundamentally differs from the methods of affine combination of multiple adaptive filters [9] and the affine projection algorithms (APA) [55, 91]. In our approach, the usage of affine projection originates from an optimization viewpoint, since we wish to take advantage of the centering operation in interior-point methods to achieve faster convergence. In contrast, the affine combination of multiple adaptive filters suggests affine combination of the outputs of differently parameterized filters to obtain performance better than each individual filter. APA can be viewed as a multi-dimensional extension of NLMS. It is set to match a vector of desired signals in order to obtain faster convergence. The involvement of affine projection in this approach is a necessary consequence of the multi-dimensional model, rather than an optimization alternative.

5.2 Problem Formulation and Background

5.2.1 Problem Formulation

Let us assume that $\mathbf{x}[i] \in \mathbb{R}^M$ is a wide sense stationary vector random process with $R \triangleq \mathsf{E}[\mathbf{x}[i]\mathbf{x}[i]^\intercal]$. In the case of a time series problem $\mathbf{x}[i]$ is formed using samples from a window of observations. Let $y[i] \in \mathbb{R}$ be the desired signal and it is assumed to be jointly wide sense stationary with $\mathbf{x}[i]$. The cross-correlation is denoted by $\mathbf{d}$, and is given by $\mathbf{d} \triangleq \mathsf{E}[\mathbf{x}[i]y[i]]$. The goal of adaptive filtering is to learn the optimal predictor coefficient vector and to obtain a linear estimate of the desired signal $y[i]$, which is given by

$$\hat{y}[i] = \mathbf{c}[i]^\intercal \mathbf{x}[i] \tag{5.1}$$

where $\mathbf{c}[i]$ is the estimate of the true predictor coefficient vector $\mathbf{c}_0$ at the ith iteration. We are interested in the scenario where the predictor coefficient vector $\mathbf{c}_0$ is sparse, i.e., there are many zero coefficients in $\mathbf{c}_0$ and most of the energy is concentrated on only a few taps. In this work, we assume all variables are real for simplicity and the development can be readily extended to complex domain.

5.2.2 Background on Adaptive Filters Exploiting Sparsity

We focus on the LMS type filters. Usually, the starting point for the derivation of an ordinary LMS filter is to consider the following linear prediction problem [83]

$$\min_{\mathbf{c}} J(\mathbf{c}), \quad \text{where } J(\mathbf{c}) = \mathsf{E}\,|y[i] - \mathbf{c}^\mathsf{T}\mathbf{x}[i]|^2 \tag{5.2}$$

which minimizes the mean squared error in predicting the desired signal. Note that (5.2) does not incorporate any information about the sparsity structure of the predictor. To impose the explicit objective of enforcing sparsity on the solution vector, one may wish to minimize the modified cost function

$$J(\mathbf{c}) = \mathsf{E}\,|y[i] - \mathbf{c}^\mathsf{T}\mathbf{x}[i]|^2 + \beta \cdot g(\mathbf{c}) \tag{5.3}$$

where $\beta \geq 0$ is a regularization parameter. This cost function consists of two terms. The first term in (5.3), as before, measures the quality of prediction. The second term $g(\mathbf{c})$ can be chosen as a proper diversity measure whose minimization leads to a sparse $\mathbf{c}$. A popular diversity measure for sparse signal recovery is the ℓ_1 norm of $\mathbf{c}$, i.e., $\|\mathbf{c}\|_1$. The regularization parameter β balances these two objectives in the cost function in order to obtain reasonable results.

To proceed, we expand $J(\mathbf{c})$ and obtain the gradient as

$$\nabla_{\mathbf{c}} J(\mathbf{c}) = -2\mathbf{d} + 2R\mathbf{c} + \beta \cdot \nabla_{\mathbf{c}} g(\mathbf{c}).$$

This gives us the following steepest descent algorithm for minimizing (5.3)

$$\mathbf{c}^{(j+1)} = \mathbf{c}^{(j)} - \mu\nabla_{\mathbf{c}} J(\mathbf{c}) = \mathbf{c}^{(j)} - \mu(-2\mathbf{d} + 2R\mathbf{c}^{(j)} + \beta \cdot \nabla_{\mathbf{c}} g(\mathbf{c}^{(j)})).$$

where $\mu > 0$ is the step size parameter. In practice, an LMS type algorithm replaces the statistical quantities $\mathbf{d}$ and R by their instantaneous estimates, i.e., $\mathbf{x}[i]y[i]$ and $\mathbf{x}[i]\mathbf{x}[i]^{\intercal}$, respectively. The resulting adaptive filtering algorithm is given by

$$\mathbf{c}[i+1] = \mathbf{c}[i] + 2\mu\mathbf{x}[i]e[i] + \mu\beta \cdot \nabla_{\mathbf{c}} g(\mathbf{c}[i])) \tag{5.4}$$

$$e[i] = y[i] - \mathbf{c}[i]^{\intercal}\mathbf{x}[i]. \tag{5.5}$$

To instantiate the algorithm, one should choose an appropriate diversity measure $g(\mathbf{c})$ and obtain the corresponding adaptive filter.

As an example, let us choose $g(\mathbf{c}) = \|\mathbf{c}\|_1$, the popular ℓ_1 norm diversity measure [24]. Note that $\nabla_{c_k} g(\mathbf{c}) = \text{sign}(c_k)$. Thus, according to the procedure above, we obtain the following adaptive algorithm

$$\mathbf{c}[i+1] = \mathbf{c}[i] + 2\mu\mathbf{x}[i]e[i] + \mu\beta \cdot \text{sign}(\mathbf{c})$$

where $\text{sign}(\mathbf{c})$ is understood elementwise, and $e[i]$ is calculated from (5.5). This algorithm is indeed the zero attracting LMS (ZA-LMS) derived by Chen, Gu, and Hero [25, 26]. Motivated by the iteratively reweighted ℓ_1 minimization [20], the reweighted zero attracting LMS (RZA-LMS) [25, 26] was derived by letting $g(\mathbf{c}) = \sum_{k=1}^{M} \log(1 +$

$|c_k|/\epsilon)$. With the selection $g(\mathbf{c}) = \sum_{k=1}^{M}(1-\exp(-\eta|c_k|))$, which can be viewed an approximation of the ℓ_0 norm, the resultant adaptive filter can be recognized as the ℓ_0-LMS proposed by Gu, Jin, and Mei [61].

A class of LMS variants, proposed by Martin, Sethares, Williamson, and Johnson [84], uses a natural gradient framework to deduce adaptive filters that can exploit the sparsity structure. Of practical interest is the approximate natural gradient (ANG) algorithm, which is characterized by a cost function $H(\mathbf{c})$, a reparameterization function $c_k = F(z_k)$, and a prior distribution of the unknown parameters $\phi_k(z_k)$. Note that the prior distribution reflects the prior knowledge of z_k, and it can be improper, i.e., the integral over the entire domain is not required to be one. Then, the update equation in the $\mathbf{z}$ domain is [79, 84]

$$z_k[i+1] = z_k[i] - \mu\frac{\partial H(\mathbf{c}[i])}{\partial z_k}\frac{1}{\phi_k^2(z_k[i])}. \tag{5.6}$$

For example, by letting $F(z_k) = z_k$ and $\phi_k(z_k) = 1$ for $k \in [M]$, (5.6) recovers the LMS algorithm. Proper choices of $H(\mathbf{c})$, $F(z_k)$, and $\phi_k(z_k)$ can lead to adaptive algorithms that exploit the sparsity structure in the prediction coefficients.

Besides the LMS type adaptive algorithms, normalized LMS (NLMS) type algorithms are also developed for exploiting the sparse nature of the predictor. The proportionate NLMS (PNLMS) algorithm proposed by Duttweiler [41] employs the following update equations

$$\mathbf{c}[i+1] = \mathbf{c}[i] + \mu \frac{W^2(\mathbf{c}[i])}{\frac{1}{M}\sum_{k=1}^{M}[W^2(\mathbf{c}[i])]_{k,k}} \frac{1}{\|\mathbf{x}[i]\|_2^2}\mathbf{x}[i]e[i] \tag{5.7}$$

$$W^2(\mathbf{c}[i]) = \mathrm{diag}(w_k) \tag{5.8}$$

$$w_k = \max(\rho_{\text{PNLMS}} d_k, |c_k[i]|) \tag{5.9}$$

$$d_k = \max(\delta_{\text{PNLMS}}, |c_1[i]|, ..., |c_M[i]|) \tag{5.10}$$

where $\rho_{\text{PNLMS}}, \delta_{\text{PNLMS}} > 0$ are regularization parameters. Different from NLMS, PNLMS introduces a weighting factor of roughly $\frac{M|c_k[i]|}{\sum_{k=1}^{M}|c_k[i]|}$ to differently scale the update for different taps. The intuition behind this modification is that when $c_k[i]$ has large magnitude, it is likely that $c_k[i+1]$ could also be large. Thus, such a weighting scheme rewards large taps with more adaptation energy and drives them to convergence faster than the taps with smaller magnitudes. Although this algorithm was not derived formally by minimizing an underlying objective function, it is well motivated and analysis were provided to support its effectiveness. Experiments are shown to demonstrate that the PNLMS algorithm does exploit the sparsity structure. It converges faster than NLMS at early stages, but it may converge slower than NLMS at the later stages of the adaptation [54], due to the observation that taps with small magnitudes typically receive less adaptation energy than their NLMS counterparts. To take advantages of both NLMS and PNLMS at different stages, a series of work have proposed various methods to combine these two filters for better performance. For example, PNLMS++ [54] suggested using

NLMS update and PNLMS update alternatively for different sample periods. Improved PNLMS (IPNLMS) [7] proposed a weighted combination of NLNS and PNLMS updates via the following update

$$\mathbf{c}[i+1] = \mathbf{c}[i] + \mu \frac{W^2(\mathbf{c}[i])}{\mathbf{x}[i]^\intercal W^2(\mathbf{c}[i])\mathbf{x}[i] + \delta_{\text{IPNLMS}}} \mathbf{x}[i]e[i] \tag{5.11}$$

$$W^2(\mathbf{c}[i]) = \operatorname{diag}(w_k) \tag{5.12}$$

$$w_k = \frac{1-\alpha_{\text{IPNLMS}}}{2M} + (1+\alpha_{\text{IPNLMS}}) \frac{|c_k[i]|}{2\|\mathbf{c}[i]\|_1 + \epsilon_{\text{IPNLMS}}} \tag{5.13}$$

where $\alpha_{\text{IPNLMS}} \in [-1, 1)$ controls the weights between NLMS and PNLMS, and δ_{IPNLMS}, $\epsilon_{\text{IPNLMS}} > 0$ are regularization parameters. The improved IPNLMS (IIPNLMS) [33] proposed more complex weighting scheme between NLMS and PNLMS which weighs active and inactive taps differently. Although these algorithms are heuristically motivated, simulations show that they exhibit faster convergence than NLMS when applied to sparse predictor estimation.

5.3 An Algorithmic Framework for Adaptive Filtering with Sparsity Concerns

5.3.1 Prototype LMS Adaptive Filtering Algorithm

We derive a general LMS type adaptive algorithm that has the capability to exploit the sparsity structure underlying the prediction problem. We start with the cost function (5.3). Assume $\nabla_{\mathbf{c}} g(\mathbf{c})$ can be written as

$$\nabla_{\mathbf{c}} g(\mathbf{c}) = \alpha(\mathbf{c}) \cdot \Pi(\mathbf{c}) \cdot \mathbf{c} \tag{5.14}$$

where $\alpha(\mathbf{c}) \in \mathbb{R}^{+}$, and $\Pi(\mathbf{c}) \in \mathbb{R}^{M \times M}$ is a diagonal matrix with positive diagonal elements. This can be shown to hold for a wide variety of diversity measures used for sparse signal recovery [93]. Thus

$$\nabla_{\mathbf{c}} J(\mathbf{c}) = -2\mathbf{d} + 2R\mathbf{c} + \beta\alpha(\mathbf{c})\Pi(\mathbf{c})\mathbf{c}.$$

By setting $\nabla_{\mathbf{c}} J(\mathbf{c}) = \mathbf{0}$, the optimal solution $\mathbf{c}_*$ should satisfy

$$\begin{aligned} \mathbf{c}_* &= \left(R + \frac{1}{2}\beta\alpha(\mathbf{c}_*)\Pi(\mathbf{c}_*)\right)^{-1} \mathbf{d} \\ &= W(\mathbf{c}_*) \left(W(\mathbf{c}_*)RW(\mathbf{c}_*) + \frac{1}{2}\beta\alpha(\mathbf{c}_*)I\right)^{-1} W(\mathbf{c}_*)\mathbf{d} \end{aligned} \tag{5.15}$$

where $\Pi(\mathbf{c}) \equiv W^{-2}(\mathbf{c})$. Note that (5.15) suggests the following iterative procedure for computing $\mathbf{c}_*$

$$\mathbf{c}^{(j+1)} = W(\mathbf{c}^{(j)}) \left(W(\mathbf{c}^{(j)})RW(\mathbf{c}^{(j)}) + \frac{1}{2}\beta\alpha(\mathbf{c}^{(j)})I\right)^{-1} W(\mathbf{c}^{(j)})\mathbf{d} \tag{5.16}$$

with some reasonable initialization $\mathbf{c}^{(0)}$.

