Studies on Discrete-Valued Vector Reconstruction from Underdetermined Linear Measurements

Ryo Hayakawa

Department of Systems Science,
Graduate School of Informatics,
Kyoto University

March 2020

Abstract

Reconstruction of an unknown discrete-valued vector from its linear measurements is a common problem in communication systems. When the number of measurements is greater than or equal to the dimension of the unknown vector, the low complexity linear methods, such as minimum mean-square-error (MMSE) method, might achieve satisfactory reconstruction performance. In the underdetermined case with an insufficient number of measurements, however, their performance is severely degraded. On the other hand, although the maximum likelihood (ML) approach with the exhaustive search can achieve an excellent performance, it requires huge computational complexity in large-scale problems. This thesis proposes an efficient algorithm for the discrete-valued vector reconstruction and provides asymptotic performance analyses for some reconstruction methods.

Chapter 1 describes the discrete-valued vector reconstruction and its application in communication systems. Moreover, the conventional methods are briefly reviewed. Finally, the outline of this thesis is explained.

In Chapter 2, we focus on the reconstruction of the binary vector as the simplest example of the discrete-valued vector reconstruction. We extend the conventional sum of absolute values (SOAV) optimization to the weighted SOAV (W-SOAV) optimization so that we can use the prior information of the unknown vector. We then propose an iterative approach named iterative weighted SOAV (IW-SOAV), where we iterate the W-SOAV optimization and the update of the weight parameters in the objective function. The W-SOAV optimization can be efficiently solved with proximal splitting methods for convex optimization. Simulation results show that the reconstruction performance of the proposed IW-SOAV is better than several conventional methods in massive overloaded multiple-input multiple-output (MIMO) signal detection and the decoding of non-orthogonal space-time block codes.

In Chapter 3, we propose an algorithm for the reconstruction of a complex discrete-valued vector. The proposed method can be considered as an extension of the conventional SOAV optimization in the real-valued domain. The proposed approach in the complex-valued domain can utilize the dependency between the real part and the imaginary part of the unknown vector. It is shown that an optimization algorithm based on alternating direction method of multipliers (ADMM) can provide a sequence converging

to the solution of the optimization problem. We have shown via computer simulations that the proposed method can achieve good performance in MIMO signal detection and channel equalization.

Chapter 4 proposes a possibly nonconvex optimization problem for the discrete-valued vector reconstruction. The proposed sum of sparse regularizers (SSR) optimization problem can be regarded as a generalization of the convex SOAV optimization. For the proposed SSR optimization, two optimization algorithms based on ADMM and primal-dual splitting are proposed. Simulation results show that the proposed algorithms using nonconvex optimization can achieve better reconstruction performance than several conventional approaches using convex optimization.

In Chapter 5, we analyze the asymptotic performance of the SOAV optimization. We firstly propose the Box-SOAV optimization by adding a box constraint to the conventional SOAV optimization. By using convex Gaussian min-max theorem (CGMT), we evaluate the asymptotic performance of the estimate obtained by the Box-SOAV optimization. We also propose an approach to optimize the parameters of the Box-SOAV optimization on the basis of the theoretical result.

In Chapter 6, we analyze the performance of the SOAV optimization from a different perspective. We firstly propose a message passing-based algorithm using the idea of the SOAV optimization. Although the proposed method requires some assumptions on the measurement matrix, it can achieve good performance with low computational complexity. Moreover, we evaluate the asymptotic performance of the proposed algorithm on the basis of the state evolution. We also obtain the required measurement ratio for the perfect reconstruction in the noise-free case.

In Chapter 7, we present the conclusion of the thesis.

Acknowledgments

I would like to express my sincere gratitude to my supervisor Prof. Kazunori Hayashi for his long-term support. I feel that his valuable advice and constructive feedbacks significantly improved my research skills from various aspects. His incisive comments and encouragements were essential to complete this work.

I would like to show my deep appreciation to Prof. Hidetoshi Shimodaira, who is the head of the laboratory to which I belong in my doctoral course, for his help. I am also grateful to the members of my thesis committee, Prof. Toshiyuki Tanaka and Prof. Nobuo Yamashita for their time and valuable comments.

In my research, many people gave me enormous supports and valuable feedbacks. I appreciate Associate Prof. Megumi Kaneko for her thoughtful advice and much support in my bachelor and master course. Prof. Masaaki Nagahara and Dr. Hampei Sasahara gave me valuable inspiration for this work. I had deep discussions with Prof. Tadashi Wadayama and Assistant Prof. Satoshi Takabe for the topic related to my research. I am also greatly benefited from Prof. Takeo Ohgane, Associate Prof. Toshihiko Nishimura, Associate Prof. Shinsuke Ibi, Associate Prof. Koji Ishibashi, Assistant Prof. Takumi Takahashi, and the members in their laboratory. Their insightful comments and suggestions greatly improved the quality of my research. I would also like to thank other members in my laboratory.

In my visit to Aalborg University in Denmark, Prof. Petar Popovski, Prof. Elisabeth De Carvalho, and Associate Prof. Thomas Arildsen kindly supported me and gave useful advice. I had precious overseas experience thanks to their support. I owe my deep gratitude to them for their constructive suggestion and discussion.

I am also grateful to the Japan Society for the Promotion and Science (JSPS) for their financial support.

Finally, I would also like to express my cordial gratitude to my parents, Takashi Hayakawa and Megumi Hayakawa, for their tremendous supports in my life. Thanks to their enormous help, I really enjoyed the study and research in Kyoto University.

Ryo Hayakawa

List of Publications

Related Journal Papers

1. **R. Hayakawa** and K. Hayashi, "Reconstruction of complex discrete-valued vector via convex optimization with sparse regularizers," *IEEE Access*, vol. 6, pp. 66499–66512, Dec. 2018. (Copyright© 2018 IEEE) (Chapter 3)

2. **R. Hayakawa** and K. Hayashi, "Discreteness-aware approximate message passing for discrete-valued vector reconstruction," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6443–6457, Dec. 2018. (Copyright© 2018 IEEE) (Chapter 6)

3. **R. Hayakawa** and K. Hayashi, "Discreteness-aware decoding for overloaded non-orthogonal STBCs via convex optimization," *IEEE Communications Letters*, vol. 22, no. 10, pp. 2080–2083, Oct. 2018. (Copyright© 2018 IEEE) (Chapter 2)

4. **R. Hayakawa** and K. Hayashi, "Error recovery for massive MIMO signal detection via reconstruction of discrete-valued sparse vector," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E100-A, no. 12, pp. 2671–2679, Dec. 2017. (Copyright© 2017 IEICE) (Chapter 2)

5. **R. Hayakawa** and K. Hayashi, "Convex optimization-based signal detection for massive overloaded MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7080–7091, Nov. 2017. (Copyright© 2017 IEEE) (Chapter 2)

Related Conference Papers

1. K. Hayashi, A. Nakai-Kasai, and **R. Hayakawa**, "An overloaded SC-CP IoT signal detection method via sparse complex discrete-valued vector reconstruction," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2019*, Lanzhou, China, Nov. 2019. (Copyright© 2019 IEEE) (Chapter 1)

2. **R. Hayakawa** and K. Hayashi, "Discrete-valued vector reconstruction by optimization with sum of sparse regularizers," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO 2019)*, A Coruña, Spain, Sep. 2019. (Copyright© 2019 IEEE) (Chapter 4)

3. **R. Hayakawa** and K. Hayashi, "Performance analysis of discrete-valued vector reconstruction based on box-constrained sum of L1 regularizers," in *Proceedings of the 44th IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019. (Copyright© 2019 IEEE) (Chapter 5)

4. K. Hayashi, A. Nakai, **R. Hayakawa**, and S. Ha, "Uplink overloaded MU-MIMO OFDM signal detection methods using convex optimization," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2018*, Honolulu, USA, Nov. 2018. (Copyright© 2018 IEEE) (Chapter 1)

5. **R. Hayakawa** and K. Hayashi, "Binary vector reconstruction via discreteness-aware approximate message passing," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2017*, Kuala Lumpur, Malaysia, Dec. 2017. (Copyright© 2017 IEEE) (Chapter 6)

6. **R. Hayakawa** and K. Hayashi, "Discreteness-aware AMP for reconstruction of symmetrically distributed discrete variables," in *Proceedings of the 18th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2017)*, Sapporo, Japan, Jul. 2017. (Copyright© 2017 IEEE) (Chapter 6)

7. **R. Hayakawa** and K. Hayashi, "Error recovery with relaxed MAP estimation for massive MIMO signal detection," in *Proceedings of the International Symposium on Information Theory and Its Applications (ISITA) 2016*, California, USA, Oct.–Nov. 2016. (Copyright© 2016 IEEE) (Chapter 2)

8. **R. Hayakawa**, K. Hayashi, H. Sasahara, and M. Nagahara, "Massive overloaded MIMO signal detection via convex optimization with proximal splitting," in *Proceedings of the 24th European Signal Processing Conference (EUSIPCO 2016)*, Budapest, Hungary, Aug.–Sep. 2016. (Copyright© 2016 IEEE) (Chapter 2)

Other Journal Papers

1. S. Takabe, M. Imanishi, T. Wadayama, **R. Hayakawa**, and K. Hayashi, "Trainable projected gradient detector for massive overloaded MIMO channels: Data-driven tuning approach," *IEEE Access*, vol. 7, pp. 93326–93338, Jul. 2019.

2. **R. Hayakawa**, K. Hayashi, and M. Kaneko, "Lattice reduction-aided detection for overloaded MIMO using slab decoding," *IEICE Transactions on Communications*, vol. E99-B, no. 8, pp. 1697–1705, Aug. 2016.

3. K. Matsuoka, Y. Yushima, **R. Hayakawa**, R. Kawasaki, K. Hayashi, and M. Kaneko, "An RFID tag identification protocol via Boolean compressed sensing," *IEICE Communications Express*, vol. 5, no. 5, pp. 118–123, May 2016.

Other Conference Papers

1. **R. Hayakawa**, A. Nakai, and K. Hayashi, "Distributed approximate message passing with summation propagation," in *Proceedings of the 43rd International Conference on Acoustic, Speech, and Signal Processing (ICASSP 2018)*, Calgary, Canada, Apr. 2018.

2. **R. Hayakawa**, K. Hayashi, and M. Kaneko, "An overloaded MIMO signal detection scheme with slab decoding and lattice reduction," in *Proceedings of the 21st Asia-Pacific Conference on Communications (APCC 2015)*, Kyoto, Japan, Oct. 2015.

Acronyms

5G 5th generation.

8PSK 8-phase shift keying.

ADC analog-to-digital converter.

ADMM alternating direction method of multipliers.

AMP approximate message passing.

AO auxiliary optimization.

AWGN additive white Gaussian noise.

BER bit error rate.

Box-SOAV box-constrained sum of absolute values.

BP belief propagation.

BPSK binary phase shift keying.

CDA cyclic division algebra.

CDF cumulative distribution function.

CDMA code division multiple access.

CGMT convex Gaussian min-max theorem.

DAMP discreteness-aware approximate message passing.

DCT discrete cosine transform.

DFT discrete Fourier transform.

EP expectation propagation.

ERTS enhanced reactive tabu search.

FTN faster-than-Nyquist.

GAMP generalized approximate message passing.

GIGD graph-based iterative Gaussian detector.

i.i.d. independent and identically distributed.

IDFT inverse discrete Fourier transform.

IO-LAMA individually-optimal large MIMO AMP.

IoT Internet of things.

IRLS iterative reweighted least squares.

IW-SCSR iterative weighted sum of complex sparse regularizers.

IW-SOAV iterative weighted sum of absolute values.

KKT Karush-Kuhn-Tucker.

LAS likelihood ascent search.

LDPC low density parity check.

LLR log likelihood ratio.

LMMSE linear minimum mean square error.

M2M machine-to-machine.

MAP maximum a posteriori.

MIMO multiple-input multiple-output.

ML maximum likelihood.

MSE mean squared error.

MU-MIMO multi user multiple-input multiple-output.

NO-STBC non-orthogonal space-time block code.

NSE normalized squared error.

OAMP orthogonal approximate message passing.

OFDM orthogonal frequency division multiplexing.

PAM pulse amplitude modulation.

PDF probability density function.

PDS primal-dual splitting.

PO primary optimization.

PVC pre-voting cancellation.

QAM quadrature amplitude modulation.

QPSK quadrature phase shift keying.

RIP restricted isometry property.

RTS reactive tabu search.

SC-CP single carrier block transmission with cyclic prefix.