Next, at the jth iteration, for $\mathbf{c} \in \mathbb{R}^{M}$, we define the corresponding affinely scaled variable

$$\mathbf{q}(\mathbf{c}) \triangleq W(\mathbf{c}^{(j)})^{-1}\mathbf{c} \tag{5.17}$$

and the auxiliary variables $R^{(j)} \triangleq W(\mathbf{c}^{(j)})RW(\mathbf{c}^{(j)})$ and $\mathbf{d}^{(j)} \triangleq W(\mathbf{c}^{(j)})\mathbf{d}$. Such transformations were first used in the work of Karmarkar in the development of algorithms for solving linear programming problems [72]. It was viewed as a mechanism to transform the problem into variables where the current estimate of the transformed variable was at the center of the feasible region. With the transformation into the affine scaling domain, the update (5.16) can be rewritten as

$$\mathbf{q}^{(j+1)} \triangleq \mathbf{q}(\mathbf{c}^{(j+1)}) = \left(R^{(j)} + \frac{1}{2}\beta\alpha(\mathbf{c}^{(j)})I\right)^{-1}\mathbf{d}^{(j)}. \tag{5.18}$$

One can observe that (5.18) is actually the minimizer of the following quadratic cost function

$$J(\mathbf{q}) = \mathsf{E}\left|y[i] - \mathbf{q}^{\mathsf{T}}W(\mathbf{c}^{(j)})\mathbf{x}[i]\right|^2 + \frac{1}{2}\beta\alpha(\mathbf{c}^{(j)})\|\mathbf{q}\|_2^2. \tag{5.19}$$

We can also view the approach as belonging to the class of iteratively reweighted least squares approaches. Note that

$$\nabla_{\mathbf{q}}J(\mathbf{q}) = -2\mathbf{d}^{(j)} + 2R^{(j)}\mathbf{q} + \beta\alpha(\mathbf{c}^{(j)})\mathbf{q}.$$

Then, we can obtain a steepest descent algorithm (SDA) which replaces (5.18),

$$\mathbf{q}^{(j+1)} = \mathbf{q}^{(j)} - \frac{\mu}{2}\nabla_{\mathbf{q}} J(\mathbf{q}^{(j)}) \tag{5.20}$$

$$= \left(1 - \frac{1}{2}\mu\beta\alpha(\mathbf{c}^{(j)})\right)\mathbf{q}^{(j)} + \mu\mathbf{d}^{(j)} - \mu R^{(j)}\mathbf{q}^{(j)} \tag{5.21}$$

where $\mu > 0$ is the step size parameter. According to (5.17), we can transform (5.21) back to the original coefficient domain as follows

$$\mathbf{c}^{(j+1)} = \left(1 - \frac{1}{2}\mu\beta\alpha(\mathbf{c}^{(j)})\right)\mathbf{c}^{(j)} + \mu W^2(\mathbf{c}^{(j)})\mathbf{d} - \mu W^2(\mathbf{c}^{(j)})R\mathbf{c}^{(j)}. \tag{5.22}$$

The following theorem sheds light on the convergence of the algorithm (5.22) and its relationship to the cost function (5.3). The proof is presented in Section 5.7.1.

Theorem 9. *Let $g(\mathbf{c})$ be a diversity measure with the following properties:*

1. $g : \mathbb{R}^M \to \mathbb{R}$ *is separable, i.e.,* $g(\mathbf{c}) = \sum_{k=1}^{M} g_k(c_k)$*;*

2. $g_k(c) = g_k(-c)$*;*

3. $g_k(c)$ *is increasing with* $|c|$*;*

4. $g_k(c)$ *is concave in* c^2*.*

Then, there exists a sequence of $\{\mu^{(k)}\}_{k=1}^{\infty}$ such that the algorithm (5.22), using $\mu^{(k)}$ at the kth iteration, converges to a local minimum of (5.3).

To derive the adaptive algorithm, we follow the usual steps of replacing R and $\mathbf{d}$ by their instantaneous estimates, i.e., $\mathbf{x}[i]\mathbf{x}[i]^\intercal$ and $\mathbf{x}[i]y[i]$, respectively. Set $j = i$ in

(5.22). We obtain the update equation

$$\mathbf{c}[i+1] = \left(1 - \frac{1}{2}\mu\beta\alpha(\mathbf{c}[i])\right)\mathbf{c}[i] + \mu W^2(\mathbf{c}[i])\mathbf{x}[i]y[i] - \mu W^2(\mathbf{c}[i])\mathbf{x}[i]\mathbf{x}[i]^\intercal\mathbf{c}[i]. \tag{5.23}$$

Therefore, by absorbing the factor $\frac{1}{2}$ into β, the resulting prototype LMS algorithm is

$$\mathbf{c}[i+1] = (1 - \mu\beta\alpha(\mathbf{c}[i]))\mathbf{c}[i] + \mu W^2(\mathbf{c}[i])\mathbf{x}[i]e[i] \tag{5.24}$$

$$W^2(\mathbf{c}[i]) = \Pi(\mathbf{c}[i])^{-1} \tag{5.25}$$

and $e[i]$ is computed according to (5.5).

5.3.2 Relation to ANG Algorithms

As introduced in Section 5.2.2, a class of approximate natural gradient (ANG) algorithms for adaptive filtering has been developed using a natural gradient framework [84]. We discuss the relation between the ANG algorithms and the prototype LMS algorithm in Section 5.3.1.

Equipped with the cost $H(\mathbf{c}) = (y[i] - \hat{y}[i])^2$, the ANG algorithm can be actually cast into the affine scaling framework. To see this, let us assume the ANG algorithm adopts some reparametrization function $c_k = F(z_k)$, which is assumed to be invertible and differentiable except possibly at a finite number of points, and some prior $\phi_k(z_k)$. First, according to Proposition IV.1 in [84], if there exist some other reparametrization function $c_k = F_1(z_k)$ and prior $\psi_k(z_k)$ such that

$$\left(\frac{\partial F(z_k)}{\partial z_k}\right)^2 \cdot \left(\frac{1}{\phi_k(z_k)}\right)^2 \Bigg|_{z_k=F^{-1}(c_k)} = \left(\frac{\partial F_1(z_k)}{\partial z_k}\right)^2 \cdot \left(\frac{1}{\psi_k(z_k)}\right)^2 \Bigg|_{z_k=F_1^{-1}(c_k)}$$

for $k \in [M]$, then $\{F, \phi_k\}$ and $\{F_1, \psi_k\}$ yield the same ANG algorithm in the $\mathbf{c}$ domain, provided that the step size parameter μ is sufficiently small. Hence, we can choose $c_k = F_1(z_k) \triangleq z_k$, and $\psi_k(z_k)$ such that

$$\psi_k^2(z_k) = \left[\left(\frac{\partial F(z_k)}{\partial z_k}\right)^2 \cdot \left(\frac{1}{\phi_k(z_k)}\right)^2 \Bigg|_{z_k=F^{-1}(c_k)}\right]^{-1} \Bigg|_{c_k=F_1(z_k)}$$

to obtain the same adaptive algorithm. We use $\{F_1, \psi_k\}$ in the discussion to follow. Based on (33) of [84] and the discussion preceding it, we have the update equation in the coefficient domain

$$\mathbf{c}[i+1] = \mathbf{c}[i] - \mu D[i]\mathbf{x}[i] \frac{\partial (y[i] - \hat{y}[i])^2}{\partial \hat{y}[i]} \Bigg|_{\hat{y}[i]=\mathbf{c}[i]^\intercal \mathbf{x}[i]}$$

where[1]

$$\begin{aligned} D[i] &\triangleq \operatorname{diag}\left(\left(\frac{\partial F_1(z_k)}{\partial z_k}\right)^2 \cdot \left(\frac{1}{\psi_k(z_k)}\right)^2 \Bigg|_{z_k=F_1^{-1}(c_k[i])}\right) \\ &= \operatorname{diag}\left(\psi_k^{-2}(z_k)\Big|_{z_k=F_1^{-1}(c_k[i])}\right). \end{aligned} \tag{5.26}$$

[1] Note that in [84] $D[i]$ is defined as an ordinary matrix. However, the $\mathbf{c}$ domain update only uses its diagonal elements, as indicated by the equation preceding (33) therein.

This equation above naturally suggests the diagonal weighting matrix $W^2(\mathbf{c}[i]) \equiv D[i]$. Then, the corresponding choice for $\Pi(\mathbf{c})$ can be found as

$$\Pi(\mathbf{c}) = W^{-2}(\mathbf{c}) = \text{diag}\left(\left[\left(\frac{\partial F(z_k)}{\partial z_k}\right)^2 \cdot \left(\frac{1}{\phi_k(z_k)}\right)^2 \Bigg|_{z_k=F^{-1}(c_k)}\right]^{-1}\right).$$

Thus, we can construct a corresponding diversity measure $g(\mathbf{c}) \triangleq \sum_{k=1}^{M} g_k(c_k)$ where

$$g_k(c_k) \triangleq \int c_k \cdot \left[\left(\frac{\partial F(z_k)}{\partial z_k}\right)^2 \cdot \left(\frac{1}{\phi_k(z_k)}\right)^2 \Bigg|_{z_k=F^{-1}(c_k)}\right]^{-1} dc_k \qquad (5.27)$$

is an indefinite integral.

As a result, an ANG based adaptive filtering algorithm can be recognized as an affine scaling algorithm developed using the diversity measure $g(\mathbf{c})$ defined in (5.27) under the limiting case $\beta = 0$. Further, the interpretation of the ANG based algorithms through the affine scaling perspective indicates that $[W^2(\mathbf{c})]_{k,k}$ is not a function of c_j for $j \neq k$. Consequently, $g(\mathbf{c})$ exhibits a separable form indicating that it can be written as the sum of the costs on individual variables. In contrast, the affine scaling framework proposed in this paper indeed permits the potential of employing more general $g(\mathbf{c})$ which cannot be written in a separate manner. This is important because there are sparse recovery methods such as the sparse Bayesian learning that employ nonseparable diversity measures [114, 129, 130, 133]. A broader class of adaptive filters could be derived using the affine scaling methodology.

5.3.3 Prototype NLMS Adaptive Filtering Algorithm

We can develop a general NLMS type algorithm using the affine scaling transform. The derivation is similar to that of the NLMS, and is presented in Section 5.7.2. We directly state the proposed NLMS type adaptive algorithm as follows

$$\mathbf{c}[i+1] = \mathbf{c}[i] + \frac{\mu}{\mathbf{x}[i]^{\intercal} W^2(\mathbf{c}[i])\mathbf{x}[i]} W^2(\mathbf{c}[i])\mathbf{x}[i]e[i] \tag{5.28}$$

$$W^2(\mathbf{c}[i]) = \Pi(\mathbf{c}[i])^{-1} \tag{5.29}$$

where $\mu > 0$ is the step size parameter.

5.3.4 Steady-State Performance Analysis of Prototype Algorithms

The analysis of adaptive filters is quite complicated because of the nonlinear nature of the update equations and the long-term dependence on the data. In spite of this difficulty, simple and intuitive analysis methods have been quite useful in understanding the behavior of adaptive algorithms. For the purpose of performance analysis, we introduce several useful assumptions. Although these assumptions may seem restrictive, they make meaningful analysis possible without significant loss of insight and are also commonly adopted in performance analysis of adaptive filters. The following signal model is employed for performance analysis:

$$y[i] = \mathbf{c}_0^{\intercal}\mathbf{x}[i] + e_0[i]. \tag{5.30}$$

In addition to the assumptions in Section 5.2.1, we further assume $e_0[i] \in \mathbb{R}$ to be the noise, which is i.i.d. according to $\mathcal{N}(0, \sigma_e^2)$.

Assumption 1: The input data vector $\mathbf{x}[i]$ is independent of $\mathbf{x}[j]$ for $i \neq j$. Further, $\mathbf{x}[i]$ is independent of $e_0[j]$ for all i, j.

Remark. Note that in the scenario of where the input signal is a time series, an input data vector is actually a specific section of this time series. In this case, Assumption 1 will not hold for $\mathbf{x}[i]$ and $\mathbf{x}[i-k]$ for $k \in [M-1]$, since they share an overlap of $(M-k)$ elements. However, in the applications related to spatial filtering where the input data vector represents the readings collected by an array of sensors, this assumption may become more meaningful. Nevertheless, in practice and from past experience in adaptive filters, this assumption simplifies the analysis and does lead to useful insights [83, 106].

Assumption 2: The input data obeys $\mathbf{x}[i] \sim \mathcal{N}(\mathbf{0}, R)$ for all i.

Remark. This technical assumption facilitates the analysis by taking advantage of the useful results on Gaussian random variables [83].

Assumption 3: At steady state, the diagonal matrix $W(\mathbf{c}[i])$ in the update equations can be viewed as a fixed matrix.

Remark. As suggested in [41, 84], after the system enters its steady state and when μ is sufficiently small, the coefficients converge in both mean and mean-squared sense. Thus, the replacement of $W(\mathbf{c}[i])$ by a fixed matrix becomes reasonable and convenient. We shall see that these assumptions lead to theoretical results that are supported by experiments.

Assumption 4: At steady state, for a fixed diagonal matrix $G \in \mathbb{R}^{M\times M}$, let $\mathbf{x}[i]^{\intercal}G\mathbf{x}[i]$ be independent of $(\mathbf{c}_0 - \mathbf{c}[i])^{\intercal}\mathbf{x}[i]$.

Remark. This is actually an extension of the Separation Principle, for which $G = I$, introduced by Sayed [106].

For the purpose of steady-state analysis, we first consider the performance of an adaptive filtering algorithm of the general form

$$\mathbf{c}[i+1] = (I - \mu\beta)\,\mathbf{c}[i] + \mu G\mathbf{x}[i]e[i] \tag{5.31}$$

$$e[i] = y[i] - \mathbf{c}[i]^{\intercal}\mathbf{x}[i] \tag{5.32}$$

where G is a fixed diagonal matrix with positive diagonal entries. Define the steady-state excess mean squared error (EMSE) as [83]

$$P_{\text{ex}} \equiv P_{\text{ex}}[\infty] \triangleq \lim_{i\to\infty} \mathsf{E}\left[|(\mathbf{c}_0 - \mathbf{c}[i])^{\intercal}\mathbf{x}[i]|^2\right]. \tag{5.33}$$

It can be shown that [83] the steady-state mean squared error (MSE) under Assumptions 1 is given by

$$P \triangleq \lim_{i\to\infty} \mathsf{E}\left[e^2[i]\right] = \sigma_e^2 + P_{\text{ex}}. \tag{5.34}$$

We have the following theorem that characterizes the steady-state EMSE of the adaptive filtering algorithm (5.31) and (5.32). The proof is presented in Section 5.7.3, which follows the discussion in [41, 83, 84].

Theorem 10. *Under Assumptions 1 and 2, for an adaptive algorithm of the form of (5.31) and (5.32) with a sufficiently small μ:*

i) if $R = I$, the steady-state EMSE is

$$P_{ex} = \frac{\mu \sum_{k=1}^{M} \frac{B_k}{C_k}}{1 - \mu \sum_{k=1}^{M} \frac{A_k}{C_k}} \tag{5.35}$$

where $A_k \triangleq G_{k,k}^2$, $B_k \triangleq G_{k,k}^2 \sigma_e^2 + \frac{2}{\mu\beta + \mu G_{k,k}} \beta^2 c_{0,k}^2 - \beta^2 c_{0,k}^2, C_k \triangleq 2\beta - \mu\beta^2 + 2(1-\mu\beta)G_{k,k} - 2\mu G_{k,k}^2$.

ii) [84, Theorem III.1] if $\beta = 0$, the steady-state EMSE is

$$P_{ex} = \mu \cdot \mathrm{tr}(RG(2I - \mu RG)^{-1}) \cdot \sigma_e^2. \tag{5.36}$$

Remark: For the case where $R = I$ and $\beta = 0$, note that part i) of Theorem 10 does not exactly agree with part ii) , due to additional simplifications in deriving part ii) for a general R [84]. We suggest using part i) in this case, although the difference could be negligible for sufficiently small μ. Consequently, the formula of the steady-state EMSE in this case is given by

$$P_{\text{ex}} = \frac{\mu \sum_{k=1}^{M} \frac{G_{k,k}}{2 - 2\mu G_{k,k}}}{1 - \mu \sum_{k=1}^{M} \frac{G_{k,k}}{2 - 2\mu G_{k,k}}} \cdot \sigma_e^2, \quad \text{for } R = 1, \beta = 0. \tag{5.37}$$

Meanwhile, the following theorem presents stability results for the prototype algorithm, whose proof is also provided in Section 5.7.3. Let $\lambda_{\text{max}}(A)$ denote the largest eigenvalue of the matrix A in magnitude.

Theorem 11. *i) (Convergence of Mean) Under Assumptions 1 and 2, the adaptive filtering algorithm of the form of (5.31) and (5.32) converges in the mean sense if $|\lambda_{max}((1-\mu\beta)I-\mu GR)|<1$.*

ii) (Convergence of MSE) Under Assumptions 1 and 2, for the adaptive filtering algorithm of the form of (5.31) and (5.32):

ii-a) When $R=I$, the steady-state MSE converges if

$$0<\mu<\left(\sum_{k=1}^{M}\frac{A_k}{C_k}\right)^{-1}. \tag{5.38}$$

ii-b) When $\beta=0$, and in addition under Assumption 4, the steady-state MSE converges if

$$0<\mu<\frac{2}{R_{1,1}\cdot\sum_{k=1}^{M}G_{k,k}}. \tag{5.39}$$

Remark: Similarly, for the case $\beta=0$ and $R=I$, we suggest using part ii-a) of Theorem 11 as the condition for convergence in MSE. Thus, in this case the algorithm converges if

$$\mu \sum_{k=1}^{M} \frac{G_{k,k}}{2 - 2\mu G_{k,k}} < 1 \tag{5.40}$$

which recovers the results in [41].

To utilized these theorems for the steady-state analysis for the prototype LMS algorithm proposed in Section 5.3.1, we can replace $W^2(\mathbf{c}[i])$ by a fixed matrix according to Assumption 3. It naturally suggests setting

$$G = W^2(\mathbf{c}_0)$$

in Theorems 10 and 11 to obtain the corresponding versions for the prototype LMS algorithm.

For the prototype NLMS algorithm, the recognition of a proper G becomes difficult. One possible approximation would be

$$G = \frac{W^2(\mathbf{c}_0)}{\mathsf{E}\left[\mathbf{x}[i]^\intercal W^2(\mathbf{c}_0)\mathbf{x}[i]\right]}.$$

5.4 The pALMS and pANLMS Algorithms

5.4.1 Derivation and Discussion of pALMS

We choose the ℓ_p norm as the diversity measure, i.e.,

$$g(\mathbf{c}) = \sum_{k=1}^{M} |c_k|^p \tag{5.41}$$

where $p \in (0, 2]$ is a constant.[2] It is worth noting that the ℓ_p norm is a very popular diversity measure in the literature of sparse signal recovery. Many existing algorithms utilize it in their formulations. For example, basis pursuit [19, 24] and Lasso [113] employ the ℓ_1 norm, which essentially leads to solving certain convex optimization problems. In contrast, FOCUSS [60] works with ℓ_p with $p \leq 1$, which in general results in a non-convex optimization problem which can be effectively solved by iteratively reweighted procedures [21, 100]. Meanwhile, the ℓ_p norm satisfies the conditions in Theorem 9.

To continue, we note that

$$\nabla_{\mathbf{c}} g(\mathbf{c}) = \alpha(\mathbf{c}) \cdot \Pi(\mathbf{c}) \cdot \mathbf{c}$$

where $\alpha(\mathbf{c}) \equiv p$ and $\Pi(\mathbf{c}) = \text{diag}(|c_k|^{p-2})$. Accordingly, based on the derivation for the prototype LMS algorithm, we recognize

$$W(\mathbf{c}) = \Pi^{-\frac{1}{2}}(\mathbf{c}) = \text{diag}(|c_k|^{\frac{2-p}{2}}).$$

[2]Strictly speaking, for $p < 1$, it is not a norm.

Using these specifications, we reach the following LMS type adaptive algorithm

$$\mathbf{c}[i+1] = (1-\mu\beta p)\mathbf{c}[i] + \mu W^2(\mathbf{c}[i])\mathbf{x}[i]e[i] \tag{5.42}$$

$$W^2(\mathbf{c}[i]) = \text{diag}(|c_k[i]|^{2-p}). \tag{5.43}$$

We term this algorithm as the pALMS algorithm, where "p" indicates the relationship to ℓ_p norm, and "A" indicates the algorithm is derived using the Affine scaling transform.

A closer inspection actually reveals a practical issue with this algorithm. To see this, let us suppose that at the ith iteration there exists some coefficient $c_k[i] = 0$. As an example, this may occur at the first iteration if all the coefficients are initialized with zeros. According to (5.42) and (5.43), it can be readily seen that $c_k[j] \equiv 0$ for all $j \geq i$, which means that the coefficient gets stuck at zero for all subsequent updates. This is an undesirable effect, especially for the adaptation of a time-varying system, and it is caused by the fact that $[W^2(\mathbf{c})]_{k,k} = 0$ for $c_k = 0$. To address this issue, we suggest an effective remedy by replacing $[W^2(\mathbf{c})]_{k,k}$ with $[W^2(\mathbf{c})]_{k,k} + \epsilon_k$, where $\epsilon_k > 0$, $k \in [M]$, is chosen to be some small constant. In practice, a simpler approach would be letting $\epsilon_1 = ... = \epsilon_M = \epsilon > 0$. This is equivalent to the alternative update equation

$$W^2(\mathbf{c}[i]) = \text{diag}(|c_k[i]|^{2-p} + \epsilon). \tag{5.44}$$

We adopt this modified update to replace (5.43) for pALMS. Note that the pALMS update with $\beta = 0$ can be written as

$$\mathbf{c}[i+1] = \mathbf{c}[i] + \mu(W^2(\mathbf{c}[i]) + \epsilon I)\mathbf{x}[i]e[i]. \tag{5.45}$$

Another supporting argument for the modification can be obtained by rewriting (5.45) as

$$\mathbf{c}[i+1] = \frac{1}{1+\epsilon}\big(\mathbf{c}[i] + \mu(1+\epsilon)W^2(\mathbf{c}[i])\mathbf{x}[i]e[i]\big) + \frac{\epsilon}{1+\epsilon}\left(\mathbf{c}[i] + \mu((1+\epsilon))\mathbf{x}[i]e[i]\right).$$

The above equation suggests that the resulting filter can be viewed as a convex combination of an (unmodified) pALMS filter and an ordinary LMS filter. Although the LMS filter has small weight in determining the update direction, it effectively helps the combined filter to escape from some local optima. Note that similar practices of imposing small perturbations to the diagonal entries of a weighting matrix are also carried out in existing adaptive filtering algorithms. As an example, steps (5.9) and (5.10) in PNLMS serve to lift up the small magnitudes in the coefficient vector before forming the diagonal weighting matrix.

Further, an important nature regarding the proposed algorithmic framework can be illustrated by using pALMS as an example. Consider the limiting case where $\beta \to 0_+$. In this scenario, the cost function (5.3) exerts diminishing impact on enforcing sparse solution, which means eventually no sparse solution is favored over other possible solutions. However, the pALMS with $\beta = 0$ still bears the nature of a proportionate type algorithm, which weighs the update of each tap by some factor determined via its previous value. Hence, when the underlying predictor is sparse, pALMS with $\beta = 0$ is capable of exploiting this sparsity structure by weighing the updates proportionately, and hence it expedites the adaptation procedure. This property essentially separates affine scaling algorithms from their counterparts developed in the original coefficient domain (e.g., pALMS versus ZA-LMS). Usually, coefficient-domain algorithms work with $\beta > 0$ to enforce sparse solutions, at the expense of biased estimation of the coefficients [25]. Otherwise, these algorithms may reduce to the ordinary LMS without possible benefit for sparse predictor estimation.

5.4.2 Steady-State Performance Analysis of pALMS

We mainly consider the case for $\beta = 0$ under Assumptions 1-4. By letting $G = \text{diag}(|c_{0,k}|^{2-p} + \epsilon)$, Theorems 10 and 11 can be applied to obtain guidance on the steady-state EMSE and the convergence conditions. Since the diagonal matrix $W(\mathbf{c}[i])$ is time-varying, one should choose μ in a more conservative manner than the upper limits predicted in Theorem 10 to ensure convergence in practice.

Generally, when $\beta > 0$, it becomes difficult to obtain a useful estimate of G which can be explicitly evaluated, mainly due to the fact that nonzero β introduces bias in the adaptation of the predictor coefficients (see, e.g., [25]).

5.4.3 Derivation and Discussion of pANLMS

Based on the prototype NLMS algorithm as well as the discussion above, we can obtain an NLMS type algorithm as follows

$$\mathbf{c}[i+1] = \mathbf{c}[i] + \frac{\mu}{\mathbf{x}[i]^{\intercal} W^2(\mathbf{c}[i])\mathbf{x}[i]} W^2(\mathbf{c}[i])\mathbf{x}[i]e[i] \tag{5.46}$$

$$W^2(\mathbf{c}[i]) = \mathrm{diag}(|c_k[i]|^{2-p} + \epsilon). \tag{5.47}$$

We refer to this algorithm as pANLMS, which can be viewed as an improved version of the pNLMS algorithm proposed in [101].

It is worthwhile to note that pANLMS, PNLMS, and IPNLMS share similar features. Let us first discuss the connection between pANLMS and PNLMS. For the purpose of analysis, we consider the case where $\rho_{\text{PNLMS}}, \delta_{\text{PNLMS}} \to 0_+$ for PNLMS, and $p = 1, \epsilon \to 0_+$ for pANLMS. Thus, the PNLMS updating rule for $W(\mathbf{c}[i])$ becomes

$$W(\mathbf{c}[i]) = \mathrm{diag}(|c_k[i]|^{\frac{1}{2}})$$

which coincides with that of pANLMS. The only difference between pANLMS and PNLMS lies in the denominators of (5.46) and (5.7): pANLMS uses

$$\mathbf{x}[i]^{\intercal} W^2(\mathbf{c}[i])\mathbf{x}[i] \equiv \sum_{k=1}^{M} w_k^2 x_k^2$$

whereas PNLMS employs

$$\frac{1}{M}\sum_{k=1}^{M}[W^2(\mathbf{c}[i])]_{k,k}\cdot\|\mathbf{x}[i]\|_2^2 \equiv \frac{1}{M}\sum_{k=1}^{M}w_k^2\sum_{k=1}^{M}x_k^2$$

where w_k denotes the kth diagonal element of $W(\mathbf{c}[i])$, and x_k represents the kth element of $\mathbf{x}[i]$. Let $w_{<k>}^2$ denote the ordered sequence with $w_{<1>}^2 \geq w_{<2>}^2 \geq \cdots \geq w_{<M>}^2$. Then, standard results on inequalities involving real sequences [8] gives

$$\sum_{k=1}^{M}w_{<k>}^2x_{<M-k+1>}^2 \leq \sum_{k=1}^{M}w_k^2x_k^2 \leq \sum_{k=1}^{M}w_{<k>}^2x_{<k>}^2 \tag{5.48}$$

and

$$\sum_{k=1}^{M}w_{<k>}^2x_{<M-k+1>}^2 \leq \frac{1}{M}\sum_{k=1}^{M}w_k^2\sum_{k=1}^{M}x_k^2 \leq \sum_{k=1}^{M}w_{<k>}^2x_{<k>}^2. \tag{5.49}$$

Note that, according to (5.48) and (5.49), the denominators of PNLMS and pANLMS fall into the same interval. The bounds could be tighter for sparse predictors since most of the diagonal elements of $W(\mathbf{c}[i])$ may be very close to zero as the filters enters their steady states. Therefore, we can intuitively interpret PNLMS as a variation of pANLMS with $p = 1$, and thus establish its connection to sparse signal recovery techniques.

Next, we compare pANLMS with $p = 1$ and IPNLMS. To see the connection, we let $\delta_{\text{IPNLMS}}, \epsilon_{\text{IPNLMS}} \to 0_+$. Note that, for IPNLMS, when $\|\mathbf{c}[i]\|_1 \neq 0$ and $\alpha_{\text{IPNLMS}} \neq -1$,

$$\frac{W^2(\mathbf{c}[i])}{\mathbf{x}[i]^\intercal W^2(\mathbf{c}[i])\mathbf{x}[i]} = \frac{W_1^2(\mathbf{c}[i])}{\mathbf{x}[i]^\intercal W_1^2(\mathbf{c}[i])\mathbf{x}[i]}$$

where $W_1^2(\mathbf{c}[i])$ is a diagonal matrix with

$$[W_1^2(\mathbf{c}[i])]_{k,k} = |c_k[i]| + \frac{(1-\alpha_{\text{IPNLMS}})\|\mathbf{c}[i]\|_1}{(1+\alpha_{\text{IPNLMS}})M}. \tag{5.50}$$

Note that (5.50) is very similar to the pANLMS update (5.47) with $p = 1$ by recognizing the term $\frac{(1-\alpha_{\text{IPNLMS}})\|\mathbf{c}[i]\|_1}{(1+\alpha_{\text{IPNLMS}})M}$ as ϵ in pANLMS. This is especially the case when the algorithm enters its steady state and $\|\mathbf{c}[i]\|_1$ does not fluctuate much. This observation establishes the connection between pANLMS with $p = 1$ and IPNLMS.