SCSR sum of complex sparse regularizers.

SER symbol error rate.

SNR signal-to-noise ratio.

SOAV sum of absolute values.

SSR sum of sparse regularizers.

STBC space-time block code.

SVD singular value decomposition.

VAMP vector approximate message passing.

W-SCSR weighted sum of complex sparse regularizers.

W-SOAV weighted sum of absolute values.

Contents

Chapter 1

Introduction

Discrete-valued vector reconstruction from its linear measurements is a common problem in signal processing for communications systems, e.g., multiple-input multiple-output (MIMO) signal detection [1–3] and multiuser detection in machine-to-machine (M2M) communications [4]. In some applications such as overloaded MIMO systems [5–9] and faster-than-Nyquist (FTN) signaling [10], the number of measurements is less than that of the unknown variables. In such underdetermined problems, simple linear methods, such as linear minimum mean square error (LMMSE) method, have poor performance. Although the maximum likelihood (ML) method with the exhaustive search can achieve good performance in terms of the error rate, the computational complexity increases exponentially along with the problem size. Thus, a low-complexity algorithm is required for the underdetermined discrete-valued vector reconstruction, especially in large-scale problems.

This chapter provides a short introduction of the discrete-valued vector reconstruction. Section 1.1 describes the reconstruction of the discrete-valued vector from its linear measurements. In Section 1.2, we briefly review several examples of the discrete-valued vector reconstruction in communication systems. Section 1.3 presents conventional approaches for the discrete-valued vector reconstruction. Finally, Section 1.4 explains the outline of this thesis.

In this thesis, we use the following notations. We denote the set of all real numbers by $\mathbb{R}$ and the set of all complex numbers by $\mathbb{C}$. $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ indicate the real part and the imaginary part, respectively. We represent the imaginary unit by j, the transpose by $(\cdot)^{\mathsf{T}}$, the Hermitian transpose by $(\cdot)^{\mathsf{H}}$, the $N \times N$ identity matrix by $\boldsymbol{I}_N$, the vector whose elements are all 1 by $\mathbf{1}$, and the matrix whose elements are all 0 by $\mathbf{0}$. For a vector $\boldsymbol{u} = [u_1 \ \cdots \ u_N]^{\mathsf{T}} \in \mathbb{K}^N$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$), we define the ℓ_1 and ℓ_2 norms of $\boldsymbol{u}$ as $\|\boldsymbol{u}\|_1 = \sum_{n=1}^{N} |u_n|$ and $\|\boldsymbol{u}\|_2 = \sqrt{\sum_{n=1}^{N} |u_n|^2}$, respectively. We also define $\|\boldsymbol{u}\|_0$ as the number of nonzero elements in $\boldsymbol{u}$. We represent the sample mean of the elements of $\boldsymbol{u}$ by $\langle \boldsymbol{u} \rangle = \frac{1}{N} \sum_{n=1}^{N} u_n$. $[\boldsymbol{u}]_n$ denotes the nth element of $\boldsymbol{u}$. $\mathrm{diag}(u_1, \ldots, u_N) \in \mathbb{K}^{N \times N}$ denotes the diagonal matrix

whose (n, n) element is u_n. We represent the Kronecker product as $\otimes$ and the sign function as $\mathrm{sign}(\cdot)$. For a lower semicontinuous function $\zeta : \mathbb{K}^N \to \mathbb{R} \cup \{\infty\}$, we define the Moreau envelope and the proximity operator as $\mathrm{env}_\zeta(\boldsymbol{u}) = \min_{\boldsymbol{s}\in\mathbb{K}^N} \left\{\zeta(\boldsymbol{s}) + \frac{1}{2}\|\boldsymbol{s}-\boldsymbol{u}\|_2^2\right\}$ and $\mathrm{prox}_\zeta(\boldsymbol{u}) = \arg\min_{\boldsymbol{s}\in\mathbb{K}^N} \left\{\zeta(\boldsymbol{s}) + \frac{1}{2}\|\boldsymbol{s}-\boldsymbol{u}\|_2^2\right\}$, respectively. $p_{\mathrm{G}}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$ and $P_{\mathrm{G}}(z) = \int_{-\infty}^{z} p_{\mathrm{G}}(z')dz'$ are the probability density function (PDF) and the cumulative distribution function (CDF) of the standard Gaussian distribution, respectively. When a sequence of random variables $\{Z_n\}$ $(n = 1, 2, \dots)$ converges in probability to Z, we denote $Z_n \xrightarrow{\mathrm{P}} Z$ as $n \to \infty$ or $\mathrm{plim}_{n\to\infty} Z_n = Z$.

1.1 Discrete-Valued Vector Reconstruction from Linear Measurements

1.1.1 Real-Valued Case

We consider the reconstruction of a discrete-valued vector $\boldsymbol{x} = [x_1 \ \cdots \ x_N]^{\mathsf{T}} \in \mathcal{R}^N \subset \mathbb{R}^N$ from its linear measurement given by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v} \in \mathbb{R}^M. \tag{1.1}$$

In this thesis, we mainly focus on the underdetermined case with $M < N$. Here, $\mathcal{R} = \{r_1, \dots, r_L\}$ is the set of possible values that the elements of the unknown vector $\boldsymbol{x}$ take, where $L \ll N$. The distribution of x_n is assumed to be known and given by $\Pr(x_n = r_\ell) = p_\ell$ $(\ell = 1, \dots, L)$, where $\sum_{\ell=1}^{L} p_\ell = 1$. $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ is a measurement matrix and $\boldsymbol{v} \in \mathbb{R}^M$ is an additive Gaussian noise vector with mean $\boldsymbol{0}$ and covariance matrix $\sigma_{\mathrm{v}}^2 \boldsymbol{I}_M$.

1.1.2 Complex-Valued Case

We consider the reconstruction of complex discrete-valued vector $\tilde{\boldsymbol{x}} = [\tilde{x}_1 \ \cdots \ \tilde{x}_N]^{\mathsf{T}} \in \mathcal{C}^N \subset \mathbb{C}^N$ from its linear measurement given by

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{v}} \in \mathbb{C}^M, \tag{1.2}$$

where $M < N$. Here, $\mathcal{C} = \{c_1, \dots, c_L\}$ is the set of possible values for the elements of the unknown vector $\tilde{\boldsymbol{x}}$. The distribution of $\tilde{\boldsymbol{x}}$ is given by $\Pr(\tilde{x}_n = c_\ell) = p_\ell$ $(\ell = 1, \dots, L)$, where $\sum_{\ell=1}^{L} p_\ell = 1$. $\tilde{\boldsymbol{A}} \in \mathbb{C}^{M\times N}$ is a measurement matrix and $\tilde{\boldsymbol{v}} \in \mathbb{C}^M$ is an additive Gaussian noise vector with mean $\boldsymbol{0}$ and covariance matrix $\sigma_{\mathrm{v}}^2 \boldsymbol{I}_M$.

The discrete-valued vector reconstruction algorithm in the real-valued domain is not appropriate for (1.2) in general. When the real part and the imaginary part are independent on $\mathcal{C}$, e.g., $\mathcal{C} = \{1 + j, -1 + j, -1 - j, 1 - j\}$, we can convert the signal

model (1.2) in the complex-valued domain into the equivalent model in the real-valued domain as

$$\check{\boldsymbol{y}} = \check{\boldsymbol{A}}\check{\boldsymbol{x}} + \check{\boldsymbol{v}}, \tag{1.3}$$

where we define $\check{\boldsymbol{y}} = \left[\mathrm{Re}\{\tilde{\boldsymbol{y}}\}^{\mathsf{T}}\ \mathrm{Im}\{\tilde{\boldsymbol{y}}\}^{\mathsf{T}}\right]^{\mathsf{T}} \in \mathbb{R}^{2M}$, $\check{\boldsymbol{x}} = \left[\mathrm{Re}\{\tilde{\boldsymbol{x}}\}^{\mathsf{T}}\ \mathrm{Im}\{\tilde{\boldsymbol{x}}\}^{\mathsf{T}}\right]^{\mathsf{T}} \in \mathbb{R}^{2N}$, $\check{\boldsymbol{v}} = \left[\mathrm{Re}\{\tilde{\boldsymbol{v}}\}^{\mathsf{T}}\ \mathrm{Im}\{\tilde{\boldsymbol{v}}\}^{\mathsf{T}}\right]^{\mathsf{T}} \in \mathbb{R}^{2M}$, and

$$\check{\boldsymbol{A}} = \begin{bmatrix} \mathrm{Re}\{\tilde{\boldsymbol{A}}\} & -\mathrm{Im}\{\tilde{\boldsymbol{A}}\} \\ \mathrm{Im}\{\tilde{\boldsymbol{A}}\} & \mathrm{Re}\{\tilde{\boldsymbol{A}}\} \end{bmatrix} \in \mathbb{R}^{2M\times 2N}. \tag{1.4}$$

In this case, we can reconstruct the original complex-valued vector $\tilde{\boldsymbol{x}}$ via the reconstruction of the real-valued vector $\check{\boldsymbol{x}}$. When the real part and the imaginary part are dependent, however, such approach is inappropriate. For example, when $C = \left\{e^{j(\ell-1)\pi/4} \mid \ell = 1, \ldots, 8\right\}$, we need to estimate the real-valued vector in $\left\{1, \frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}, -1\right\}^{2N}$. Hence, we cannot use the dependency between the real part and the imaginary part in the reconstruction. It would be better in such cases to directly reconstruct the vector $\tilde{\boldsymbol{x}}$ in the complex-valued domain.

1.2 Applications in Communication Systems

In this section, we present several applications of discrete-valued vector reconstruction in communication systems.

MIMO Signal Detection

MIMO communications use multiple antennas at both the transmitter and the receiver as in Fig. 1.1 to achieve high spectral efficiency and reliability. As the required data rate and throughput have been significantly increasing, massive MIMO using tens or hundreds of antennas are gathering attention as one of key technologies in the 5th generation (5G) mobile communication systems [1, 2].

MIMO signal detection is to estimate the transmitted symbols from the received signals, which is distorted by the channel and the additive noise. Since the transmitted symbols belong to an finite-sized alphabet in digital communications, MIMO signal detection with N_{t} transmit antennas and N_{r} receive antennas can be modeled as the discrete-valued vector reconstruction with $N = N_{\mathrm{t}}$ and $M = N_{\mathrm{r}}$. The unknown vector $\tilde{\boldsymbol{x}} \in C^{N_{\mathrm{t}}}$ is composed of the transmitted symbols from N_{t} transmit antennas, where C is the alphabet of transmitted symbols. For quadrature phase shift keying (QPSK), we have $(c_1, c_2, c_3, c_4) = (1 + j, -1 + j, -1 - j, 1 - j)$ and $p_\ell = 1/4$ $(\ell = 1, \ldots, 4)$. For 8-phase shift keying (8PSK), we have $c_\ell = e^{j(\ell-1)\pi/4}$ and $p_\ell = 1/8$ $(\ell = 1, \ldots, 8)$. For simplicity,

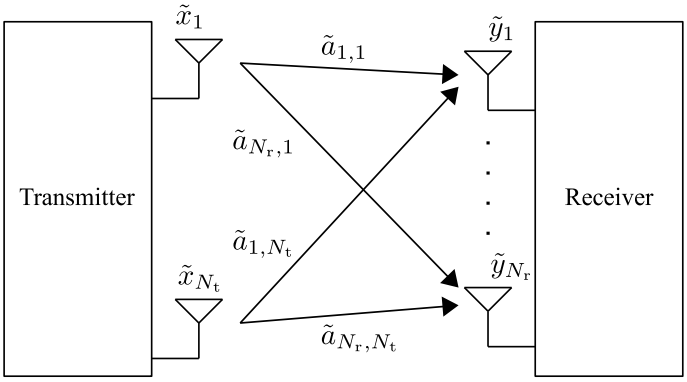


Figure 1.1: System model of MIMO communications

precoding is not considered and the number of transmitted streams is assumed to be equal to that of transmit antennas in this thesis. The measurement vector $\tilde{\boldsymbol{y}} \in \mathbb{C}^{N_\mathrm{r}}$ denotes the received signals at N_r receive antennas and $\tilde{\boldsymbol{A}} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ represents the channel matrix between the transmitter and the receiver.