5.4.4 Steady-State Performance Analysis of pANLMS

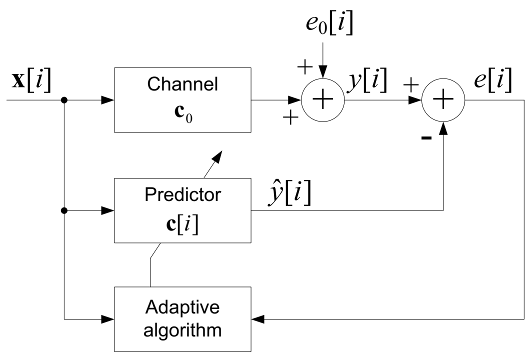
Let us suppose that Assumptions 1-4 are in position. We need to find a fixed G to approximate the term $\frac{W^2(\mathbf{c}[i])}{\mathbf{x}[i]^\intercal W^2(\mathbf{c}[i])\mathbf{x}[i]}$, for which an exact characterization seems difficult, if at all possible, to obtain. Note that, for a fixed W, $\mathsf{E}[\mathbf{x}[i]^\intercal W^2 \mathbf{x}[i]] = R_{1,1}\text{tr}(W^2)$. Hence, we replace $\mathbf{x}[i]^\intercal W^2 \mathbf{x}[i]$ by $R_{1,1}\text{tr}(W^2)$ and reach the approximation that $G \to \frac{1}{R_{1,1}\text{tr}(W^2(\mathbf{c}_0))} W^2(\mathbf{c}_0)$, where $\mathbf{c}[i]$ is substituted with the true coefficient vector $\mathbf{c}_0$. A useful fact from this approximation is that $\text{tr}(G) = R_{1,1}^{-1}$. As a result, when $R = I$ and μ is small, Theorem 10 and 11 imply that

$$P_{\text{ex}} = \frac{\mu}{2-\mu}\sigma_e^2 \tag{5.51}$$

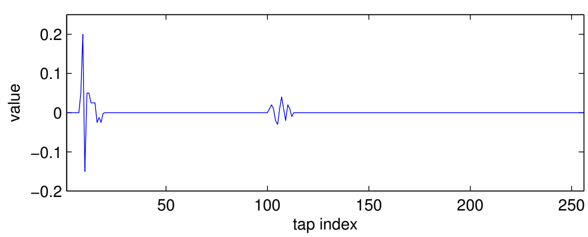
and $0 < \mu < \frac{2}{3}$ is needed to ensure convergence in MSE.

5.5 Experiments

We perform experiments to demonstrate the performance of the pALMS and pANLMS algorithms. The application scenario for the experiments is illustrated in Figure 5.1(a). The sparse channel $\mathbf{c}_0$ has 256 taps with 24 nonzero taps, which is shown in Figure 5.1(b). The goal is to estimate the underlying sparse channel and match the desired signal $y[i]$.



(a) The application scenario.



(b) The sparse channel.

Figure 5.1: The application scenario for experimental study.

In our experiments, the predictor coefficients are initialized with all zeros. For each set of experiments, the adaptation is initially turned off for a short period and then simultaneously turned on for the convenience of comparison. Each performance curve is obtained by averaging over 20000 random trials.

5.5.1 Experimental Justification of Theorem 10 for pALMS

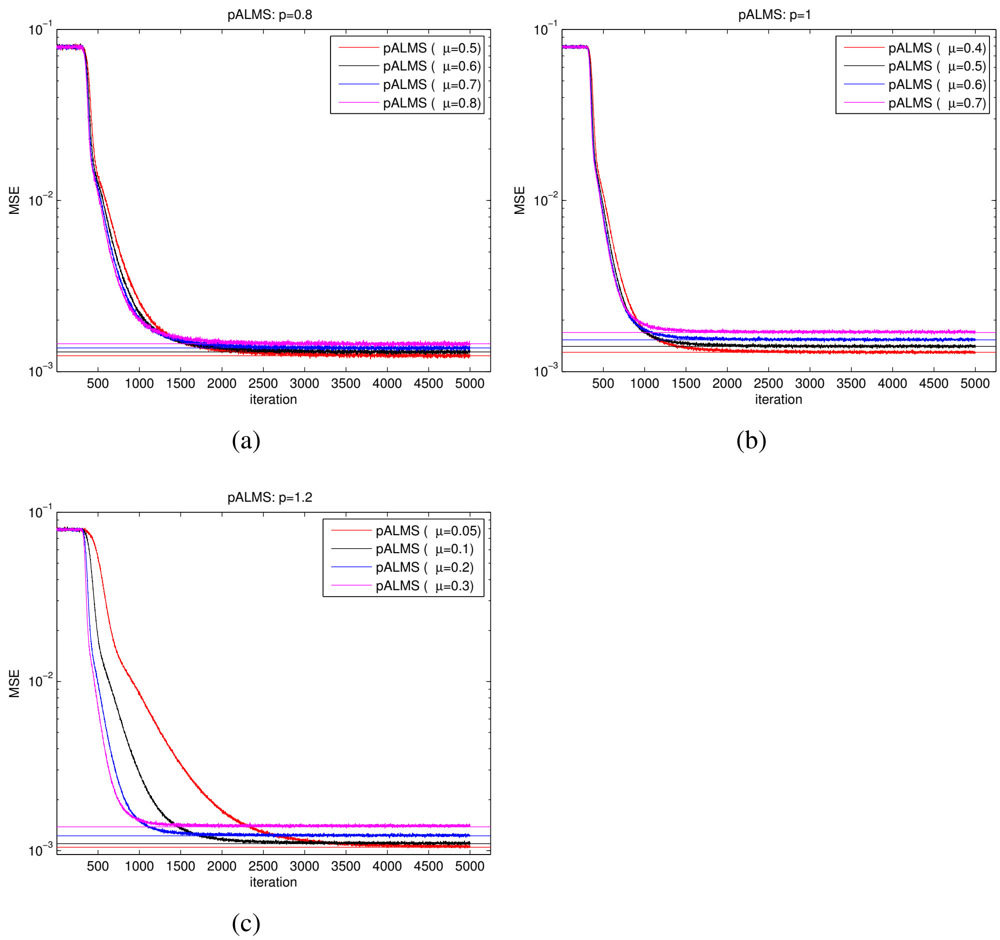
We justify the effectiveness of Theorem 10 on estimating the excess MSE of pALMS. In this set of experiments, the input signal $\mathbf{x}[i]$ is a Gaussian white noise sequence where each element is independently drawn from $\mathcal{N}(0,1)$. The noise $e_0[i]$ is i.i.d. according to $\mathcal{N}(0,0.001)$. We consider pALMS with $\beta = 0$, $p = 0.8, 1, 1.2$, respectively. For each choice of p, we test different choices of μ, and compare the empirical MSE with the estimated values (using (5.34)) by Theorem 10. The results are presented in Figure 5.2.

From Figure 5.2, we can see that the prediction of MSE made by Theorem 10 agrees well with the steady-state MSE of the pALMS in most cases, especially when the step size parameter μ is relatively small.

5.5.2 The Effect of Parameters

We start with the study of the roles of parameters of pALMS, which gives us guidance for further experimentation. The experiment setup is the same as in Section 5.5.1.

First, we study the effect of p in pALMS. For comparison, we run the LMS algorithm with $\mu_{\text{LMS}} = 0.001$. For pALMS, we set $\beta = 0, \epsilon = 0.001$. We study the cases for $p = 0.5, 0.8, 1, 1.2$. Using Theorem 10, we obtain $\mu_{\text{pALMS}} = 0.513$, 0.334, 0.226, 0.141, respectively, to yield matched steady-state MSE to that of the LMS algorithm. Figure 5.3 summarizes the results.



(a) (b) (c)

Figure 5.2: Experimental justification of the theoretic estimation of MSE for pALMS. In each plot, the horizontal lines indicates MSE predicted by the theorems. For each choice of μ, the color of the predicted MSE matches the color of the empirical learning curve of pALMS.

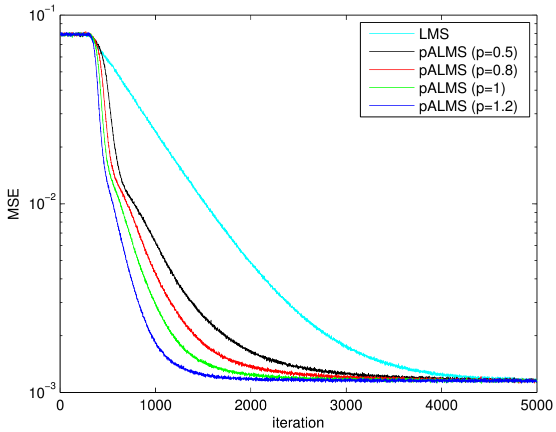
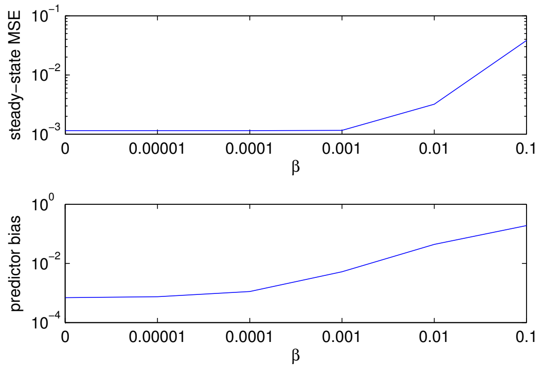


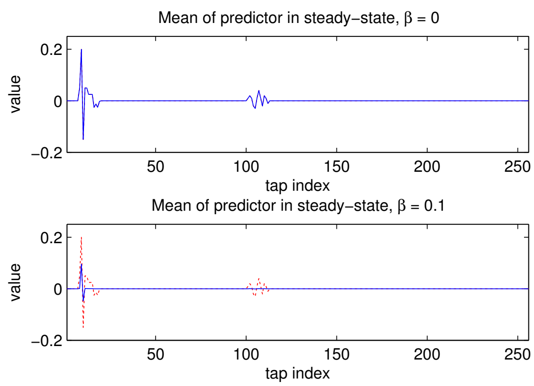
Figure 5.3: The effect of p in pALMS.

From Figure 5.3, we note that the steady-state MSE of this set of algorithms agree well, which justifies the validity of Theorem 10 in this case. At the very initial stage of adaptation, the pALMS algorithms seem to adapt slowly, due to the fact that the coefficients are initialized with all zeros, thereby it is the ϵ term in (5.44) that helps pALMS to break out of the all-zero state. Once the coefficients have reasonable magnitudes, the adaptation becomes very fast. Overall, the set of pALMS algorithms outperform LMS, indicating their usefulness in sparse channel estimation. Further, it can be seen that as p increases from 0.5 to 1.2, the performance of pALMS improves since the convergence becomes faster. Note that, as $p \to 2$, the behavior of the pALMS algorithm will theoretically converge to an ordinary LMS algorithm.

Next, we examine the effect of β in pALMS. Specifically, pALMS with $p = 1, \epsilon = 0.001$, and $\mu_{\text{pALMS}} = 0.226$ is performed with different choices of β. We focus on the steady-state MSE and the bias in the estimated channel coefficients. Note that the bias is computed as $\|\mathbf{c}[\infty] - \mathbf{c}_0\|_2$, where $\mathbf{c}[\infty]$ is calculated as the average of 500



(a) Effect on steady-state MSE and predictor bias



(b) Effect on predictor sparsity (red dotted curve indicates true channel coefficients $\mathbf{c}_0$)

Figure 5.4: The effect of β in pALMS.

iterations after the algorithm enters its steady state. Figure 5.4 summarizes the results.

According to Figure 5.4(a), we note that as β increases, the steady-state MSE increases. This is reasonable because of the tradeoff between the prediction quality and the sparsity of the predictor as indicated in (5.3). As we enforce more sparsity in the predictor by increasing β, the bias in the estimated channel coefficients increases. This is further illustrated in Figure 5.4(b). We can see that when $\beta = 0$, the mean of the predictor agrees well with the true channel coefficients. In contrast, when $\beta = 0.1$, most predictor coefficients are enforced to be very close to zero, resulting in more sparsity with large bias in channel estimation.

Now, let us focus on the effect of p in pANLMS. For the purpose of comparison, we also run NLMS algorithm with $\mu_{\text{NLMS}} = 0.2$. The results with different choices of p are shown in Figure 5.5.

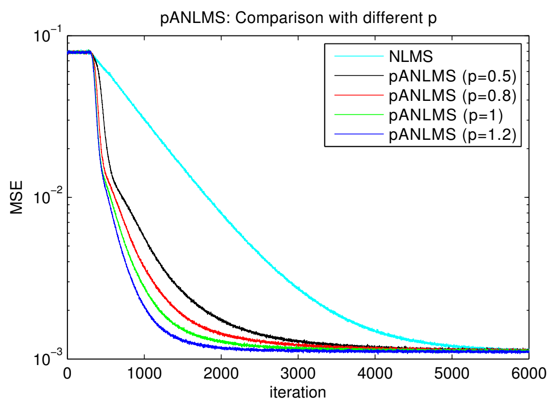


Figure 5.5: The effect of p in pANLMS.

From Figure 5.5, we can draw conclusions for pANLMS which are very similar to the discussions we had for pALMS.

5.5.3 Comparisons to Existing LMS Type Algorithms

We investigate the performance of the pALMS algorithm, and compare pALMS with LMS, L_0-LMS, ZA-LMS, and RZA-LMS. The sparse channel in Figure 5.1(b) is employed. Two types of input signals are considered as follows.

White Input Signal

We consider the case where $\mathbf{x}[i]$ is obtain from a white time series, which is the same with that used in Section 5.5.1. The parameter selection is based on the following procedure.

1. We pick $\mu_{L_0\text{-LMS}} = 0.0015$ for L_0-LMS. We choose $Q_{L_0\text{-LMS}} = 4$ and $\alpha_{L_0\text{-LMS}} = 5$ as suggested in [61].[3] Using (24) of [110], we compute the optimal $\kappa = 2.03 \times 10^{-7}$ that gives the lowest steady-state MSE. This steady-state MSE can be computed using (11) of [110], and we have $\text{MSE}_{L_0\text{-LMS}} = 1.1157 \times 10^{-3}$. In the next, we will choose the parameters of other algorithms to match this steady-state MSE.

2. We determine the parameters for ZA-LMS and RZA-LMS. For a given μ, the selection of ρ in each iteration is determined using (11) of [26]. Note that there is no formula available for steady-state MSE of ZA-LMS and RZA-LMS.[4] We found the following parameters give roughly the same steady-state MSE as L_0-LMS and fastest convergence. For ZA-LMS, we choose $\mu_{\text{ZA-LMS}} = 0.0015$, $\eta_{\text{ZA-LMS}} = 0.93445$. Note that according to [26], ZA-LMS is sensitive to the selection of $\eta_{\text{ZA-LMS}}$. We actually experiment with different values and found that this current selection, given by $\eta_{\text{ZA-LMS}} = 1.1\|\mathbf{c}_0\|_1$, gives the best result. For RZA-LMS, $\mu_{\text{RZA-LMS}} = 0.0049$, $\delta_{\text{RZA-LMS}} = 10^{-4}$, $\eta_{\text{RZA-LMS}} = 26$. According to [26], RZA-LMS is less sensitive to the selection of $\eta_{\text{RZA-LMS}}$, and our current selection reflects $\eta_{\text{RZA-LMS}} = \frac{13}{12}\|\mathbf{c}_0\|_0$.

3. We determine the parameters for pALMS. We choose $\beta = 0, \epsilon = 0.001$, and

[3] We follow the notations in [110], where the parameter α is equivalent to β in [61].

[4] Note that here we adopt the ZA-LMS and RZA-LMS presented in [26], where the adaptation can be on or off depending on the instantaneous parameters. This is different from the earlier versions of ZA-LMS and RZA-LMS presented in [25], for which the parameter selections were difficult to determine. Hence, the steady-state analysis in [107], where the earlier version of ZA-LMS was analyzed, is not employed in this experiment.

$p = 1, 1.4$, respectively. The step size μ_{pALMS} is chosen according to Theorem 10-i), which gives $\mu_{\text{pALMS}} = 0.1849, 0.0656$, respectively, for corresponding p values.

4. We determine the parameter for LMS using Theorem 10-i). The step size is given by $\mu_{\text{LMS}} = 8.1 \times 10^{-4}$.

The learning curves of different algorithms are illustrated in Figure 5.6.

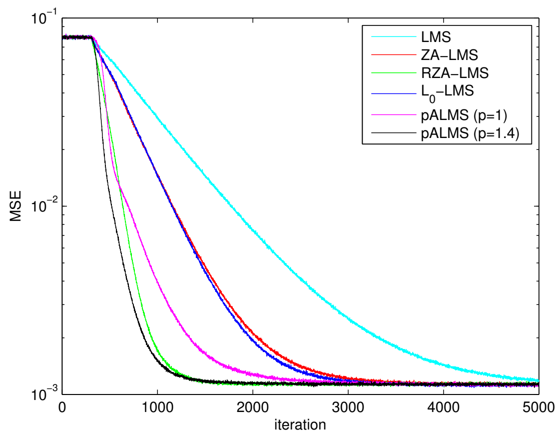


Figure 5.6: Performance comparison of LMS type algorithms. (white input)

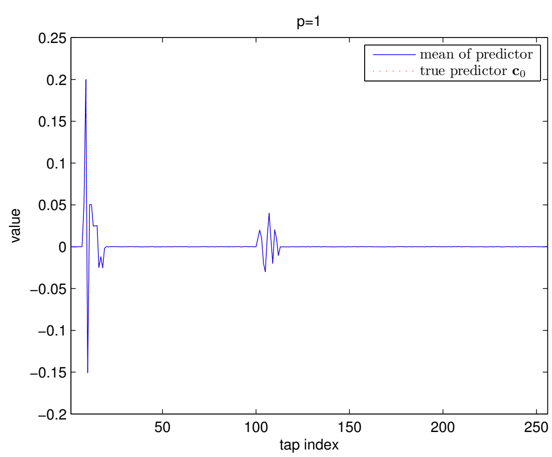


Figure 5.7: pALMS: Mean of predictor ($p = 1$). (white input)

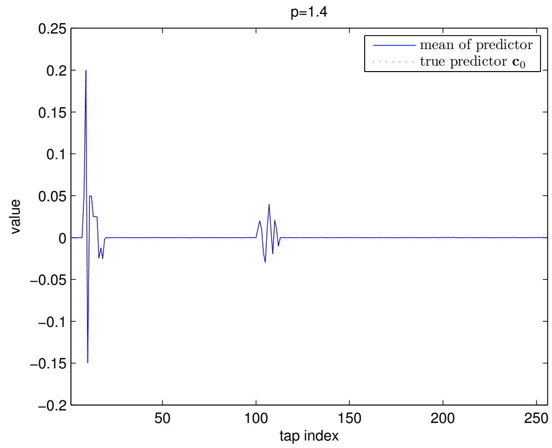


Figure 5.8: pALMS: Mean of predictor ($p = 1.4$). (white input)

From Figure 5.6, the steady-state MSE for the algorithms agree well. All adaptive algorithms with sparsity concerns outperform LMS. In this setup, ZA-LMS performs similarly to L_0-LMS, and L_0-LMS converges slightly faster than ZA-LMS in later stages. RZA-LMS has much fast convergence than ZA-LMS and L_0-LMS. Note that pALMS with $p = 1$ outperforms ZA-LMS and L_0-LMS, and pALMS with $p = 1.4$ outperforms RZA-LMS. Meanwhile, the mean of the predictor for pALMS in steady-state is shown in Figures 5.7 and 5.8 for pALMS with $p = 1$ and $p = 1.4$, respectively, using the same approach as in producing Figure 5.4(b). Comparing with the true predictor $\mathbf{c}_0$, we can see that the means of the predictor in both cases are very close to the true predictor.

Correlated Input Signal

In this case, the input signal is an AR(1) process generated by first filtering a white Gaussian noise through a system with the transfer function $\frac{1}{1-0.8z^{-1}}$ and the sequence is then normalized to have unit variance. First, we consider the LMS type algorithms. The parameter selection is based on the following procedure.

1. For LMS, we choose $\mu_{\text{LMS}} = 0.00126$.

2. For L_0-LMS, there is less guideline for choosing parameters in the case for correlated input signal. We choose the parameters such that it yields similar steady-state MSE to that of LMS. According our experiments and the parameter selection in the case for white input, we found that $Q_{L_0\text{-LMS}} = 4$ and $\alpha_{L_0\text{-LMS}} = 5$, $\kappa = 10^{-7}$, $\mu_{L_0\text{-LMS}} = 0.0015$ gives the reasonably good result under this setup.

3. For parameters with ZA-LMS and RZA-LMS, we employ (29) of [26] for choosing ρ. We choose the parameters to have the matching steady-state MSE to that of LMS. As a result, we choose $\mu_{\text{ZA-LMS}} = 0.0015, \eta_{\text{ZA-LMS}} = 0.9345, \mu_{\text{RZA-LMS}} = 0.0018, \eta_{\text{RZA-LMS}} = 26, \delta_{\text{RZA-LMS}} = 10^{-4}$.

4. For pALMS, we choose $\beta = 0, \epsilon = 0.001$, and $p = 1, 1.2$, respectively. Using Theorem 10, we choose $\mu_{\text{pALMS}} = 0.205, 0.127$, respectively, such that these two pALMS algorithm configurations yield the same steady-state MSE.

The learning curves of different algorithms are illustrated in Figure 5.9.

From Figure 5.9, we can see that LMS, L_0-LMS, ZA-LMS, and RZA-LMS have mathced steady-state MSE. L_0-LMS and ZA-LMS have very similar performance, and they outperform LMS with faster convergence. RZA-LMS outperforms L_0-LMS and ZA-LMS with faster convergence. In contrast, the pALMS algorithms achieve both

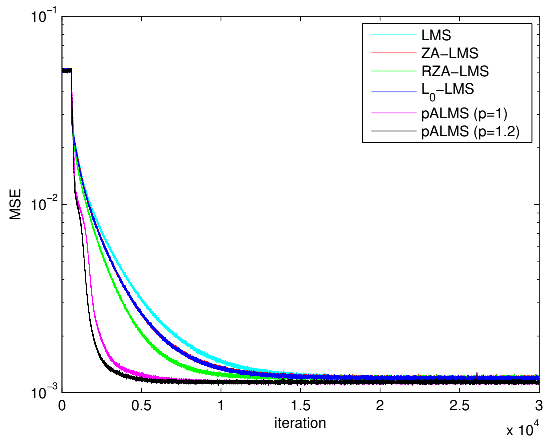


Figure 5.9: Performance comparison of LMS type algorithms. (correlated input)

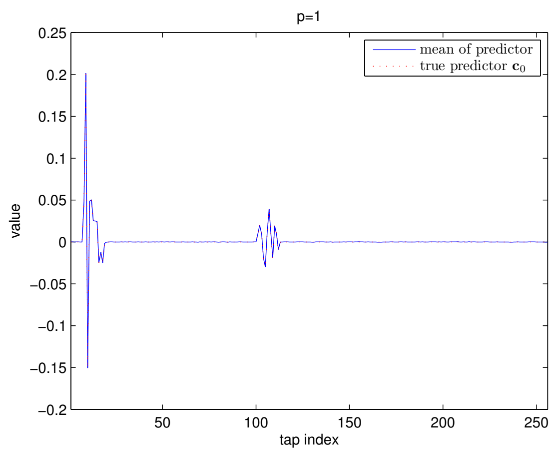


Figure 5.10: pALMS: Mean of predictor ($p = 1$). (correlated input)

faster convergence and lower steady-state MSE. In detail, the mean of the predictor for pALMS in steady-state is shown in Figures 5.10 and 5.11 for pALMS with $p = 1$ and $p = 1.2$, respectively, using the same approach as in producing Figure 5.4(b). Again, comparing with the true predictor $\mathbf{c}_0$, we can see that the means of the predictor in both cases are very close to the true predictor. Overall, the experimental results support the advantage of the affine scaling method for deriving adaptive algorithms that incorporate sparsity.

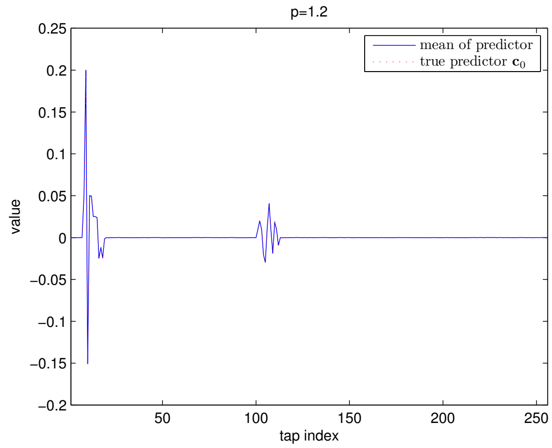


Figure 5.11: pALMS: Mean of predictor ($p = 1.2$). (correlated input)

5.5.4 Comparisons to Existing NLMS Type Algorithms

We investigate the performance of the pANLMS algorithm by comparing it with NLMS, PNLMS, IPNLMS, and IIPNLMS. Again, we employ the cases of white input and correlated input with the same experimental setups as in Section 5.5.3.

White Input Signal

For all algorithm, we choose $\mu = 0.2$ in expectation of the same steady-state MSE. For PNLMS, $\rho_{\text{PNLMS}} = \delta_{\text{PNLMS}} = 0.001$. For IPNLMS, $\alpha_{\text{IPNLMS}} = 0, \epsilon_{\text{IPNLMS}} = 0.001, \delta_{\text{IPNLMS}} = 0.001$. For IIPNLMS, $\alpha_{1,\text{IIPNLMS}} = -0.5, \alpha_{2,\text{IIPNLMS}} = 0.5, \epsilon_{\text{IIPNLMS}} = 0.001, \delta_{\text{IIPNLMS}} = 0.001, \rho_{\text{IIPNLMS}} = 0.01, \gamma_{\text{IIPNLMS}} = 0.1$. Figure 5.12 summarizes the results.

To interpret the results in Figure 5.12, we first note that the IIPNLMS and IPNLMS have slower initial convergence than PNLMS, but they outperform PNLMS with faster convergence at later stages, which is consistent with existing observations

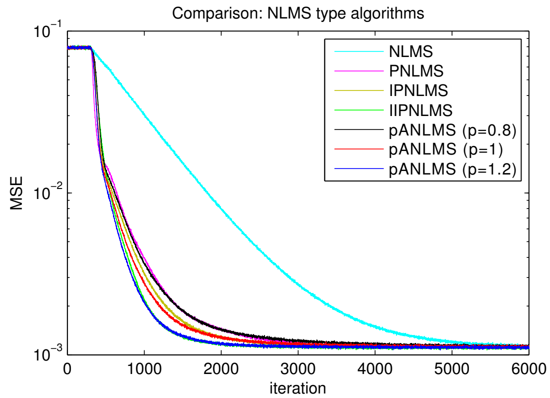


Figure 5.12: Performance comparison of NLMS type algorithms. (white input)

on these algorithms. Note that in this particular experimental setup, the pANLMS with $p = 0.8$ and PNLMS have similar behavior, whereas pANLMS with $p = 1$ and IPNLMS have comparable performance. These observations partly echoes the analysis in Section 5.4.3. The pANLMS algorithm with $p = 1.2$ performs better than other choices of p. It also has slightly faster convergence at the early stage than that of the IIPNLMS. At the late stage, their curves almost coincide.

Correlated Input Signal

For NLMS, we choose $\mu_{\text{NLMS}} = 0.2$. For PNLMS, IPNLMS, and IIPNLMS, we choose $\mu_{\text{PNLMS}} = \mu_{\text{IPNLMS}} = \mu_{\text{IIPNLMS}} = 0.2$ to expect the same stead-state MSE, and other parameters are chosen in the same manner as in the case of white input. For pANLMS, we choose $p = 1, 1.2$ with $\mu_{\text{pANLMS}} = 0.197$ in both cases to obtain the same steady-state MSE, which is predicted by Theorem 10-ii). Figure 5.13 illustrates the results.