The distribution of the channel matrix $\tilde{\boldsymbol{A}}$ depends on the channel model. In uncorrelated flat Rayleigh fading channels, the channel matrix is given by $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}}_{\mathrm{i.i.d.}}$, where $\tilde{\boldsymbol{A}}_{\mathrm{i.i.d.}} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ is composed of i.i.d. circular complex Gaussian variables with zero mean and unit variance. For spatially correlated MIMO channels with equally spaced linear arrays, the channel matrix can be modeled as $\tilde{\boldsymbol{A}} = \boldsymbol{\Psi}_\mathrm{r}^{\frac{1}{2}} \tilde{\boldsymbol{A}}_{\mathrm{i.i.d.}} \boldsymbol{\Psi}_\mathrm{t}^{\frac{1}{2}}$ [11]. Here, the (i_1, i_2) element of $\boldsymbol{\Psi}_\mathrm{r}$ and $\boldsymbol{\Psi}_\mathrm{t}$ are given by $[\boldsymbol{\Psi}_\mathrm{r}]_{i_1,i_2} = J_0\left(|i_1 - i_2| \cdot 2\pi d_\mathrm{r}/\lambda_\mathrm{w}\right)$ and $[\boldsymbol{\Psi}_\mathrm{t}]_{i_1,i_2} = J_0\left(|i_1 - i_2| \cdot 2\pi d_\mathrm{t}/\lambda_\mathrm{w}\right)$, respectively. The function $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind. We denote the wavelength by λ_w and the antenna spacing at the receiver and the transmitter by d_r and d_t, respectively. For other channel models, see [11].

In some MIMO systems, sufficient number of receive antennas may not be available due to the limited size, weight, cost and/or power consumption of the receiver. Such MIMO systems, where the number of receive antennas N_r is less than that of transmitted streams N_t, are known as overloaded (or underdetermined) MIMO systems [5, 7]. In overloaded MIMO systems, $\tilde{\boldsymbol{A}} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ $(N_\mathrm{r} < N_\mathrm{t})$ is a fat matrix and hence the signal detection problem becomes underdetermined.

Error Recovery of MIMO Signal Detection

To improve the performance of the MIMO signal detection, some error recovery method have been discussed [12, 13]. In these methods, the system model (1.2) is converted into the linear equation of the error vector. Let $\hat{\boldsymbol{x}} \in \mathbb{C}^N$ be a tentative estimate of $\tilde{\boldsymbol{x}}$ obtained by some simple detection method such as the LMMSE method. We then obtain $\hat{\boldsymbol{x}}_\mathrm{d} = Q_C(\hat{\boldsymbol{x}}) \in C^N$, where the element-wise function $Q_C(\cdot)$ maps each element into its

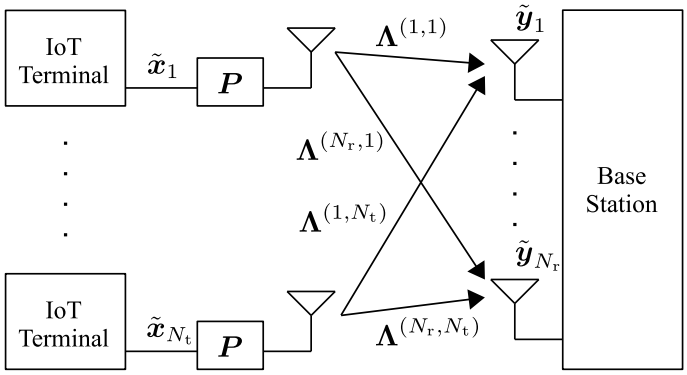


Figure 1.2: Uplink MU-MIMO OFDM system for IoT environment

closest symbol in C, i.e., it provides the hard decision of $\tilde{\boldsymbol{x}}$. The key point is that the error vector $\tilde{\boldsymbol{e}} = \tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}_{\mathrm{d}}$ is sparse, i.e., $\tilde{\boldsymbol{e}}$ has many zero elements, if the tentative estimate is reliable enough. Moreover, the error vector $\tilde{\boldsymbol{e}}$ also has the discreteness. For example, when we use QPSK with $(c_1, c_2, c_3, c_4) = (1 + j, -1 + j, -1 - j, 1 - j)$, the real part and the imaginary part of the error vector contain only 2, 0, and -2. The transformation into the equation of $\tilde{\boldsymbol{e}}$ can be performed by subtracting $\tilde{\boldsymbol{A}}\hat{\boldsymbol{x}}_{\mathrm{d}}$ from both sides of (I.2) as

$$\tilde{\boldsymbol{y}}_{\mathrm{d}} \coloneqq \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\hat{\boldsymbol{x}}_{\mathrm{d}} = \tilde{\boldsymbol{A}}\left(\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}_{\mathrm{d}}\right) + \tilde{\boldsymbol{v}} \tag{1.5}$$

$$= \tilde{\boldsymbol{A}}\tilde{\boldsymbol{e}} + \tilde{\boldsymbol{v}}. \tag{1.6}$$

From (I.6), we can reconstruct the error vector $\tilde{\boldsymbol{e}}$ via some algorithm for the sparse discrete-valued reconstruction algorithm. Denoting the estimate of the error vector $\tilde{\boldsymbol{e}}$ as $\hat{\boldsymbol{e}}$, we can obtain the improved estimate of $\tilde{\boldsymbol{x}}$ as $\hat{\boldsymbol{x}}_{\mathrm{d}} + \hat{\boldsymbol{e}}$.

Signal Detection for MU-MIMO OFDM/SC-CP

We consider uplink communications of Internet of things (IoT) environments, which is modeled as a precoded multi user multiple-input multiple-output (MU-MIMO) orthogonal frequency division multiplexing (OFDM) system. Figure I.2 shows the system model, where the number of transmit terminals is N_{t}, the number of receive antennas at the base station is N_{r}, and the number of subcarriers is Q_{c}. Given that the number of transmit terminal is typically large in IoT environments, we focus on the overloaded scenario and assume $N_{\mathrm{r}} < N_{\mathrm{t}}$ hereafter. The symbol alphabet and the frequency domain transmitted OFDM symbol vector from the n_{t}-th transmit IoT terminal are denoted by C and $\tilde{\boldsymbol{x}}_{n_{\mathrm{t}}}$, respectively. Here, taking IoT environment specific feature into consideration, we assume only N_{act} IoT terminals out of N_{t} terminals are active meaning that only N_{act} terminals transmit OFDM signal blocks. Non-active $N_{\mathrm{t}} - N_{\mathrm{act}}$ terminals actually keep silent, but we can regard they transmit all zero signal block $\boldsymbol{0}_{Q_{\mathrm{c}}}$. We thus have $\tilde{\boldsymbol{x}}_{n_{\mathrm{t}}} \in C^{Q_{\mathrm{c}}}$ when the n-th terminal is active, and otherwise $\tilde{\boldsymbol{x}}_{n_{\mathrm{t}}} = \boldsymbol{0}_{Q_{\mathrm{c}}}$. When we use the cyclic prefix

with the length greater than or equal to the channel order, the received signal vector after the removal of the cyclic prefix is given by

$$\begin{bmatrix} \tilde{\boldsymbol{y}}_1^{\mathrm{f,OFDM}} \\ \vdots \\ \tilde{\boldsymbol{y}}_{N_\mathrm{r}}^{\mathrm{f,OFDM}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}^{(1,1)}\boldsymbol{P} & \cdots & \boldsymbol{\Lambda}^{(1,N_\mathrm{t})}\boldsymbol{P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Lambda}^{(N_\mathrm{r},1)}\boldsymbol{P} & \cdots & \boldsymbol{\Lambda}^{(N_\mathrm{r},N_\mathrm{t})}\boldsymbol{P} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{x}}_1 \\ \vdots \\ \tilde{\boldsymbol{x}}_{N_\mathrm{t}} \end{bmatrix} + \begin{bmatrix} \tilde{\boldsymbol{v}}_1^{\mathrm{f}} \\ \vdots \\ \tilde{\boldsymbol{v}}_{N_\mathrm{r}}^{\mathrm{f}} \end{bmatrix}, \tag{1.7}$$

where $\tilde{\boldsymbol{y}}_{n_\mathrm{r}}^{\mathrm{f,OFDM}} \in \mathbb{C}^{Q_\mathrm{c}}$ is the frequency domain received OFDM signal block at the n_r-th receive antenna [14]. The diagonal matrix $\boldsymbol{\Lambda}^{(n_\mathrm{r},n_\mathrm{t})} = \mathrm{diag}\left(\lambda_1^{(n_\mathrm{r},n_\mathrm{t})}, \ldots, \lambda_{Q_\mathrm{c}}^{(n_\mathrm{r},n_\mathrm{t})}\right) \in \mathbb{C}^{Q_\mathrm{c}\times Q_\mathrm{c}}$ is composed of the channel frequency responses with the order of $L_\mathrm{p}-1$ between the n_t-th IoT terminal and the n_r-th receive antenna. The diagonal elements can be written as

$$\begin{bmatrix} \lambda_1^{(n_\mathrm{r},n_\mathrm{t})} \\ \vdots \\ \lambda_{Q_\mathrm{c}}^{(n_\mathrm{t},n_\mathrm{t})} \end{bmatrix} = \sqrt{Q_\mathrm{c}}\boldsymbol{D} \begin{bmatrix} h_0^{(n_\mathrm{r},n_\mathrm{t})} \\ \vdots \\ h_{L_\mathrm{p}-1}^{(n_\mathrm{r},n_\mathrm{t})} \\ \boldsymbol{0}_{Q_\mathrm{c}-L_\mathrm{p}} \end{bmatrix}, \tag{1.8}$$

where $\boldsymbol{D} \in \mathbb{C}^{Q_\mathrm{c}\times Q_\mathrm{c}}$ is a Q_c-point unitary discrete Fourier transform (DFT) matrix defined as

$$\boldsymbol{D} = \frac{1}{\sqrt{Q_\mathrm{c}}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-j\frac{2\pi\times 1\times 1}{Q_\mathrm{c}}} & \cdots & e^{-j\frac{2\pi\times 1\times (Q_\mathrm{c}-1)}{Q_\mathrm{c}}} \\ \vdots & \vdots & & \vdots \\ 1 & e^{-j\frac{2\pi\times (Q_\mathrm{c}-1)\times 1}{Q_\mathrm{c}}} & \cdots & e^{-j\frac{2\pi\times (Q_\mathrm{c}-1)\times (Q_\mathrm{c}-1)}{Q_\mathrm{c}}} \end{bmatrix} \tag{1.9}$$

and $h_0^{(n_\mathrm{r},n_\mathrm{t})}, \ldots, h_{L_\mathrm{p}-1}^{(n_\mathrm{r},n_\mathrm{t})}$ denotes the impulse response of the frequency selective channel between the n_t-th IoT terminal and n_r-th receive antenna. $\boldsymbol{P} \in \mathbb{C}^{Q_\mathrm{c}\times Q_\mathrm{c}}$ is a precoding matrix. $\boldsymbol{v}_{n_\mathrm{r}}^{\mathrm{f}} \in \mathbb{C}^{Q_\mathrm{c}}$ is the frequency domain additive white noise vector at the n_r-th receive antenna with mean $\boldsymbol{0}_{Q_\mathrm{c}}$ and covariance matrix $\sigma_\mathrm{v}^2\boldsymbol{I}_{Q_\mathrm{c}}$.