Note that based on Figure 5.13 the NLMS has not converged yet, whereas other

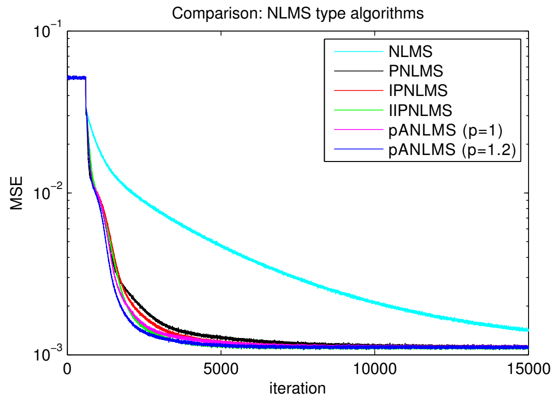


Figure 5.13: Performance comparison of NLMS type algorithms. (correlated input)

algorithms converge and have agreed steady-state MSE. The IIPNLMS and IPNLMS converge slower than PNLMS at the beginning, but they outperform PNLMS at later stages. The pANLMS with $p = 1.2$ converges faster than IIPNLMS almost throughout the adaptation, and it has the best performance in this set of algorithms. This experiment supports the usefulness of pANLMS for dealing with correlated time series.

5.6 Acknowledgements

Chapter 5 is, in part, a reprint of the paper "LMS Type Adaptive Filtering Algorithms that Incorporate Sparsity," Y. Jin, B. Song, and B. D. Rao, *submitted to IEEE Transactions on Signal Processing*. I was the primary author, B. Song contributed to the research, and B. D. Rao supervised the research.

5.7 Appendices

5.7.1 Proof of Theorem 9

The proof follows the idea in [99, Theorem 2]. We wish to show that the cost function (5.3) is decreased at each iteration. Note that (5.3) can be expanded as

$$J(\mathbf{c}) = \mathsf{E}\left|y[i] - \mathbf{c}^\mathsf{T}\mathbf{x}[i]\right|^2 + \beta \cdot g(\mathbf{c}) = \mathbf{c}_0 R\mathbf{c}_0 + \sigma_e^2 - 2\mathbf{d}^\mathsf{T}\mathbf{c} + \mathbf{c}^\mathsf{T} R\mathbf{c} + \beta \cdot g(\mathbf{c}).$$

Meanwhile, with the assumptions on $g(\mathbf{c})$, according to [92] (absorbing $\alpha(\mathbf{c}^{(j)})$ into $\Pi(\mathbf{c}^{(j)})$)

$$g(\mathbf{c}^{(j+1)}) - g(\mathbf{c}^{(j)}) \leq \frac{1}{2}\left((\mathbf{c}^{(j+1)})^\mathsf{T}\Pi(\mathbf{c}^{(j)})\mathbf{c}^{(j+1)} - (\mathbf{c}^{(j)})^\mathsf{T}\Pi(\mathbf{c}^{(j)})\mathbf{c}^{(j)}\right). \tag{5.52}$$

Thus

$$\begin{aligned}
&J(\mathbf{c}^{(j+1)}) - J(\mathbf{c}^{(j)}) \\
&= \left(\mathbf{c}_0 R\mathbf{c}_0 + \sigma_e^2 - 2\mathbf{d}^\mathsf{T}\mathbf{c}^{(j+1)} + (\mathbf{c}^{(j+1)})^\mathsf{T} R\mathbf{c}^{(j+1)} + \beta \cdot g(\mathbf{c}^{(j+1)})\right) \\
&\quad - \left(\mathbf{c}_0 R\mathbf{c}_0 + \sigma_e^2 - 2\mathbf{d}^\mathsf{T}\mathbf{c}^{(j)} + (\mathbf{c}^{(j)})^\mathsf{T} R\mathbf{c}^{(j)} + \beta \cdot g(\mathbf{c}^{(j)})\right) \\
&\leq \left(\mathbf{c}_0 R\mathbf{c}_0 + \sigma_e^2 - 2\mathbf{d}^\mathsf{T}\mathbf{c}^{(j+1)} + (\mathbf{c}^{(j+1)})^\mathsf{T} R\mathbf{c}^{(j+1)} + \frac{\beta}{2} \cdot (\mathbf{c}^{(j+1)})^\mathsf{T}\Pi(\mathbf{c}^{(j)})\mathbf{c}^{(j+1)}\right) \\
&\quad - \left(\mathbf{c}_0 R\mathbf{c}_0 + \sigma_e^2 - 2\mathbf{d}^\mathsf{T}\mathbf{c}^{(j)} + (\mathbf{c}^{(j)})^\mathsf{T} R\mathbf{c}^{(j)} + \frac{\beta}{2} \cdot (\mathbf{c}^{(j)})^\mathsf{T}\Pi(\mathbf{c}^{(j)})\mathbf{c}^{(j)}\right) \\
&= J(\mathbf{q}^{(j+1)}) - J(\mathbf{q}^{(j)})
\end{aligned} \tag{5.53}$$

where $J(\mathbf{q})$ is defined in (5.19) at the jth iteration. From (5.19), $J(\mathbf{q})$ is quadratic for all j and there exists a sequence of $\{\mu^{(j)}\}_{j=1}^{\infty}$, such that $J(\mathbf{q})$ is guaranteed to decrease at every iteration, i.e. $J(\mathbf{q}^{(j+1)}) - J(\mathbf{q}^{(j)}) < 0$. This choice of $\{\mu^{(j)}\}_{j=1}^{\infty}$ ensures the decrease in $J(\mathbf{c})$ from (5.53), and the algorithm (5.22) converges to a local minimum of (5.3).

5.7.2 Derivation of a General NLMS Type Algorithm Using Affine Scaling Transform

With the affinely scaled variable (5.17), an NLMS type algorithm can be derived by solving following optimization problem

$$\min_{\mathbf{c}[i+1]} \|\mathbf{q}[i+1] - \mathbf{q}[i]\|_2^2 \quad \text{subject to} \quad \mathbf{c}[i+1]^{\mathsf{T}}\mathbf{x}[i] = y[i]. \tag{5.54}$$

The Lagrangian is given by

$$\begin{aligned} L &= \|\mathbf{q}[i+1] - \mathbf{q}[i]\|_2^2 + \lambda(\mathbf{c}[i+1]^{\mathsf{T}}\mathbf{x}[i] - y[i]) \\ &= \|W^{-1}(\mathbf{c}[i])\mathbf{c}[i+1] - W^{-1}(\mathbf{c}[i])\mathbf{c}[i]\|_2^2 + \lambda(\mathbf{c}[i+1]^{\mathsf{T}}\mathbf{x}[i] - y[i]) \end{aligned}$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier. Thus

$$\nabla_{\mathbf{c}[i+1]} L = 2W^{-2}(\mathbf{c}[i])(\mathbf{c}[i+1] - \mathbf{c}[i]) + \lambda\mathbf{x}[i]$$

and by letting $\nabla_{\mathbf{c}[i+1]} L = \mathbf{0}$ we obtain

$$\mathbf{c}[i+1] = \mathbf{c}[i] - \frac{1}{2}\lambda W^2(\mathbf{c}[i])\mathbf{x}[i]. \tag{5.55}$$

Multiply both sides by $\mathbf{x}[i]^\intercal$ from the left and using the constraint in (5.54), we have

$$\mathbf{x}[i]^\intercal \mathbf{c}[i+1] = \mathbf{x}[i]^\intercal \mathbf{c}[i] - \frac{1}{2}\lambda \mathbf{x}[i]^\intercal W^2(\mathbf{c}[i])\mathbf{x}[i] = y[i]$$

which leads to

$$\lambda = -\frac{y[i] - \mathbf{x}[i]^\intercal \mathbf{c}[i]}{\frac{1}{2}\mathbf{x}[i]^\intercal W^2(\mathbf{c}[i])\mathbf{x}[i]} = -\frac{2e[i]}{\mathbf{x}[i]^\intercal W^2(\mathbf{c}[i])\mathbf{x}[i]}. \tag{5.56}$$

Substituting (5.56) into (5.55), we have

$$\mathbf{c}[i+1] = \mathbf{c}[i] + \frac{e[i]}{\mathbf{x}[i]^\intercal W^2(\mathbf{c}[i])\mathbf{x}[i]} W^2(\mathbf{c}[i])\mathbf{x}[i].$$

In practice, we have the prototype NLMS filter as given in (5.28) and (5.29).

5.7.3 Proofs of Theorems 10 and 11

We consider the following model

$$\mathbf{c}[i+1] = (1-\mu_1)\mathbf{c}[i] + \mu_2 G\mathbf{x}[i]e[i] \tag{5.57}$$

$$e[i] = y[i] - \mathbf{c}[i]^\intercal \mathbf{x}[i] \tag{5.58}$$

where $\mu_1, \mu_2 > 0$ and $G \in \mathbb{R}^{M\times M}$ is a fixed diagonal matrix. Note that by letting $\mu_1 = \mu_2\beta$, this model agrees with (5.31) and (5.32). Substituting (5.58) into (5.57) and using (5.30) gives

$$\mathbf{c}[i+1] = (1-\mu_1)\,\mathbf{c}[i] - \mu_2 G\mathbf{x}[i]\mathbf{x}[i]^\mathsf{T}\mathbf{c}[i] + \mu_2 G\mathbf{x}[i]\mathbf{x}[i]^\mathsf{T}\mathbf{c}_0 + \mu_2 G\mathbf{x}[i]e_0[i]. \quad (5.59)$$

Define the misalignment vector as $\tilde{\mathbf{c}}[i] \triangleq \mathbf{c}_0 - \mathbf{c}[i]$. Then, from (5.59), we have

$$\tilde{\mathbf{c}}[i+1] = ((1-\mu_1)I - \mu_2 G\mathbf{x}[i]\mathbf{x}[i]^\mathsf{T})\,\tilde{\mathbf{c}}[i] + \mu_1\mathbf{c}_0 - \mu_2 G\mathbf{x}[i]e_0[i]. \quad (5.60)$$

Based on (5.60), we apply different approaches to prove different parts of the theorems.

Proofs of Theorem 10-i), Theorem 11-i) and Theorem 11-ii-a)

Assumption 1 ensures that $\mathbf{x}[i]$, $\mathbf{c}[i]$, and $e_0[i]$ are mutually independent. Thus, taking expectation of both sides of (5.60) gives

$$\mathsf{E}\,\tilde{\mathbf{c}}[i+1] = ((1-\mu_1)I - \mu_2 GR)\,\mathsf{E}\,\tilde{\mathbf{c}}[i] + \mu_1\mathbf{c}_0. \quad (5.61)$$

Hence, the following condition is sufficient for convergence in mean sense [83]

$$|\lambda_{\text{max}}((1-\mu_1)I - \mu_2 GR)| < 1. \quad (5.62)$$

This justifies Theorem 3-i).

Next, based on (5.60), we have

$$
\begin{aligned}
&\tilde{\mathbf{c}}[i+1]\tilde{\mathbf{c}}[i+1]^\intercal \\
&= \left((1-\mu_1)I - \mu_2 G\mathbf{x}[i]\mathbf{x}[i]^\intercal\right)\tilde{\mathbf{c}}[i]\tilde{\mathbf{c}}[i]^\intercal\left((1-\mu_1)I - \mu_2\mathbf{x}[i]\mathbf{x}[i]^\intercal G\right) \\
&+ \mu_1^2\mathbf{c}_0\mathbf{c}_0^\intercal + \mu_2^2 e_0^2[i]G\mathbf{x}[i]\mathbf{x}[i]^\intercal G + \mu_1\left((1-\mu_1)I - \mu_2 G\mathbf{x}[i]\mathbf{x}[i]^\intercal\right)\tilde{\mathbf{c}}[i]\mathbf{c}_0^\intercal \\
&+ \mu_1\mathbf{c}_0\tilde{\mathbf{c}}[i]^\intercal\left((1-\mu_1)I - \mu_2\mathbf{x}[i]\mathbf{x}[i]^\intercal G\right) + Q
\end{aligned}
\tag{5.63}
$$

where Q represents the remaining cross terms whose expectations are zeros. Taking expectation of both sides, and employing the notation $\Phi[i] = \mathsf{E}\,\tilde{\mathbf{c}}[i]\tilde{\mathbf{c}}[i]^\intercal$, we have

$$
\begin{aligned}
\Phi[i+1] = &\underbrace{\mathsf{E}[\left((1-\mu_1)I - \mu_2 G\mathbf{x}[i]\mathbf{x}[i]^\intercal\right)\tilde{\mathbf{c}}[i]\tilde{\mathbf{c}}[i]^\intercal\left((1-\mu_1)I - \mu_2\mathbf{x}[i]\mathbf{x}[i]^\intercal G\right)]}_{\triangleq T_1} \\
&+ \mu_1^2\mathbf{c}_0\mathbf{c}_0^\intercal + \mu_2^2\sigma_e^2 GRG + \mu_1\left((1-\mu_1)I - \mu_2 GR\right)\mathsf{E}\,\tilde{\mathbf{c}}[i]\mathbf{c}_0^\intercal \\
&+ \mu_1\mathbf{c}_0\,\mathsf{E}\,\tilde{\mathbf{c}}[i]^\intercal\left((1-\mu_1)I - \mu_2 RG\right).
\end{aligned}
\tag{5.64}
$$

Let us inspect T_1. Note that

$$
\begin{aligned}
T_1 = &(1-\mu_1)^2\Phi[i] - \mu_2 GR\Phi[i](1-\mu_1) - \mu_2(1-\mu_1)\Phi[i]RG \\
&+ \mu_2^2 G\,\mathsf{E}[\mathbf{x}[i]\mathbf{x}[i]^\intercal\tilde{\mathbf{c}}[i]\tilde{\mathbf{c}}[i]^\intercal\mathbf{x}[i]\mathbf{x}[i]^\intercal]G.
\end{aligned}
$$

With Assumptions 1 and 2, it can be shown that [83], $\mathsf{E}[\mathbf{x}[i]\mathbf{x}[i]^\intercal\tilde{\mathbf{c}}[i]\tilde{\mathbf{c}}[i]^\intercal\mathbf{x}[i]\mathbf{x}[i]^\intercal] =$

$2R\Phi[i]R + R\text{tr}(R\Phi[i])$. Thus,

$$T_1 = (1-\mu_1)^2\Phi[i] - \mu_2 GR\Phi[i](1-\mu_1) - \mu_2(1-\mu_1)\Phi[i]RG + \mu_2^2 G(2R\Phi[i]R + R\text{tr}(R\Phi[i]))G.$$

Meanwhile, when μ_1, μ_2 satisfy (5.62), according to (5.61), we have

$$\mathsf{E}\,\tilde{\mathbf{c}}[\infty] = \mu_1 \left(\mu_1 I + \mu_2 GR\right)^{-1} \mathbf{c}_0 \tag{5.65}$$

where the notation $\mathsf{E}\,\tilde{\mathbf{c}}[\infty]$ denotes the mean value in steady state. Then, in steady state, i.e., $i \to \infty$,

$$\begin{aligned}\Phi[\infty] = {} & (1-\mu_1)^2\Phi[\infty] - \mu_2 GR\Phi[i](1-\mu_1) - \mu_2(1-\mu_1)\Phi[\infty]RG \\ & + 2\mu_2^2 GR\Phi[\infty]RG + \mu_2^2 GRG\text{tr}(R\Phi[\infty]) + \mu_1^2\mathbf{c}_0\mathbf{c}_0^{\mathsf{T}} + \mu_2^2\sigma_e^2 GRG \\ & + \mu_1^2\left((1-\mu_1)I - \mu_2 GR\right)\left(\mu_1 I + \mu_2 GR\right)^{-1}\mathbf{c}_0\mathbf{c}_0^{\mathsf{T}} \\ & + \mu_1^2\mathbf{c}_0\mathbf{c}_0^{\mathsf{T}}\left(\mu_1 I + \mu_2 GR\right)^{-1}\left((1-\mu_1)I - \mu_2 RG\right).\end{aligned}$$

We only consider the case where $R = I$. Thus

$$\begin{aligned}\phi[\infty] = {} & (1-\mu_1)^2\phi[\infty] - 2\mu_2(1-\mu_1)G\phi[\infty] + 2\mu_2^2 G^2\phi[\infty] + \mu_2^2\boldsymbol{g}^2\cdot\mathbf{1}^{\mathsf{T}}\phi[\infty] \\ & + \mu_1^2\mathbf{c}_0^2 + \mu_2^2\sigma_e^2\boldsymbol{g}^2 + 2\mu_1^2\left((1-\mu_1)I - \mu_2 G\right)\left(\mu_1 I + \mu_2 G\right)^{-1}\mathbf{c}_0^2\end{aligned} \tag{5.66}$$

where $\phi[\infty]$ is the vector of the diagonal elements of $\Phi[\infty]$, $\boldsymbol{g}$ is the vector of the diagonal elements of G, $\mathbf{c}_0^2$ and $\boldsymbol{g}^2$ denote the element-wise squares, and $\mathbf{1}$ denotes a vector of all 1's with a proper size. The steady-state EMSE is

$$P_{\text{ex}}[\infty] = \lim_{i\to\infty} \mathsf{E}[(\tilde{\mathbf{c}}[i]^\intercal \mathbf{x}[i])^2] = \lim_{i\to\infty} \text{tr}(\Phi[i]\,\mathsf{E}[\mathbf{x}[i]\mathbf{x}[i]^\intercal]) = \text{tr}(\Phi[\infty]) = \mathbf{1}^\intercal \phi[\infty].$$

Hence, according to (5.66),

$$\begin{aligned}
&\phi_k[\infty] \\
&= \frac{\mu_2^2 G_{k,k}^2 P_{\text{ex}}[\infty] + \mu_1^2 \mathbf{c}_{0,k}^2 + \mu_2^2\sigma_e^2 G_{k,k}^2 + 2\mu_1^2\left(1-\mu_1-\mu_2 G_{k,k}\right)\left(\mu_1+\mu_2 G_{k,k}\right)^{-1}\mathbf{c}_{0,k}^2}{1-(1-\mu_1)^2 + 2\mu_2(1-\mu_1)G_{k,k} - 2\mu_2^2 G_{k,k}^2} \\
&\triangleq \frac{A_k P_{\text{ex}} + B_k}{C_k}
\end{aligned}$$

where A_k, B_k, C_k are defined in Theorem 10-i) with proper recognition of μ_1, μ_2. Therefore

$$P_{\text{ex}}[\infty] = \sum_{k=1}^{M} \phi_k[\infty] = \sum_{k=1}^{M} \frac{A_k P_{\text{ex}} + B_k}{C_k}$$

which gives

$$P_{\text{ex}}[\infty] = \frac{\sum_{k=1}^{M} \frac{B_k}{C_k}}{1-\sum_{k=1}^{M}\frac{A_k}{C_k}}.$$

This justifies Theorem 10-i). By requiring $1-\sum_{k=1}^{M}\frac{A_k}{C_k} > 0$, we justify Theorem 11-ii-a).

Proof of Theorem 11-ii-b)

To handle the case of R, we apply the approach introduced by Sayed [106]. Assume $\mu_1 = 0$. Using the result from part (b) of Problem 9.18 in [106], by choosing $\Sigma = G^{-1}$, we have

$$\tilde{\mathbf{c}}[i+1]^\intercal G^{-1}\tilde{\mathbf{c}}[i+1] + \frac{(\tilde{\mathbf{c}}[i]^\intercal \mathbf{x}[i])^2}{\mathbf{x}[i]^\intercal G\mathbf{x}[i]} = \tilde{\mathbf{c}}[i]^\intercal G^{-1}\tilde{\mathbf{c}}[i] + \frac{(\tilde{\mathbf{c}}[i+1]^\intercal \mathbf{x}[i])^2}{\mathbf{x}[i]^\intercal G\mathbf{x}[i]}. \tag{5.67}$$

By using the following approximation in the steady state

$$\mathsf{E}[\tilde{\mathbf{c}}[i+1]^\intercal G^{-1}\tilde{\mathbf{c}}[i+1]] = \mathsf{E}[\tilde{\mathbf{c}}[i]^\intercal G^{-1}\tilde{\mathbf{c}}[i]] \tag{5.68}$$

we obtain

$$\mathsf{E}\left[\frac{(\tilde{\mathbf{c}}[i]^\intercal \mathbf{x}[i])^2}{\mathbf{x}[i]^\intercal G\mathbf{x}[i]}\right] = \mathsf{E}\left[\frac{(\tilde{\mathbf{c}}[i+1]^\intercal \mathbf{x}[i])^2}{\mathbf{x}[i]^\intercal G\mathbf{x}[i]}\right] \tag{5.69}$$

Note that, according to (5.60) with $\mu_1 = 0$,

$$\tilde{\mathbf{c}}[i+1]^\intercal \mathbf{x}[i] = \tilde{\mathbf{c}}[i]^\intercal \mathbf{x}[i](1 - \mu\mathbf{x}[i]^\intercal G\mathbf{x}[i]) - \mu\mathbf{x}[i]^\intercal G\mathbf{x}[i]e_0[i]. \tag{5.70}$$

Substitute (5.70) into (5.69) and rearranging terms, we have

$$P_{\text{ex}} = \frac{\mu}{2}\,\mathsf{E}\left[\mathbf{x}[i]^\intercal G\mathbf{x}[i]((\tilde{\mathbf{c}}[i]^\intercal \mathbf{x}[i])^2 + e_0[i]^2)\right] \tag{5.71}$$

Invoking Assumption 4, we have

$$P_{\text{ex}} = \frac{\mu}{2}\,\mathsf{E}\left[\mathbf{x}[i]^{\mathsf{T}} G \mathbf{x}[i]\right](P_{\text{ex}} + \sigma_e^2) = \frac{\mu}{2}\,\mathrm{tr}(GR)(P_{\text{ex}} + \sigma_e^2). \tag{5.72}$$

Solving (5.72) for P_{ex} leads to Theorem 11-ii-b). Note that this also leads to an expression for P_{ex}. We suggest using Theorem 10-ii) for P_{ex}, which is derived by [84].

Chapter 6

Concluding Remarks

The past several decades have witnessed a fast growth in the area of sparse signal recovery. Many applications have been shown to have intimate relations to sparse signal recovery, promoting the explosive development for practical algorithms. Especially, recently theoretical advancements regarding the performance issues in sparse signal recovery have enhanced the confidence of many practitioners from different science and engineering disciplines to apply the sparsity recovery techniques to applications of their interests.

6.1 Summary of Contributions

With the privilege of having access to a plethora of previous developments in this area, we wish to advance our understandings on the theory, algorithm, and applications for sparse signal recovery. In this thesis, first, we established the connection between multiple access communication and sparse signal recovery, and leveraged the techniques for the former problem to unveil performance limits of the latter. Sharp sufficient and necessary conditions were derived for guaranteeing successful support recovery of sparse signals in the asymptotical sense. We also demonstrated that this methodology is flexible and can incorporate with different signal and measurement models for sparse signal recovery.

Next, we extended the techniques for sparse signal recovery to meet new challenges in practical applications. To this end, we observed the strengths and weaknesses of sequential selection methods and joint recovery methods, respectively, and proposed the MultiPass algorithmic framework that aimed to make the best of both worlds. The MultiPass Lasso algorithm and the Reweighted MultiPass Lasso algorithm were derived. Experimental results demonstrated their performance improvement in terms of estimation accuracy and computational complexity.

Then, we applied our experiences in sparse signal recovery to two practical applications, i.e., robust regression and adaptive filtering. For robust regression, the key was properly recognizing a sparse component, which represents the outliers in the observations. For adaptive filtering, the contribution was transforming batch estimation algorithms, such as FOCUSS, into corresponding online versions which can facilitate the learning of dynamic systems. Experimental study showed that the proposed algorithms displayed attractive performance improvements.

To conclude, let us briefly discuss several potential research directions enabled by the theoretical and algorithmic developments presented in this thesis.

6.2 Suggestions for Future Research

First, we focus on the information theoretic perspective for performance analysis. Note that we established a useful connection between sparse signal recovery and multiuser communication, and leveraged the information theoretic techniques to tackle the performance issues in sparse signal recovery. Based on the vast knowledge base and many ongoing developments in multiuser communication, new opportunities are opened for exploration.

As one example, the design of channel codes and the development of decoding methods have been extensively studied in the contexts of information theory and wireless communication. Some of these ideas have been transformed into design principles for sparse signal recovery [1, 6, 70, 96, 137]. Thus far, however, the efforts in utilizing the codebook designs and decoding methods are mainly focused on the point-to-point channel model, which implies that the recovery methods iterate between first recovering one nonzero entry or a group of nonzero entries by treating the rest of them as noise and then removing the recovered nonzero entries from the residual signal. It motivates us to envision opportunities beyond a point-to-point channel model. As one important question, for example, can we develop practical codes for joint decoding and reconstruction techniques to simultaneously recover all the nonzero entries?

Another direction of interest is motivated by the role of W, i.e., the nonzero value matrix, in support recovery. On one hand, we demonstrated that a W with orthogonal columns could lead to performance improvement compared to a W with highly correlated columns. By improvement, we mean the fewer measurements are needed given other parameters fixed. On the other hand, the AR-SBL algorithm [138] models the temporal (row-wise) correlation of W by a first order AR model, and it explicitly learns the coefficient for this AR model among other parameters. By learning the structure of the underlying sparse signal, performance improvement can be attained. Here, by improvement, we mean the performance of support recovery in terms of success rate. These observations from theoretical analysis and algorithmic development pose an interesting question: Is correlation in the nonzero entries a blessing of a curse for sparse signal recovery?

We think that the estimation quality of the nonzero entries plays a linking role between theory and practice. In our analytical framework, we considered asymptotical scenarios, in which the estimation quality does not play a negative role since it can be made arbitrarily accurate as the problem size grows to infinity. However, the estimation quality becomes a cause for performance degradation for any finite-size problem. According to estimation theory, as the dependency among columns increases, the estimation quality improves. Although a W with more correlated columns will hurt the performance in the limiting case, it does make the estimation problem easier in a finite setting. Therefore, it is conceivable that there exists a performance tradeoff, regarding the structure of W, between the asymptotic performance limit and the performance that can be achieved in a finite setting. It will be interesting to characterize the impact of estimation quality on the performance in finite settings, which is not only of theoretical importance but also enlightening for the design of practical algorithms.

Next, the MultiPass algorithmic framework leaves us with several interesting research potentials. First, as observed in the experiments, the MultiPass Lasso exhibits higher computational efficiency and better estimation accuracy than Lasso. Our theoretical analysis addresses the performance improvement in terms of support recovery. However, it is still unclear why the computational complexity of MPL is lower than Lasso. It will be helpful to fully understand the theoretical justification behind this observation. Second, MultiPass algorithmic framework is general, and it can also work with other joint recovery methods. An interesting possibility would be sparse Bayesian learning (SBL). Note that in SBL there also exists certain parameter that, in effect, controls the sparsity of the resultant vector. It is generally observed that SBL exerts better performance than Lasso while having higher computational complexity. Can we derive a MultiPass version of SBL, which may enhance its estimation accuracy and, in the meantime, reduce its computational cost?

Further, we note that the proposed framework for deriving adaptive filtering algorithms is general. We studied in detail the adaptive filters using the ℓ_p norm diversity measure. It is worth noting that the ℓ_p norm diversity measure belongs to the family of separable diversity measures, which can be written as

$$g(\mathbf{c}) = \sum_{k=1}^{M} g_1(c_k)$$

for some function $g_1 : \mathbb{R} \mapsto \mathbb{R}$. This means that $g(\mathbf{c})$ is a sum of the costs on individual elements of $\mathbf{c}$. In contrast, there also exist nonseparable diversity measures which cannot be expressed as above. As an example, the sparse Bayesian learning algorithm for sparse signal recovery employs [129]

$$g(\mathbf{c}) = \min_{\gamma_k \geq 0, k \in [M]} \mathbf{c}^\intercal \operatorname{diag}(\gamma_k^{-1})\mathbf{c} + \log\det(\alpha I + A \operatorname{diag}(\gamma_k) A^\intercal)$$

where for some $\alpha > 0$, and $A \in \mathbb{R}^{M \times M}$ is the measurement matrix. It has been pointed out that [129] working with nonseparable diversity measures can lead to performance improvement in sparse signal recovery than using separable diversity measures. It would be interesting to study such an adaptive version based on the SBL diversity measure and explore its potential benefit for adaptive filtering applications.

Bibliography

[1] M. Akcakaya and V. Tarokh, “A frame construction and a universal distortion bound for sparse representations,” *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2443–2450, June 2008.

[2] M. Akçakaya and V. Tarokh, “Shannon theoretic limits on noisy compressive sampling,” *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 492–504, 2010.

[3] D. Angelosante and G. B. Giannakis, “Rls-weighted lasso for adaptive estimation of sparse signals,” *ICASSP*, pp. 3245–3248, 2009.

[4] S. Baillet, J. C. Mosher, and R. M. Leahy, “Electromagnetic brain mapping,” *IEEE Signal Processing Magazine*, pp. 14–30, 2001.