Here, we show a non-precoded MU-MIMO single carrier block transmission with cyclic prefix (SC-CP) signal model. Assuming the length of cyclic prefix is greater than or equal to the channel order $L_\mathrm{p}-1$, the time domain received signal block at the n_r-th receive antenna of the base station is written as

$$\tilde{\boldsymbol{y}}_{n_\mathrm{r}}^{\mathrm{t,SC\text{-}CP}} = \sum_{n_\mathrm{t}=1}^{N_\mathrm{t}} \boldsymbol{D}^{\mathsf{H}}\boldsymbol{\Lambda}^{(n_\mathrm{r},n_\mathrm{t})}\boldsymbol{D}\tilde{\boldsymbol{x}}_{n_\mathrm{t}} + \tilde{\boldsymbol{v}}_{n_\mathrm{r}}^{\mathrm{t}}, \tag{1.10}$$

where $\tilde{\boldsymbol{v}}_{n_\mathrm{t}}^{\mathrm{t}} \in \mathbb{C}^{Q_\mathrm{c}}$ is the time domain additive white noise vector at the n_r-th receive antenna having mean $\boldsymbol{0}_{Q_\mathrm{c}}$ and covariance matrix $\sigma_\mathrm{v}^2 \boldsymbol{I}_{Q_\mathrm{c}}$ [15, 16]. By stacking from $\tilde{\boldsymbol{y}}_1^{\mathrm{t,SC\text{-}CP}}$ to $\tilde{\boldsymbol{y}}_{N_\mathrm{r}}^{\mathrm{t,SC\text{-}CP}}$ in (1.10), and multiplying a unitary matrix of

$$\begin{bmatrix} \boldsymbol{D} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & & \vdots \\ \vdots & & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} \in \mathbb{C}^{Q_\mathrm{c} N_\mathrm{r} \times Q_\mathrm{c} N_\mathrm{r}} \tag{1.11}$$

from the left of both sides, we have the frequency domain received non-precoded SC-CP IoT signal vector at the base station as

$$\begin{bmatrix} \tilde{\boldsymbol{y}}_1^{\mathrm{f,SC\text{-}CP}} \\ \vdots \\ \tilde{\boldsymbol{y}}_{N_\mathrm{r}}^{\mathrm{f,SC\text{-}CP}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & & \vdots \\ \vdots & & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{y}}_1^{\mathrm{t,SC\text{-}CP}} \\ \vdots \\ \tilde{\boldsymbol{y}}_{N_\mathrm{r}}^{\mathrm{t,SC\text{-}CP}} \end{bmatrix} \tag{1.12}$$

$$= \begin{bmatrix} \boldsymbol{\Lambda}^{(1,1)} \boldsymbol{D} & \cdots & \boldsymbol{\Lambda}^{(1,N_\mathrm{t})} \boldsymbol{D} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Lambda}^{(N_\mathrm{r},1)} \boldsymbol{D} & \cdots & \boldsymbol{\Lambda}^{(N_\mathrm{r},N_\mathrm{t})} \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{x}}_1 \\ \vdots \\ \tilde{\boldsymbol{x}}_{N_\mathrm{t}} \end{bmatrix} + \begin{bmatrix} \tilde{\boldsymbol{v}}_1^{\mathrm{f}} \\ \vdots \\ \tilde{\boldsymbol{v}}_{N_\mathrm{r}}^{\mathrm{f}} \end{bmatrix}, \tag{1.13}$$

where $\tilde{\boldsymbol{y}}_{n_\mathrm{r}}^{\mathrm{f,SC\text{-}CP}} \in \mathbb{C}^{Q_\mathrm{c}}$ is the frequency domain received SC-CP signal vector at the n_r-th base station antenna and $\tilde{\boldsymbol{v}}_{n_\mathrm{r}}^{\mathrm{f}} = \boldsymbol{D} \tilde{\boldsymbol{v}}_{n_\mathrm{r}}^{\mathrm{t}}$ $(n_\mathrm{r} = 1, \ldots, N_\mathrm{r})$. It should be noted here that this received signal model can be regarded as a special case of (1.7), where the precoding matrix $\boldsymbol{P}$ is set to be $\boldsymbol{D}$. Thus, if DFT matrix $\boldsymbol{D}$ is appropriate for the precoding matrix of overloaded MU-MIMO OFDM system with the convex optimization-based signal detection, then the choice of non-precoded SC-CP signaling is extremely suited for IoT environments because this approach requires neither the inverse discrete Fourier transform (IDFT) operation nor the precoding operation at the IoT node (transmitter side).

Channel Equalization

Channel equalization in the single carrier block transmission [15] can also be modeled as the complex discrete-valued vector reconstruction. We here consider a MIMO system with N_t transmit antennas and N_r receive antennas. When we use the cyclic prefix to remove inter-block interference, the resultant channel matrix $\tilde{\boldsymbol{A}} \in \mathbb{C}^{N_\mathrm{r} Q_\mathrm{b} \times N_\mathrm{t} Q_\mathrm{b}}$ corresponding to an information block can be written as a block circulant matrix given

by

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{\Gamma}^{(0)} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{\Gamma}^{(L_\mathrm{p}-1)} & \cdots & \boldsymbol{\Gamma}^{(1)} \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \boldsymbol{\Gamma}^{(L_\mathrm{p}-2)} & & \ddots & \ddots & & \ddots & \boldsymbol{\Gamma}^{(L_\mathrm{p}-1)} \\ \boldsymbol{\Gamma}^{(L_\mathrm{p}-1)} & \ddots & & \ddots & \ddots & & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{\Gamma}^{(L_\mathrm{p}-1)} & \boldsymbol{\Gamma}^{(L_\mathrm{p}-2)} & \cdots & \boldsymbol{\Gamma}^{(0)} \end{bmatrix}, \tag{1.14}$$

where Q_b is the information block length,

$$\boldsymbol{\Gamma}^{(i)} = \begin{bmatrix} \gamma_{1,1}^{(i)} & \cdots & \gamma_{1,N_\mathrm{t}}^{(i)} \\ \vdots & \ddots & \vdots \\ \gamma_{N_\mathrm{r},1}^{(i)} & \cdots & \gamma_{N_\mathrm{r},N_\mathrm{t}}^{(i)} \end{bmatrix} \in \mathbb{C}^{N_\mathrm{r}\times N_\mathrm{t}}, \tag{1.15}$$

and $\left\{\gamma_{n_\mathrm{r},n_\mathrm{t}}^{(i)}\right\}$ $(i = 0, \ldots, L_\mathrm{p}-1)$ is the impulse response of the channel between the n_tth transmit antenna and the n_rth receive antenna ($n_\mathrm{t} = 1, \ldots, N_\mathrm{t}$ and $n_\mathrm{r} = 1, \ldots, N_\mathrm{r}$) [17,18]. In the Rayleigh fading channels, $\gamma_{n_\mathrm{r},n_\mathrm{t}}^{(i)}$ is a circular complex Gaussian variable with zero mean. It should be noted that the channel equalization problem also becomes underdetermined in overloaded MIMO systems.

Decoding of NO-STBC

For MIMO communications, non-orthogonal space-time block code (NO-STBC) has been studied to achieve both high rate and high diversity order [19]. In [20], for example, a NO-STBC has been proposed by using cyclic division algebra (CDA), which can achieve both the full diversity and the information losslessness under the ML decoding. Moreover, the rate of the code is equal to the number of transmit antennas.

We consider MIMO communications using space-time block code (STBC) with N_t transmit antennas and N_r receive antennas. By using a STBC, we send K_d complex data symbols $\tilde{x}_1, \ldots, \tilde{x}_{K_\mathrm{d}} \in \mathbb{C}$ during T_d time slots. We define the STBC matrix as $\tilde{\boldsymbol{B}} = \left[\tilde{\boldsymbol{b}}_1 \cdots \tilde{\boldsymbol{b}}_{T_\mathrm{d}}\right] \in \mathbb{C}^{N_\mathrm{t}\times T_\mathrm{d}}$, where $\tilde{\boldsymbol{b}}_t = [\tilde{b}_{1,t} \cdots \tilde{b}_{N_\mathrm{t},t}]^\mathsf{T} \in \mathbb{C}^{N_\mathrm{t}}$ $(t = 1, \ldots, T_\mathrm{d})$ indicates the transmitted signal vector at the tth time slot and $\tilde{b}_{n_\mathrm{t},t}$ is the transmitted symbol from the n_tth transmit antenna ($n_\mathrm{t} = 1, \ldots, N_\mathrm{t}$). In linear dispersion STBCs, the STBC matrix $\tilde{\boldsymbol{B}}$ is given by

$$\tilde{\boldsymbol{B}} = \sum_{k=1}^{K_\mathrm{d}} \tilde{\boldsymbol{C}}_k \tilde{x}_k, \tag{1.16}$$

where $\tilde{\boldsymbol{C}}_k \in \mathbb{C}^{N_\mathrm{t} \times T_\mathrm{d}}$ is a weight matrix corresponding to the data symbol $\tilde{s}_k$. In [20], for example, the NO-STBC matrix

$$\tilde{\boldsymbol{B}} = \sum_{n_\mathrm{t}=0}^{N_\mathrm{t}-1} \begin{bmatrix} \bar{x}_{0,n_\mathrm{t}} & \delta \bar{x}_{N_\mathrm{t}-1,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{n_\mathrm{t}} & \cdots & \delta \bar{x}_{1,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{(N_\mathrm{t}-1)n_\mathrm{t}} \\ \bar{x}_{1,n_\mathrm{t}} & \bar{x}_{0,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{n_\mathrm{t}} & \cdots & \delta \bar{x}_{2,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{(N_\mathrm{t}-1)n_\mathrm{t}} \\ \bar{x}_{2,n_\mathrm{t}} & \bar{x}_{1,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{n_\mathrm{t}} & \cdots & \delta \bar{x}_{3,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{(N_\mathrm{t}-1)n_\mathrm{t}} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{N_\mathrm{t}-2,n_\mathrm{t}} & \bar{x}_{N_\mathrm{t}-3,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{n_\mathrm{t}} & \cdots & \delta \bar{x}_{N_\mathrm{t}-1,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{(N_\mathrm{t}-1)n_\mathrm{t}} \\ \bar{x}_{N_\mathrm{t}-1,n_\mathrm{t}} & \bar{x}_{N_\mathrm{t}-2,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{n_\mathrm{t}} & \cdots & \bar{x}_{0,n_\mathrm{t}} \omega_{N_\mathrm{t}}^{(N_\mathrm{t}-1)n_\mathrm{t}} \end{bmatrix} \rho^{n_\mathrm{t}} \tag{1.17}$$

has been proposed by using CDA, where $\bar{x}_{n_\mathrm{t},n_\mathrm{t}'} = \tilde{x}_{n_\mathrm{t}N_\mathrm{t}+n_\mathrm{t}'+1} \in \mathbb{C}$ $(n_\mathrm{t}, n_\mathrm{t}' = 0, \ldots, N_\mathrm{t}-1)$ are the complex data symbols to be sent, and $\omega_{N_\mathrm{t}} = e^{j\frac{2\pi}{N_\mathrm{t}}}$. Since we use $T_\mathrm{d} = N_\mathrm{t}$ time slots to send $K_\mathrm{d} = N_\mathrm{t}^2$ symbols in (1.17), the rate of this NO-STBC is $K_\mathrm{d}/T_\mathrm{d} = N_\mathrm{t}$. Moreover, when $\delta = e^{\sqrt{5}j}$ and $\rho = e^{j}$, the full diversity is also achieved under ML decoding [20].

The decoding of NO-STBC [21–24] to estimate the transmitted symbols can also be regarded as the discrete-valued vector reconstruction. The received signal matrix $\tilde{\boldsymbol{Y}} \in \mathbb{C}^{N_\mathrm{r} \times T_\mathrm{d}}$ corresponding to $\tilde{\boldsymbol{B}}$ during T_d time slots is given by

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{H}}\tilde{\boldsymbol{B}} + \tilde{\boldsymbol{V}}, \tag{1.18}$$

where $\tilde{\boldsymbol{H}} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ is the channel matrix and $\tilde{\boldsymbol{V}} \in \mathbb{C}^{N_\mathrm{r} \times T_\mathrm{d}}$ is the zero mean additive white Gaussian noise matrix. From (1.16) and (1.18), we have $\tilde{\boldsymbol{Y}} = \sum_{k=1}^{K_\mathrm{d}} \tilde{\boldsymbol{H}}\tilde{\boldsymbol{C}}_k \tilde{x}_k + \tilde{\boldsymbol{V}}$ and hence $\tilde{\boldsymbol{y}} := \mathrm{vec}(\tilde{\boldsymbol{Y}}) \in \mathbb{C}^{N_\mathrm{r}T_\mathrm{d}}$ can be written as

$$\tilde{\boldsymbol{y}} = \sum_{k=1}^{K_\mathrm{d}} (\boldsymbol{I}_{T_\mathrm{d}} \otimes \tilde{\boldsymbol{H}}) \mathrm{vec}(\tilde{\boldsymbol{C}}_k) \tilde{x}_k + \mathrm{vec}(\tilde{\boldsymbol{V}}) \tag{1.19}$$

$$= (\boldsymbol{I}_{T_\mathrm{d}} \otimes \tilde{\boldsymbol{H}}) \tilde{\boldsymbol{C}} \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{v}} \tag{1.20}$$

$$= \tilde{\boldsymbol{A}} \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{v}}, \tag{1.21}$$

where $\tilde{\boldsymbol{x}} = [\tilde{x}_1 \cdots \tilde{x}_{K_\mathrm{d}}]^\mathsf{T} \in \mathbb{C}^{K_\mathrm{d}}$, $\tilde{\boldsymbol{v}} = \mathrm{vec}(\tilde{\boldsymbol{V}}) \in \mathbb{C}^{N_\mathrm{r}T_\mathrm{d}}$, $\tilde{\boldsymbol{C}} = \left[\mathrm{vec}(\tilde{\boldsymbol{C}}_1) \cdots \mathrm{vec}(\tilde{\boldsymbol{C}}_{K_\mathrm{d}})\right] \in \mathbb{C}^{N_\mathrm{t}T_\mathrm{d} \times K_\mathrm{d}}$, and $\tilde{\boldsymbol{A}} = (\boldsymbol{I}_{T_\mathrm{d}} \otimes \tilde{\boldsymbol{H}})\tilde{\boldsymbol{C}} \in \mathbb{C}^{N_\mathrm{r}T_\mathrm{d} \times K_\mathrm{d}}$ [21]. Note that the size of the effective channel matrix $\tilde{\boldsymbol{A}} \in \mathbb{C}^{N_\mathrm{r}T_\mathrm{d} \times K_\mathrm{d}}$ is much larger than that of $\tilde{\boldsymbol{H}} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$. When we use the NO-STBC given by (1.17) and assume the overloaded scenario with $N_\mathrm{r} < N_\mathrm{t}$, the decoding is an underdetermined problem because $K = N_\mathrm{t}T_\mathrm{d} > N_\mathrm{r}T_\mathrm{d}$ and hence $\tilde{\boldsymbol{A}}$ becomes a fat matrix.