[5] W. U. Bajwa, J. Haupt, G. Raz, and R. Nowak, “Compressed channel sensing,” *CISS*, 2008.

[6] D. Baron, S. Sarvotham, and R. G. Baraniuk, “Bayesian compressive sensing via belief propagation,” *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 269–280, Jan 2010.

[7] J. Benesty and S. L. Gay, “An improved pnlms algorithm,” *ICASSP*, pp. 1881–1884, 2002.

[8] D. S. Bernstein, *Matrix Mathematics: Theorey, facts, and formulas (second edition).* Princeton University Press, 2009.

[9] N. J. Bershad, J. C. M. Bermudez, and J.-Y. Tourneret, “An affine combination of two lms adaptive filters—transient mean-square analysis,” *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1853–1864, 2008.

[10] N. J. Bershad and A. Bist, “Fast coupled adaptation for sparse impulse responses using a partial haar transform,” *IEEE Transactions on Signal Processing*, vol. 53, no. 3, pp. 966–976, 2005.

[11] G. E. P. Box and D. R. Cox, “An analysis of transformations,” *Journal of the Royal Statistical Society*, vol. 26, no. 2, pp. 211–252, 1964.

[12] L. Breiman, “Statistical modeling: the two cultures,” *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[13] S. D. Cabrera and T. W. Parks, “Extrapolation and spectral estimation with iterative weighted norm modification,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 4, pp. 842–851, 1991.

[14] E. Candes, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Mathematique*, 2008.

[15] E. Candes and T. Tao, “The dantzig selector: Statistical estimation when p is much larger than n,” *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[16] E. J. Candes, “Compressive sampling,” *Proceedings of the International Congress of Mathematicians*, pp. 1433–1452, 2006.

[17] E. J. Candes and J. Romberg, “ℓ1-magic : Recovery of sparse signals via convex programming,” *Technical Report*, 2005.

[18] E. J. Candes, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[19] E. J. Candes and T. Tao, “Decoding by linear programming,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[20] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, 2008.

[21] R. Chartrand and W. Yin, “Iteratively reweighted algorithms for compressive sensing,” *ICASSP*, 2008.

[22] C.-P. Chen and F. Qi, “Completely monotonic function associated with the Gamma functions and proof of Wallis’ inequality,” *Tamkang Journal of Mathematics*, vol. 36, pp. 303–307, 2005.

[23] J. Chen and X. Huo, “Theoretical results on sparse representations of multiple-measurement vectors,” *IEEE Transactions on Signal Processing*, vol. 54, pp. 4634–4643, 2006.

[24] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIREV*, vol. 43, no. 1, pp. 129–159, 2001.

[25] Y. Chen, Y. Gu, and A. O. Hero, "Sparse lms for system identification," *ICASSP*, pp. 3125–3128, 2009.

[26] ——, "Regularized least-mean-square algorithms," *ArXiv:1012.5066v2 [stat.ME]*, 2010.

[27] W. C. Chu, *Speech coding algorithms*. Wiley-Interscience, 2003.

[28] A. Cichocki, "Blind source separation: New tools for extraction of source signals and denoising," *SPIE*, pp. 11–25, 2005.

[29] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. on Communications*, vol. 50, pp. 374–377, 2002.

[30] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[31] S. Cotter, J. Adler, B. D. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 146, no. 5, pp. 235–244, 1999.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2006.

[33] J. Cui, P. A. Naylor, and D. T. Brown, "An improved ipnlms algorithm for echo cancellation in packet-switched networks," *ICASSP*, pp. 141–144, 2004.

[34] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[35] H. Deng and M. Doroslovački, "Proportionate adaptive algorithms for network echo cancellation," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1794–1803, 2006.

[36] D. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presense of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[37] D. Donoho, Y. Tsaig, I. Drori, and J. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," *preprint*, 2006.

[38] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Preprint*, 2004.

[39] ——, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[40] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 83–91, 2008.

[41] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 8, pp. 508–518, 2000.

[42] H. Dym, *Linear Algebra in Action (Graduate Studies in Mathematics, Volume 78).* American Mathematical Society, 2007.

[43] F. Y. Edgeworth, "On observations relating to several quantities," *Philosophical Magazine*, 1887.

[44] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[45] M. Elad and A. Feuer, "Superresolution restoration of an image sequence: Adaptive filtering approach," *IEEE Transactions on Image Processing*, vol. 8, no. 3, pp. 387–395, 1999.

[46] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, Nov 2009.

[47] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519.

[48] M. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, pp. 586–597, 2007.

[49] A. K. Fletcher, S. Rangan, and V. K. Goyal, "On the rate-distortion performance of compressed sensing," *ICASSP*, pp. 885–888, April 2007.

[50] ——, "Resolution limits of sparse coding in high dimensions," *NIPS*, 2008.

[51] ——, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5758–5772, Dec 2009.

[52] ——, "On-off random access channels: A compressed sensing framework," *Preprint*, 2009.

[53] H. Garudadri, P. K. Baheti, and S. Majumdar, "Low complexity sensors for body area networks," *Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments*, 2010.

[54] S. L. Gay, "An efficient, fast converging adaptive filter for network echo cancellation," *The Thirty-Second Asilomar Conference on Signals, Systems & Computers*, pp. 394–398, 1998.

[55] ——, "Affine projection algorithms," *in Least-mean-square adaptive filters, Editors: S. haykin and B. Widrow*, 2003.

[56] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin., *Bayesian Data Analysis*. Chapman & Hall, 2003.

[57] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of mimo channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, 2003.

[58] R. C. Gonzalez and R. E. Woods, *Digital image processing, 3rd Ed.* Prentice Hall, 2008.

[59] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroencephalography and Clinical Neurophysiology*, vol. 95, pp. 231–251, 1995.

[60] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted norm minimization algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[61] Y. Gu, J. Jin, and S. Mei, "ℓ_0 norm constraint lms algorithm for sparse system identification," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 774–777, 2009.

[62] D. Guo, "Neighbor discovery in ad hoc networks as a compressed sensing problem," *Presented at Information Theory and Application Workshop, UCSD*, 2009.

[63] E. T. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing," *TR07-07, available at http://www.caam.rice.edu/ optimization/L1/fpc/*.

[64] F. R. Hampel, "Robust estimation: A condensed partial survey," *Probability Theory and Related Fields*, 1973.

[65] F. Hasegawa, J. Luo, K. R. Pattipati, P. Willett, and D. Pham, "Speed and accuracy comparison of techniques for multiuser detection in synchronous cdma," *IEEE Transactions on Communications*, vol. 52, pp. 540–545, 2004.

[66] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, pp. 951–963, 2000.

[67] D. M. Hawkins, *Identification of Outliers*. Chapman and Hall Ltd, 1980.

[68] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.

[69] P. J. Huber, *Robust Statistics*. Wiley-Interscience, 2004.

[70] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, “Efficient and robust compressed sensing using optimized expander graphs,” *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4299–4308, Sep 2009.

[71] B. D. Jeffs, “Sparse inverse solution methods for signal and image processing applications,” *ICASSP*, pp. 1885–1888, 1998.

[72] N. Karmarkar, “New polynomial time algorithm for linear programming,” *Combinatorica*, vol. 4, no. 4, pp. 373–395, 1984.

[73] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “An interior-point method for large-scale l1-regularized least squares,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, pp. 606–617, 2007.

[74] H. O. Lancaster, *The* χ^2 *distribution*. Wiley, 1969.

[75] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, “Robust statistical modeling using the t distribution,” *Journal of the American Statistical Association*, vol. 84, pp. 881–896, 1989.

[76] A. Lapidoth, “Nearest neighbor decoding for additive non-Gaussian noise channels,” *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 1520–1529, 1996.

[77] J. Luo, K. Pattipati, P.Willett, and G. Levchuk, “Optimal grouping algorithm for a group decision-feedback detector in synchronous code-division multiaccess communications,” *IEEE Transactions on Communications*, vol. 41, no. 3, pp. 341–346, March 2003.

[78] D. J. C. MacKay, “Comparison of approximate methods for handling hyperparameters,” *Neural Computation*, vol. 11, pp. 1035–1068, 1999.

[79] R. E. Mahony and R. C. Williamson, “Prior knowledge and preferential structures in gradient descent learning algorithms,” *Journal of Machine Learning Research*, vol. 1, pp. 311–355, 2001.

[80] D. Malioutov, M. Cetin, and A. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[81] D. M. Malioutov, M. Çetin, and A. S. Willsky, “Optimal sparse representations in general overcomplete bases,” *ICASSP*, 2004.

[82] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[83] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and adaptive signal processing*. McGraw-Hill, 2000.

[84] R. K. Martin, W. A. Sethares, R. C. Williamson, and C. R. J. Jr., “Exploiting sparsity in adaptive filters,” *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1883–1894, 2002.

[85] K. Murano, S. Unagami, and F. Amano, “Echo cancellation and applications,” *IEEE Communications Magazine*, vol. 28, no. 1, pp. 49–55, 1990.

[86] J. Nagumo and A. Noda, “A learning method for system identification,” *IEEE Transactions on Automatic Control*, vol. 12, no. 3, pp. 282–287, 1967.

[87] S. G. Nash and A. Sofer, *Linear and nonlinear programming*. McGraw-Hill, 1996.

[88] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” *Communications of the ACM*, pp. 93–100, Dec 2010.

[89] G. Obozinski, M. J. Wainwright, and M. I. Jordan, “Support union recovery in high-dimensional multivariate regression,” *preprint*, 2010.

[90] M. R. Osborne, B. Presnell, and B. A. Turlach, “On the lasso and its dual,” *Journal of Computational and Graphical Statistics*, vol. 9, pp. 319–337, 1999.

[91] K. Ozeki and T. Umeda, “An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties,” *Electronics and Communications in Japan*, vol. 67-A, no. 5, 1984.

[92] J. A. Palmer and K. Kreutz-Delgado, “A globally convergent algorithm for map estimation with non-gaussian priors,” *The 36th Asilomar Conference on Signals and Systems*, pp. 1772–1776, 2002.

[93] J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao, “Variational em algorithms for non-gaussian latent variable models,” *NIPS*, pp. 1059–1066, 2006.

[94] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” *27th Annual Asilomar Conference on Signals, Systems, and Computers*, 1993.

[95] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge University Press, 2003.

[96] H. V. Pham, W. Dai, and O. Milenkovic, “Sublinear compressive sensing reconstruction via belief propagation decoding,” *ISIT*, 2009.

[97] S. U. H. Qureshi, "Adaptive equalization," *Proceedings of the IEEE*, vol. 73, no. 9, pp. 1349–1387, 1985.

[98] K. R. Rad, "Sharp upper bound on error probability of exact sparsity recovery," *The 43rd Annual Conference on Information Sciences and Systems*, pp. 14–17, 2009.

[99] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 760–770, 2003.

[100] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 187–200, 1999.

[101] B. D. Rao and B. Song, "Adaptive filtering algorithms for promoting sparsity," *ICASSP*, pp. 361–364, 2003.

[102] S. I. Resnick, *A probability path*. Birkhauser Boston, 1999.

[103] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Wiley, 2003.

[104] M. R. Sambur, "Adaptive noise canceling for speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-26, no. 5, pp. 419–423, 1978.

[105] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," *Proceedings of 44th Allerton Conference on Communication, Control, and Computing*, 2006.

[106] A. H. Sayed, *Fundamentals of Adaptive Filtering*. John Wiley & Sons, 2003.

[107] K. Shi and P. Shi, "Convergence analysis of sparse lms algorithms with l_1-norm penalty based on white input signal," *Signal Processing*, vol. 90, no. 12, pp. 3289–3293, 2010.

[108] J. W. Silverstein, "The smallest eigenvalue of a large dimensional Wishart matrix," *Annals of Probability*, vol. 13, pp. 1364–1368, 1985.

[109] M. Skolnik, *Introduction to radar systems*. McGraw-Hill, 1980.

[110] G. Su, J. Jin, and Y. Gu, "Performance analysis of l_0-lms with gaussian input signal," *ICSP*, pp. 235–238, 2010.

[111] G. Tang and A. Nehorai, "Performance analysis for sparse support recovery," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1383–1399, 2010.

[112] Z. Tian and G. B. Giannakis, "Compressed sensing for wideband cognitive radios," *ICASSP*, pp. 1357–1360, 2007.

[113] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[114] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, 2001.

[115] J. A. Tropp, "Greedy is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[116] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[117] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," *ICASSP*, 2005.

[118] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[119] M. K. Varanasi, "Group detection for synchronous gaussian code-division multiple-access channels," *IEEE Transactions on Information Theory*, vol. 41, no. 4, pp. 1083–1096, July 1995.

[120] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1417–1428, 2002.

[121] M. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," *ISIT*, June 2007.

[122] ——, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Transactions on Information Theory*, no. 12, pp. 5728–5741, Dec 2009.

[123] ——, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[124] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement," *ISIT*, pp. 2197–2201, 2008.

[125] ——, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2967–2979, June 2008.

[126] B. Widrow and M. E. Hoff, "Adaptive switching circuits," *IRE WESCON Convention Record, Part 4*, pp. 96–104, 1960.

[127] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Prentice Hall, 1985.

[128] D. Wipf and S. Nagarajan, "A unified bayesian framework for MEG/EEG source imaging," *NeuroImage*, pp. 947–966, 2008.

[129] ——, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[130] D. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, August 2004.

[131] ——, "Comparing the effects of different weight distributions on finding sparse representations," *NIPS*, 2006.

[132] D. P. Wipf, J. A. Palmer, and B. D. Rao, "Perspectives on sparse bayesian learning," *NIPS*, 2004.

[133] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[134] ——, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, July 2007.

[135] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 1–18, 2009.

[136] R. Zdunek and A. Cichocki, "Improved M-FOCUSS algorithm with overlapping blocks for locally smooth sparse signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4752–4761, Oct 2008.

[137] F. Zhang and H. D. Pfister, "Compressed sensing and linear codes over real numbers," *Information Theory and Applications Workshop, UCSD*, 2008.

[138] Z. Zhang and B. D. Rao, "Sparse signal recovery in the presence of correlated multiple measurement vectors," *ICASSP*, 2010.

[139] ——, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE Journal of Selected Topics in Signal Processing (in press)*, 2011.

[140] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.