Multiuser Detection

Multiuser detection is an important issue in M2M communications, where a number of transmit node simultaneously transmit signals with low data rates [4, 25, 26]. We here

consider the following received signal model

$$y(t) = \sum_{n=1}^{N_\mathrm{t}} h_n x_n s_n(t) + v(t), \tag{1.22}$$

in the transmission period $\left[0, T_\mathrm{p}\right]$, where $x_n \in \{1, 0, -1\}$ and $s_n(t)$ are the transmitted symbol and the signature waveform of the nth transmit node, respectively. h_n is the channel gain between the nth node and the receiver, and $v(n)$ is the additive white Gaussian noise (AWGN) with zero mean. Note that $x_n = 0$ means that the nth node is not active in the transmission period.

Multiuser detection to reconstruct the transmitted symbols x_n can be considered as the discrete-valued vector reconstruction. When we use M_f filters $\varphi_m(t)$ $(m = 1, \ldots, M_\mathrm{f})$ at the receiver, the output of the filter is $y_m \coloneqq \int_0^{T_\mathrm{p}} y(t)\varphi_m(T - t)dt$. Letting $s_{m,n} \coloneqq \int_0^{T_\mathrm{p}} s_n(t)\varphi_m(T - t)dt$ and $v_m \coloneqq \int_0^{T_\mathrm{p}} v(t)\varphi_m(T - t)dt$, we have

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{v}, \tag{1.23}$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 & \cdots & y_{M_\mathrm{f}} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{M_\mathrm{f}}$, $\boldsymbol{H} = \mathrm{diag}(h_1, \ldots, h_{N_\mathrm{t}}) \in \mathbb{R}^{N_\mathrm{t} \times N_\mathrm{t}}$, $\boldsymbol{x} = \begin{bmatrix} x_1 & \cdots & x_{N_\mathrm{t}} \end{bmatrix}^\mathsf{T} \in \{1, -1\}^{N_\mathrm{t}}$, $\boldsymbol{v} = \begin{bmatrix} v_1 & \cdots & v_{M_\mathrm{f}} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{M_\mathrm{f}}$, and

$$\boldsymbol{S} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,N_\mathrm{t}} \\ \vdots & \ddots & \vdots \\ s_{M_\mathrm{f},1} & \cdots & s_{M_\mathrm{f},N_\mathrm{t}} \end{bmatrix} \in \mathbb{R}^{M_\mathrm{f} \times N_\mathrm{t}}. \tag{1.24}$$

FTN Signaling

To achieve high speed data transmission, FTN signaling has attracted much attention [10, 27–29]. In FTN signaling, the transmitter transmits signals beyond the Nyquist rate. For example, when we consider the binary phase shift keying (BPSK) signals $x_1, \ldots, x_{N_\mathrm{s}} \in \{1, -1\}$, the modulated signal in the transmission period $\left[0, T_\mathrm{p}\right]$ can be written as

$$x(t) = \sum_{n=1}^{N_\mathrm{s}} x_n a_n(t), \tag{1.25}$$

where N_s is the number of transmitted symbols, T_p is the interval of one period, and $a_n(t)$ $(n = 1, \ldots, N_\mathrm{s})$ is the modulation pulse. Hence, the received signal through the AWGN channel is given by

$$y(t) = \sum_{n=1}^{N_\mathrm{s}} x_n a_n(t) + v(t), \tag{1.26}$$

where $v(t)$ is the AWGN with zero mean.

Detection of the transmitted symbols x_n from the received signal $y(t)$ results in the discrete-valued vector reconstruction. Let $\varphi_m(t)$ $(m = 1, \ldots, M_\mathrm{b})$ be an orthogonal basis in the time-frequency space to which $x(t)$ belongs. We also define $y_m \coloneqq \langle y(t), \varphi_m(t) \rangle$, $a_{m,n} \coloneqq \langle a_n(t), \varphi_m(t) \rangle$, and $v_m \coloneqq \langle v(t), \varphi_m(t) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product. We then obtain

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}, \tag{1.27}$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 & \cdots & y_{M_\mathrm{b}} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{M_\mathrm{b}}$, $\boldsymbol{x} = \begin{bmatrix} x_1 & \cdots & x_{N_\mathrm{s}} \end{bmatrix}^\mathsf{T} \in \{1, -1\}^{N_\mathrm{s}}$, $\boldsymbol{v} = \begin{bmatrix} v_1 & \cdots & v_{M_\mathrm{b}} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{M_\mathrm{b}}$, and

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,N_\mathrm{s}} \\ \vdots & \ddots & \vdots \\ a_{M_\mathrm{b},1} & \cdots & a_{M_\mathrm{b},N_\mathrm{s}} \end{bmatrix} \in \mathbb{R}^{M_\mathrm{b} \times N_\mathrm{s}}. \tag{1.28}$$

Since we have $M_\mathrm{b} < N_\mathrm{s}$ in FTN signaling, the detection problem can be regarded as the underdetermined discrete-valued vector reconstruction.

1.3 Conventional Methods

In this section, we briefly review several conventional approaches for the discrete-valued vector reconstruction.

LMMSE Method

Linear reconstruction methods obtain the estimate of the unknown vector $\tilde{\boldsymbol{x}}$ as $\hat{\boldsymbol{x}} = \boldsymbol{W}\tilde{\boldsymbol{y}}$, where $\boldsymbol{W} \in \mathbb{C}^{N \times M}$ is a weight matrix. In the LMMSE method, for example, the weight matrix $\boldsymbol{W}_\mathrm{LMMSE}$ is determined by

$$\boldsymbol{W}_\mathrm{LMMSE} = \underset{\boldsymbol{W} \in \mathbb{C}^{N \times M}}{\arg\min}\ \mathrm{E}\left[\| \boldsymbol{W}\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{x}} \|_2^2 \right] \tag{1.29}$$

$$= \boldsymbol{R}_\mathrm{x} \tilde{\boldsymbol{A}}^\mathsf{H} \left(\tilde{\boldsymbol{A}} \boldsymbol{R}_\mathrm{x} \tilde{\boldsymbol{A}}^\mathsf{H} + \sigma_\mathrm{v}^2 \boldsymbol{I}_M \right)^{-1}, \tag{1.30}$$

where $\boldsymbol{R}_\mathrm{x} = \mathrm{E}\left[\tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^\mathsf{H} \right]$. Although the linear reconstruction methods have low computational complexity, the reconstruction performance becomes poor in underdetermined problems.

Maximum Likelihood Method

The ML method obtains the vector $\boldsymbol{s} \in \mathcal{C}^N$ that maximizes the likelihood function

$$p\left(\tilde{\boldsymbol{y}} \mid \tilde{\boldsymbol{x}} = \boldsymbol{s}\right) \propto \exp\left(-\frac{\left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\boldsymbol{s}\right\|_2^2}{\sigma_{\mathrm{v}}^2}\right). \tag{1.31}$$

This approach is equivalent to the minimization problem

$$\underset{\boldsymbol{s}\in\mathcal{C}^N}{\text{minimize}} \ \left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\boldsymbol{s}\right\|_2^2. \tag{1.32}$$

The ML method can achieve the optimal performance in terms of the symbol error rate (SER) when the distribution of the unknown vector $\boldsymbol{x}$ is uniform. However, since the optimization in (1.32) is a combinatorial optimization problem, the required computational complexity becomes prohibitive in large-scale problems. Although some complexity reduction methods such as sphere decoding [30, 31] and slab sphere decoding [5] have been proposed, their complexity is still huge for tens or hundreds of N.

Local Neighborhood Search

In the context of MIMO signal detection and the decoding of STBC, several reconstruction methods have been proposed on the basis of local neighborhood search. These algorithms starts from an initial estimate and updates each element in an iterative manner. The likelihood ascent search (LAS) [21, 32, 33] simply updates the element so that we have a larger likelihood. The reactive tabu search (RTS) [34] incorporate the tabu search technique to escape from the local optima. Since the performance of these methods severely degrade for overloaded scenario because of many local optima, the enhanced reactive tabu search (ERTS) has been proposed in [7]. ERTS is an extension of RTS and employs RTS iteratively while randomly varying the initial point of the search until a certain condition is satisfied. It is shown in [7] that ERTS can achieve comparable performance to the optimal ML detection with affordable computational complexity for overloaded MIMO systems with around 30 transmit antennas. If the number of antennas further increases, however, ERTS requires prohibitive computational complexity to achieve such performance because the required number of RTSs significantly increases.

Message Passing-Based Methods

Low complexity algorithms have been proposed for the discrete-valued vector reconstruction on the basis of belief propagation (BP) [35, 36]. For binary vector reconstruction, approximated BP has been proposed with application to code division multiple access (CDMA) multiuser detection [37]. A similar algorithm named approximate message passing (AMP) [38, 39] has also been proposed for compressed sensing [40, 41], and then

applied to discrete-valued vector reconstruction [42] and more general scenario [43]. The AMP-based methods can be used for complex-valued vectors and its asymptotic performance in the large system limit can be predicted by state evolution technique [38, 44]. However, the empirical performance of the AMP algorithm is degraded for small-scale problems because of the cycles in the factor graph considered in the original BP [45, 46]. Moreover, we require an assumption of the i.i.d. zero mean Gaussian measurement matrix in the derivation of the algorithm. In fact, the performance of the AMP algorithm might severely degrade for general measurement matrices. For example, the convergence of the AMP algorithm becomes unstable for nonzero mean i.i.d. matrix [47]. To improve the stability of the AMP algorithm, several techniques such as adaptive damping, mean removal, and sequential update have been proposed [47, 48]. The convergence of generalized approximate message passing (GAMP) [43] with the appropriate damping have also been discussed in [49] for general measurement matrices under several assumptions for the distributions of the unknown vector and the measurement noise. To overcome this unstability of the AMP algorithm, other message passing-based approaches, e.g., expectation propagation (EP) [50], vector approximate message passing (VAMP) [51], and orthogonal approximate message passing (OAMP) [52], have also been proposed. The asymptotic performance of these algorithms has been theoretically analyzed for unitary invariant measurement matrices [51, 53], which is a wider class than i.i.d. Gaussian matrices.

Convex Optimization-Based Methods

Although the ML method can achieve excellent performance, the required computational complexity becomes prohibitive when the problem size is large. To tackle this problem, several convex optimization-based approaches have been proposed for the discrete-valued vector reconstruction in the real-valued domain.

The box-relaxation method [54, 55] is a convex relaxation of the ML method under the hypercube containing all possible discrete-valued vectors. In the real-valued case, the box relaxation method uses the box constraint $\boldsymbol{s} \in [r_1, r_L]^N$ as

$$\underset{\boldsymbol{s}\in[r_1,r_L]^N}{\text{minimize}} \ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 \tag{1.33}$$

because the unknown vector satisfies $\boldsymbol{x} \in \mathcal{R}^N \subset [r_1, r_L]^N$. Since both of the objective function and the feasible region are convex, the optimization problem can be solved with several convex optimization techniques. The asymptotic SER of the box relaxation method has been derived in [56] by using the convex Gaussian min-max theorem (CGMT) framework.

In [57], the regularization-based method given by

$$\underset{\boldsymbol{s}\in\mathbb{R}^N}{\text{minimize}} \sum_{\ell=1}^{L} \|\boldsymbol{s} - r_\ell \mathbf{1}\|_1 \ \text{ subject to } \ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{s} \tag{1.34}$$

has been proposed for the discrete-valued vector reconstruction in the noise-free case. This method uses the regularizer $\sum_{\ell=1}^{L} \|\boldsymbol{s} - r_\ell \mathbf{1}\|_1$ for the unknown discrete-valued vector. The idea of the regularizer comes from compressed sensing [40, 41] for the reconstruction of the sparse vector and the fact that the vector $\boldsymbol{x} - r_\ell \mathbf{1}$ has some zero elements. As described in [57], the optimization problem (1.34) can be solved as linear programming. As for the theoretical analysis, the SER of the regularization-based method has been derived for the binary vector reconstruction. Some methods based on a similar idea have also been proposed for the noisy measurement case [58, 59]. However, these methods cannot utilize the knowledge of the distribution of the unknown vector.

The sum of absolute values (SOAV) optimization [60] for the reconstruction of $\boldsymbol{x}$ is given by

$$\underset{\boldsymbol{s}\in\mathbb{R}^N}{\text{minimize}} \sum_{\ell=1}^{L} q_\ell \|\boldsymbol{s} - r_\ell \mathbf{1}\|_1 \ \text{ subject to } \ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{s} \tag{1.35}$$

in the noise-free case, and

$$\underset{\boldsymbol{s}\in\mathbb{R}^N}{\text{minimize}} \left\{ \sum_{\ell=1}^{L} q_\ell \|\boldsymbol{s} - r_\ell \mathbf{1}\|_1 + \frac{\lambda}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 \right\} \tag{1.36}$$

in the noisy case, where q_ℓ (≥ 0) is a parameter and set as $q_\ell = p_\ell$ in [60]. λ (> 0) is also an parameter to control the balance between two terms in the objective function. Although the SOAV optimization is based on a similar idea as that of the regularization-based method (1.34), it includes the parameter q_ℓ in the objective function. By tuning these parameters, we can take the probability $p_1, \ldots, p_L$ into account. Since the objective function of the SOAV optimization is convex, we can obtain a sequence converging to the optimal solution by several convex optimization algorithms such as proximal splitting methods [61]. For example, an algorithm based on Beck-Teboulle proximal gradient algorithm [62] has been proposed in [26]. Some theoretical results about the SOAV optimization have been derived in [26] by using restricted isometry property (RIP) [63].

1.4 Outline of the Thesis

As described in the previous sections, we need a low-complexity algorithm for the large-scale discrete-valued vector reconstruction. Although the LMMSE method and the local

neighborhood search have low complexity, their reconstruction performance severely degrades for underdetermined problems. The message passing-based algorithms can achieve good performance with low complexity. However, it requires some assumptions on the measurement matrix and the large system limit. Since the measurement matrix has various structures in the applications described in Section 1.2, we require an algorithm which can achieve good performance without any assumptions on the measurement matrix. We thus propose novel reconstruction algorithms in this thesis by extending the conventional convex optimization-based methods, which does not require any explicit assumptions on the measurement matrix.

Another issue on the discrete-valued vector reconstruction is the performance analysis of the algorithms. For message passing-based methods, the asymptotic performance has been analyzed under some assumptions on the measurement matrix. For box-relaxation method and the regularization-based method, the theoretical results have also been provided in the large system limit. However, the theoretical aspects of the SOAV optimization has not been understood sufficiently. In this thesis, we thus analyze the asymptotic performance of the SOAV optimization and the corresponding message passing-based algorithm. The result of the analysis enables us to optimize the parameters in the objective function. Moreover, we can derive the required number of measurements for the perfect reconstruction in the noise-free case.

The remainder of this thesis is organized as follows. Figure 1.3 shows the overview of the thesis.

Chapter 2: Binary Vector Reconstruction via Iterative Convex Optimization

In Chapter 2, we consider the binary vector reconstruction as the simplest case of the discrete-valued vector reconstruction. We extend the conventional SOAV optimization to the weighted sum of absolute values (W-SOAV) optimization so that we can use the prior information of the unknown vector. Moreover, we propose an iterative approach referred to as iterative weighted sum of absolute values (IW-SOAV) to solve the W-SOAV optimization with the update of the parameters in the objective function. Simulation results show that the bit error rate (BER) performance of the proposed method is better than that of conventional schemes, especially in the large-scale overloaded MIMO signal detection and the large-scale decoding of NO-STBC.

Chapter 3: Reconstruction of Complex Discrete-Valued Vector via Convex Optimization with Sparse Regularizers

In Chapter 3, we propose a method for the reconstruction of a complex discrete-valued vector from its linear measurements. We propose a reconstruction approach of solving an optimization problem called sum of complex sparse regularizers (SCSR) optimization. The sum of sparse regularizers in the objective function can directly utilize the discrete

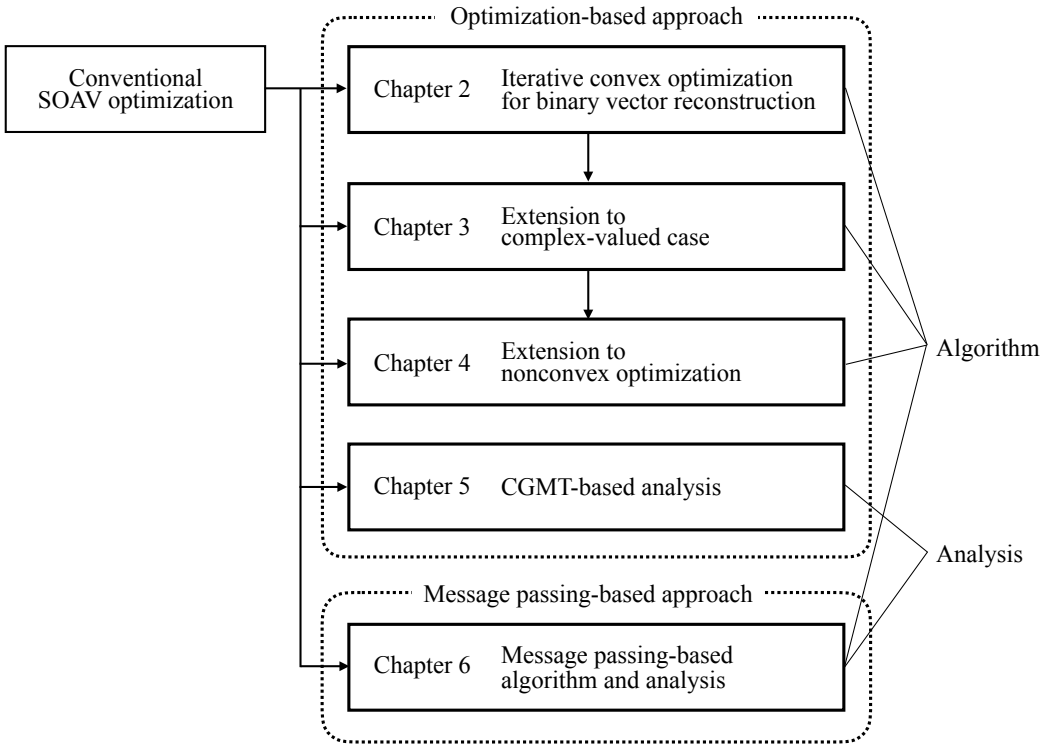


Figure 1.3: Overview of the thesis

nature of the unknown vector in the complex-valued domain. We also propose an algorithm for the SCSR optimization problem on the basis of alternating direction method of multipliers (ADMM). For the proposed convex regularizers, we analytically prove that the sequence obtained by the proposed algorithm converges to the optimal solution of the problem. To obtain better reconstruction performance, we further propose an iterative approach named iterative weighted sum of complex sparse regularizers (IW-SCSR), where we update the parameters in the objective function in each iteration by using the tentative estimate in the previous iteration. Simulation results show that IW-SCSR can reconstruct the complex discrete-valued vector from its underdetermined linear measurements and achieve good performance in the applications of overloaded MIMO signal detection and channel equalization.

Chapter 4: Discrete-Valued Vector Reconstruction by Optimization with Sum of Sparse Regularizers

In Chapter 4, we propose a possibly nonconvex optimization problem to reconstruct a discrete-valued vector from its underdetermined linear measurements. The proposed

sum of sparse regularizers (SSR) optimization uses the sum of sparse regularizers as a regularizer for the discrete-valued vector. We also propose two proximal splitting algorithms for the SSR optimization problem on the basis of ADMM and primal-dual splitting (PDS). The ADMM-based algorithm can achieve faster convergence, whereas the PDS-based algorithm does not require the computation of any inverse matrix. Moreover, we extend the ADMM-based approach for the reconstruction of complex discrete-valued vectors. Note that the proposed approach can use any sparse regularizer as long as its proximity operator can be efficiently computed. Simulation results show that the proposed algorithms with nonconvex regularizers can achieve good reconstruction performance.

Chapter 5: Asymptotic Performance Analysis of Discrete-Valued Vector Reconstruction with Sum of ℓ_1 Regularizers

In Chapter 5, we analyze the asymptotic performance of a convex optimization-based discrete-valued vector reconstruction from linear measurements. We firstly propose a box-constrained version of the conventional SOAV optimization, which uses a weighted sum of ℓ_1 regularizers as a regularizer for the discrete-valued vector. We then derive the asymptotic SER performance of the box-constrained sum of absolute values (Box-SOAV) optimization theoretically by using CGMT. We also derive the asymptotic distribution of the estimate obtained by the Box-SOAV optimization. On the basis of the asymptotic results, we can obtain the optimal parameters of the Box-SOAV optimization in terms of the asymptotic SER. Moreover, we can also optimize the quantizer to obtain the final estimate of the unknown discrete-valued vector. Simulation results show that the empirical SER performance of Box-SOAV and the conventional SOAV is very close to the theoretical result for Box-SOAV when the problem size is sufficiently large. We also show that we can obtain better SER performance by using the proposed asymptotically optimal parameters and quantizers compared to the case with some fixed parameter and a naive quantizer.

Chapter 6: Discreteness-Aware Approximate Message Passing for Discrete-Valued Vector Reconstruction

Chapter 6 considers the reconstruction of a discrete-valued random vector from possibly underdetermined linear measurements using SOAV optimization. The proposed algorithm, referred to as discreteness-aware approximate message passing (DAMP), is based on the idea of AMP, which has been originally proposed for compressed sensing. The DAMP algorithm has low computational complexity and its performance in the large system limit can be predicted analytically via state evolution framework, where we provide a condition for the exact reconstruction with DAMP in the noise-free case. From the analysis, we also propose a method to determine the parameters of the SOAV optimiza-

tion. Moreover, on the basis of the state evolution, we provide Bayes optimal DAMP, which has the minimum mean-square-error at each iteration of the algorithm. Simulation results show that the DAMP algorithms can reconstruct the discrete-valued vector from underdetermined linear measurements and the empirical performance agrees with our theoretical results in large-scale systems. When the problem size is not large enough, the SOAV optimization with the proposed parameters can achieve better performance than the DAMP algorithms for high signal-to-noise ratio.

Chapter 7: Conclusion

In Chapter 7, we provide the summary and future work of this thesis.

Chapter 2

Binary Vector Reconstruction via Iterative Convex Optimization

2.1 Introduction

In this chapter, we consider the binary vector reconstruction as the simplest example of the discrete-valued vector reconstruction. We firstly formulate the SOAV optimization [60] for the binary vector reconstruction in the noisy observation case. We then extend the SOAV optimization to the W-SOAV optimization, where the prior information on the unknown vector can be used, and propose an iterative approach, referred to as IW-SOAV. In IW-SOAV, we iterate the W-SOAV optimization and the update of the parameters in the objective function. IW-SOAV can reconstruct the unknown binary vector with low computational complexity because the W-SOAV optimization problem can be efficiently solved with proximal splitting methods [61]. Simulation results show that IW-SOAV has better BER performance than several conventional methods in signal detection and the decoding of NO-STBCs in overloaded MIMO systems.

The rest of this chapter is organized as follows. In Section 2.2, we present the proposed IW-SOAV for the binary vector reconstruction. Section 2.3 gives some simulation results to demonstrate the performance of the proposed scheme. Finally, Section 2.4 presents some conclusions.

2.2 Proposed Method

2.2.1 SOAV Optimization for Binary Vector Reconstruction

In this chapter, we consider the reconstruction of the binary vector $\boldsymbol{x} \in \{1, -1\}^N$ from its linear measurements given by (1.1). We assume that the probability distribution is uniform, i.e., $\Pr(x_n = 1) = \Pr(x_n = -1) = 1/2$. From the symmetry of the distribution,

Algorithm 2.1 Douglas-Rachford algorithm for (2.2)

1: Fix $\varepsilon \in (0, 1)$, $\gamma > 0$, and $\boldsymbol{z}_0 \in \mathbb{R}^N$
2: **for** $t = 0, 1, \ldots$ **do**
3: $\quad \boldsymbol{s}_t = \mathrm{prox}_{\gamma\phi_2}(\boldsymbol{z}_t)$
4: $\quad \theta_t \in [\varepsilon, 2 - \varepsilon]$
5: $\quad \boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \theta_t(\mathrm{prox}_{\gamma\phi_1}(2\boldsymbol{s}_t - \boldsymbol{z}_t) - \boldsymbol{s}_t))$
6: **end for**

the SOAV optimization (1.36) in this case can be written as

$$\underset{\boldsymbol{s}\in\mathbb{R}^N}{\text{minimize}} \left\{ \frac{1}{2} \|\boldsymbol{s} - \boldsymbol{1}\|_1 + \frac{1}{2} \|\boldsymbol{s} + \boldsymbol{1}\|_1 + \frac{\lambda}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 \right\}, \tag{2.1}$$

where λ (> 0) is the parameter. The solution of (2.1) can be obtained with the following theorem [61].

Theorem 2.2.1. Let $\phi_1, \phi_2 : \mathbb{R}^N \to (-\infty, \infty]$ be lower semicontinuous convex functions and $(\mathrm{ri\ dom}\ \phi_1) \cap (\mathrm{ri\ dom}\ \phi_2) \neq \emptyset$, where ri and dom denote the relative interior of the set and the domain of the function, respectively. In addition, $\phi_1(\boldsymbol{s}) + \phi_2(\boldsymbol{s}) \to \infty$ as $\|\boldsymbol{s}\|_2 \to \infty$ is assumed. A sequence $\{\boldsymbol{s}_t\}$ ($t = 0, 1, \ldots$) converging to the solution of

$$\underset{\boldsymbol{s}\in\mathbb{R}^N}{\text{minimize}} \ \{\phi_1(\boldsymbol{s}) + \phi_2(\boldsymbol{s})\} \tag{2.2}$$

can be obtained by using the following Douglas-Rachford algorithm in Algorithm 2.1.

The Douglas-Rachford algorithm is one of the proximal splitting methods [61], which can solve the optimization problem with the form of (2.2) by using the proximity operator, which can be considered as an extension of the projection onto nonempty closed convex sets for the convex function. In fact, for the indicator function $\iota_C(\boldsymbol{u})$ ($\iota_C(\boldsymbol{u}) = 0$ if $\boldsymbol{u} \in C$, and $\iota_C(\boldsymbol{u}) = \infty$ otherwise) with such a convex set C, $\mathrm{prox}_{\iota_C}(\boldsymbol{u})$ is the projection of $\boldsymbol{u}$ onto C.

In order to apply the theorem to our problem, we rewrite (2.1) as

$$\underset{\boldsymbol{s}\in\mathbb{R}^N}{\text{minimize}} \ \{f(\boldsymbol{s}) + g(\boldsymbol{s})\}, \tag{2.3}$$

where $f(\boldsymbol{s}) = \lambda \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 / 2$ and $g(\boldsymbol{s}) = \|\boldsymbol{s} - \boldsymbol{1}\|_1 / 2 + \|\boldsymbol{s} + \boldsymbol{1}\|_1 / 2$. Note that $f(\boldsymbol{s})$ and $g(\boldsymbol{s})$ are lower semicontinuous convex functions due to the continuity and the convexity of ℓ_1 and ℓ_2 norms. Moreover, we have $(\mathrm{ri\ dom}\ f) \cap (\mathrm{ri\ dom}\ g) = (\mathrm{ri}\ \mathbb{R}^N) \cap (\mathrm{ri}\ \mathbb{R}^N) = \mathbb{R}^N \neq \emptyset$ and $f(\boldsymbol{s}) + g(\boldsymbol{s}) \to \infty$ as $\|\boldsymbol{s}\|_2^2 \to \infty$. Thus, we can calculate the solution of (2.1) or (2.3) by using Algorithm 2.1 with $\phi_1(\boldsymbol{s}) = f(\boldsymbol{s})$ and $\phi_2(\boldsymbol{s}) = g(\boldsymbol{s})$. The proximity operators of $\gamma f(\boldsymbol{s})$ and $\gamma g(\boldsymbol{s})$ can be obtained as

$$\mathrm{prox}_{\gamma f}(\boldsymbol{u}) = \left(\boldsymbol{I} + \lambda\gamma \boldsymbol{H}^\mathsf{T}\boldsymbol{H}\right)^{-1} \left(\boldsymbol{u} + \lambda\gamma \boldsymbol{H}^\mathsf{T}\boldsymbol{y}\right), \tag{2.4}$$

and

$$\left[\mathrm{prox}_{\gamma g}(\boldsymbol{u})\right]_n = \begin{cases} u_n + \gamma & (u_n < -1 - \gamma) \\ -1 & (-1 - \gamma \leq u_n < -1) \\ u_n & (-1 \leq u_n \leq 1) \\ 1 & (1 \leq u_n < 1 + \gamma) \\ u_n - \gamma & (1 + \gamma \leq u_n) \end{cases} \tag{2.5}$$

respectively, where u_n indicates the nth element of $\boldsymbol{u}$.

The computational complexity of the algorithm is $O(N^3)$, which is dominated by the matrix inversion $\left(\boldsymbol{I} + \lambda\gamma \boldsymbol{H}^\top \boldsymbol{H}\right)^{-1}$ in (2.4). Note that the calculation of the inversion is required only once, and thus the corresponding computational cost does not grow with the number of iterations in the algorithm.

2.2.2 IW-SOAV

In this section, we consider to further improve the performance of the SOAV optimization by extending the SOAV optimization into the W-SOAV optimization. Moreover, we also propose the iterative approach named IW-SOAV, where we iterate the W-SOAV and the update of the parameter in the objective function.

We firstly extend the SOAV optimization in (2.1) to the W-SOAV so that we can use the prior information about $\boldsymbol{x}$ as

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{s}\in\mathbb{R}^N}{\arg\min} \left\{ \sum_{n=1}^{N} \left(q_n^+ \left| s_n - 1 \right| + q_n^- \left| s_n + 1 \right| \right) + \frac{\lambda}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{s} \right\|_2^2 \right\}, \tag{2.6}$$

where we can choose the different parameters q_n^+ and q_n^- for each n. If there is no prior information about $\boldsymbol{x}$, i.e., $q_n^+ = q_n^- = 1/2$, the optimization problem (2.6) is equivalent to (2.1). If $q_n^+ > q_n^-$, then $\underset{s_n\in\mathbb{R}}{\arg\min} \left(q_n^+ |s_n - 1| + q_n^- |s_n + 1| \right) = 1$ and hence the solution of s_n in (2.6) tends to take the value close to 1, and vice versa. Hence, if we have the prior information about the unknown vector, we can incorporate them by tuning the parameters q_n^+ and q_n^- properly. The optimization problem (2.6) can also be solved by using the Douglas-Rachford algorithm. The proximity operator of

$$\gamma g_{\mathrm{w}}(\boldsymbol{u}) \coloneqq \gamma \sum_{n=1}^{N} \left(q_n^+ |u_n - 1| + q_n^- |u_n + 1| \right) \tag{2.7}$$

Algorithm 2.2 IW-SOAV

Input: $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$
Output: $\hat{\boldsymbol{x}} \in \mathbb{R}^N$

Let $\hat{\boldsymbol{x}} = \boldsymbol{0}$.
for $k = 1$ to K_{itr} **do**
    Update q_n^+ and q_n^- from $\hat{\boldsymbol{x}}$.
    Fix $\varepsilon \in (0, 1)$, $\gamma > 0$, and $\boldsymbol{z}_0 \in \mathbb{R}^N$.
    for $t = 0$ to T_{itr} **do**
        $\boldsymbol{s}_t = \text{prox}_{\gamma g_{\text{w}}}(\boldsymbol{z}_t)$
        $\theta_t \in [\varepsilon, 2 - \varepsilon]$
        $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \theta_t(\text{prox}_{\gamma f}(2\boldsymbol{s}_t - \boldsymbol{z}_t) - \boldsymbol{s}_t)$
    end for
    $\hat{\boldsymbol{x}} = \boldsymbol{s}^{T_{\text{itr}}}$
end for
$\hat{\boldsymbol{x}} = \boldsymbol{s}^{T_{\text{itr}}}$

can be written as

$$
\begin{aligned}
&[\text{prox}_{\gamma g_{\text{w}}}(\boldsymbol{u})]_n \\
&= \begin{cases}
u_n + \gamma & (u_n < -1 - \gamma) \\
-1 & (-1 - \gamma \le u_n < -1 - d_n\gamma) \\
u_n + d_n\gamma & (-1 - d_n\gamma \le u_n < 1 - d_n\gamma) \\
1 & (1 - d_n\gamma \le u_n < 1 + \gamma) \\
u_n - \gamma & (1 + \gamma \le u_n)
\end{cases},
\end{aligned} \tag{2.8}
$$

where $d_n = q_n^+ - q_n^-$. By solving the optimization problem in (2.6) via the Douglas-Rachford algorithm with $\text{prox}_{\gamma f}$ and $\text{prox}_{\gamma g_{\text{w}}}$, a new estimate of the unknown vector $\boldsymbol{x}$ can be obtained.

To implement the idea of W-SOAV, we propose an iterative approach summarized in Algorithm 2.2, referred to as IW-SOAV. In each iteration of IW-SOAV, we use the estimate obtained in the previous iteration as the prior information. By calculating the weights q_n^+ and q_n^- from the estimate and solving the W-SOAV optimization problem, we can obtain an improved estimate of $\boldsymbol{x}$. We discuss the method of the weight update in the next subsection.

The computational complexity of IW-SOAV is the same order as that of the Douglas-Rachford algorithm for the SOAV optimization because it is dominated by the matrix inversion $\left(\boldsymbol{I} + \lambda\gamma\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}\right)^{-1}$ in (2.4).

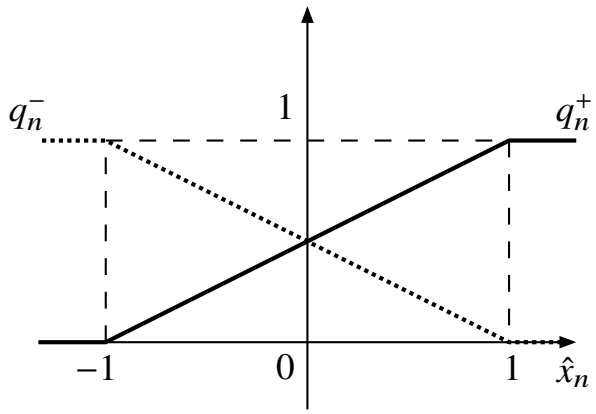


Figure 2.1: q_n^+ and q_n^- in the simple approach

2.2.3 Weight Update Rule in IW-SOAV

As a candidate for the weight update rule, we can consider the simple method given by

$$q_n^+ = \begin{cases} 0 & (\hat{x}_n < -1) \\ \dfrac{1+\hat{x}_n}{2} & (-1 \le \hat{x}_n < 1) \\ 1 & (1 \le \hat{x}_n) \end{cases} \tag{2.9}$$

and

$$q_n^- = 1 - q_n^+ = \begin{cases} 1 & (\hat{x}_n < -1) \\ \dfrac{1-\hat{x}_n}{2} & (-1 \le \hat{x}_n < 1) \\ 0 & (1 \le \hat{x}_n) \end{cases}. \tag{2.10}$$

Figure 2.1 shows q_n^+ and q_n^- as a function of $\hat{x}_n$. q_n^+ is large when $\hat{x}_n$ is large, whereas q_n^- is large when $\hat{x}_n$ is small.

Although the above approach is very simple, it does not use the previous estimate of other symbols x_i $(i \neq n)$ to obtain the weight q_n^+ and q_n^-. As a more reasonable approach, we here propose the log likelihood ratio (LLR)-based approach. We firstly consider to approximate the posterior LLR of x_n defined as

$$\Lambda_n := \log \frac{p(x_n = +1 \mid \boldsymbol{y})}{p(x_n = -1 \mid \boldsymbol{y})} \tag{2.11}$$

$$= \log \frac{p(\boldsymbol{y} \mid x_n = +1)}{p(\boldsymbol{y} \mid x_n = -1)}, \tag{2.12}$$

by using the current estimate $\hat{\boldsymbol{x}}$. To reduce the computational complexity, we firstly approximate Λ_n as

$$\Lambda_n \approx \log \frac{\prod_{m=1}^{M} p(y_m \mid x_n = +1)}{\prod_{m=1}^{M} p(y_m \mid x_n = -1)} \tag{2.13}$$

$$= \sum_{m=1}^{M} \log \frac{p(y_m \mid x_n = +1)}{p(y_m \mid x_n = -1)} \tag{2.14}$$

by assuming that the observations $y_1, \ldots, y_M$ are independent, which means $p(\boldsymbol{y} \mid x_n = +1) = \prod_{m=1}^{M} p(y_m \mid x_n = +1)$ and $p(\boldsymbol{y} \mid x_n = -1) = \prod_{m=1}^{M} p(y_m \mid x_n = -1)$. By using the similar idea to the Gaussian approximation in the BP-based detection [64], we rewrite y_m as

$$y_m = a_{m,n} x_n + \sum_{\substack{k=1 \\ k \neq n}}^{N} a_{m,k} x_k + v_m \tag{2.15}$$

$$= a_{m,n} x_n + \xi_m^n \tag{2.16}$$

where $\xi_m^n = \sum_{k=1, k\neq n}^{N} a_{m,k} x_k + v_m$. Since ξ_m^n is the sum of $N-1$ independent random variables and Gaussian noise, we can approximate it as a Gaussian random variable from the central limit theorem when N is large. We thus calculate (2.14) as

$$\sum_{m=1}^{M} \log \frac{p(y_m \mid x_n = +1)}{p(y_m \mid x_n = -1)} \approx \sum_{m=1}^{M} \frac{2a_{m,n}\left(y_m - \mu_{\xi_m^n}\right)}{\sigma^2_{\xi_m^n}}, \tag{2.17}$$

where $\mu_{\xi_m^n}$ and $\sigma^2_{\xi_m^n}$ represent the mean and the variance of ξ_m^n, respectively, which are given by

$$\mu_{\xi_m^n} = \sum_{\substack{k=1 \\ k \neq n}}^{N} a_{m,k} \mathrm{E}\left[x_k\right], \tag{2.18}$$

$$\sigma^2_{\xi_i^n} = \sum_{\substack{k=1 \\ k \neq n}}^{N} a^2_{m,k} \left(1 - \mathrm{E}\left[x_k\right]^2\right) + \sigma_{\mathrm{v}}^2. \tag{2.19}$$

Since $\mathrm{E}\left[x_k\right]$ is not available in general, we approximate $\mu_{\xi_i^n}$ and $\sigma^2_{\xi_i^n}$ using the current

estimates $\hat{x}_1, \ldots, \hat{x}_N$ as

$$\mu_{\xi_i^n} \approx \hat{\mu}_{\xi_i^n} := \sum_{\substack{k=1 \\ k \neq n}}^{N} a_{m,k} \hat{x}'_k, \tag{2.20}$$

$$\sigma^2_{\xi_m^n} \approx \hat{\sigma}^2_{\xi_m^n} := \sum_{\substack{k=1 \\ k \neq n}}^{N} a^2_{m,k} \left(1 - (\hat{x}'_k)^2\right) + \sigma_{\mathrm{v}}^2, \tag{2.21}$$

where

$$\hat{x}'_n = \begin{cases} -1 & (\hat{x}_n < -1) \\ \hat{x}_n & (-1 \leq \hat{x}_n < 1) \\ 1 & (1 \leq \hat{x}_n) \end{cases} \tag{2.22}$$

is bounded in $[-1, 1]$ so that $1 - (\hat{x}'_k)^2$ in (2.21) is not negative. From (2.14), (2.17), (2.20) and (2.21), the posterior LLR of x_n can be approximated as

$$\Lambda_n \approx \hat{\Lambda}_n := \sum_{m=1}^{M} \frac{2a_{m,n}\left(y_m - \sum_{\substack{k=1 \\ k \neq n}}^{N} a_{m,k} \hat{x}'_k\right)}{\sum_{\substack{k=1 \\ k \neq n}}^{N} a^2_{m,k} \left(1 - (\hat{x}'_k)^2\right) + \sigma_{\mathrm{v}}^2}. \tag{2.23}$$

From the approximated posterior LLR $\hat{\Lambda}_n$, we use the approximation of the posterior probabilities $\Pr(x_n = +1 \mid \boldsymbol{y})$ and $\Pr(x_n = -1 \mid \boldsymbol{y})$ as q_n^+ and q_n^-, respectively, and update

$$q_n^+ := \frac{e^{\hat{\Lambda}_n}}{1 + e^{\hat{\Lambda}_n}}, \quad q_n^- := \frac{1}{1 + e^{\hat{\Lambda}_n}}. \tag{2.24}$$

It should be noted that the LLR-based approach can be combined with the channel decoder as shown in Sec. 2.3.2.

Since the computational complexity of (2.23) is $O(MN)$, the complexity for the direct calculation of $\hat{\Lambda}_1, \ldots, \hat{\Lambda}_N$ will be $O(MN^2)$. However, we can reduce the complexity to $O(MN)$ by calculating and storing

$$\hat{\mu}_m := \sum_{k=1}^{N} a_{m,k} \hat{x}'_k, \tag{2.25}$$

$$\hat{\sigma}^2_m := \sum_{k=1}^{N} a^2_{m,k} \left(1 - (\hat{x}'_k)^2\right) + \sigma_{\mathrm{v}}^2, \tag{2.26}$$

in advance. Since (2.25) and (2.26) can be calculated with the complexity of $O(N)$, we can obtain all of $\hat{\mu}_m$ and $\hat{\sigma}_m^2$ $(m = 1, \ldots, M)$ with $O(MN)$. By using $\hat{\mu}_m$ and $\hat{\sigma}_m$, (2.20) and (2.21) are rewritten as

$$\hat{\mu}_{\xi_m^n} = \hat{\mu}_m - a_{m,n}\hat{x}'_n, \tag{2.27}$$

$$\hat{\sigma}^2_{\xi_m^n} = \hat{\sigma}_m^2 - a_{m,n}^2\left(1 - (\hat{x}'_n)^2\right), \tag{2.28}$$

which can be obtained with $O(1)$. With (2.25)–(2.28), (2.23) can be rewritten as

$$\hat{\Lambda}_n = \sum_{m=1}^{M} \frac{2a_{m,n}\left\{y_m - \left(\hat{\mu}_m - a_{m,n}\hat{x}'_n\right)\right\}}{\hat{\sigma}_m^2 - a_{m,n}^2\left(1 - (\hat{x}'_n)^2\right)} \tag{2.29}$$

and hence the complexity needed for the calculation of each $\hat{\Lambda}_n$ is reduced to just $O(M)$. As a result, we can obtain $\hat{\Lambda}_1, \ldots, \hat{\Lambda}_N$ with the complexity of $O(MN)$, while the complexity of the Douglas-Rachford algorithm for the SOAV optimization is $O(N^3)$.

2.2.4 Extension to Non-Binary Vector

Although we have focused on the binary vector reconstruction in this chapter, we can extend the proposed approach to the reconstruction of non-binary vectors. The W-SOAV optimization in (2.6) can be extended for the reconstruction of $\boldsymbol{x} \in \{r_1, \ldots, r_L\}^N$ as

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{s}\in\mathbb{R}^N} \left\{ \left(\sum_{n=1}^{N}\sum_{\ell=1}^{L} q_{\ell,n}\left|s_n - r_\ell\right| \right) + \frac{\lambda}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\right\|_2^2 \right\}, \tag{2.30}$$

where $q_{\ell,n}$ (≥ 0) is the parameter. We have proposed a method for the error recovery of MIMO signal detection with QPSK, where we reconstruct the discrete-valued vector in $\{-2, 0, 2\}^N$ [65]. In [65], the parameter $q_{\ell,n}$ is determined on the basis of the relaxed maximum a posteriori (MAP) estimation with approximated LLRs.

2.3 Simulation Results

In this section, we show the performance of LW-SOAV in two applications in communication systems via computer simulations. As the weight update rule, we use the proposed LLR-based approach.

2.3.1 Overloaded MIMO Signal Detection

As discussed in Section 1.2, massive overloaded MIMO signal detection can be regarded as the discrete-valued vector reconstruction. When we use QPSK with $(c_1, c_2, c_3, c_4) =$

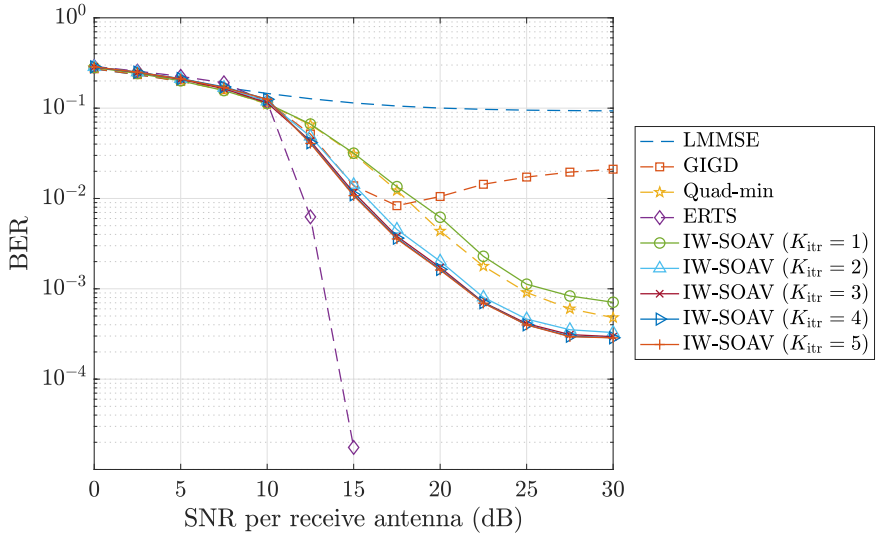


Figure 2.2: BER performance for MIMO with $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (50, 32)$

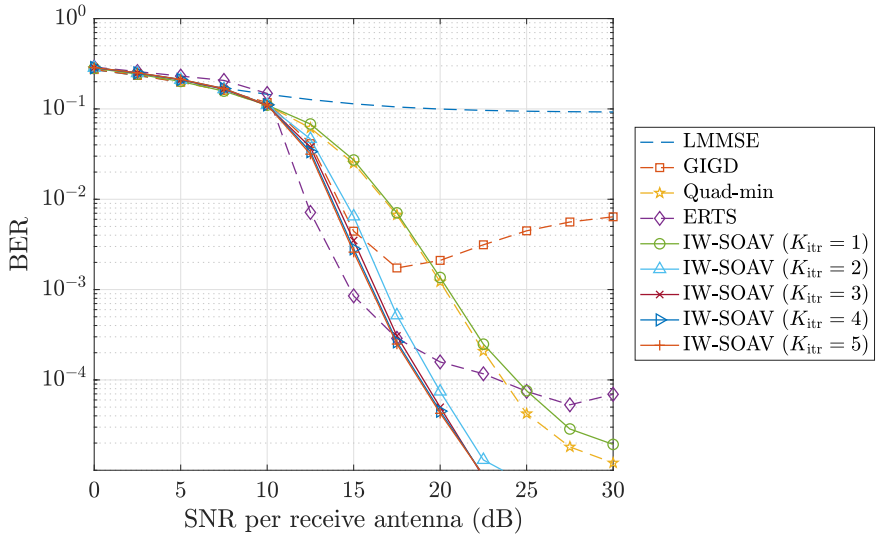


Figure 2.3: BER performance for MIMO with $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (100, 64)$

$(1+j, -1+j, -1-j, 1-j)$ as the modulation method, the signal detection problem results in the binary vector reconstruction in the real-valued domain after the transformation to the model (1.3). We here present several simulation results for massive overloaded MIMO signal detection via IW-SOAV.

Figures 2.2 and 2.3 shows the BER performance versus signal-to-noise ratio (SNR) for overloaded MIMO systems with $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (50, 32)$ and $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (100, 64)$, respectively. In the figures, we assume flat Rayleigh fading channels and set $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}}_{\mathrm{i.i.d.}}$,

Table 2.1: The value of λ in (2.6)

SNR per receive antenna (dB)	0–10	12.5–20	22.5	25–30
λ	0.01	0.1	0.3	1

which is composed of i.i.d. complex Gaussian random variables with zero mean and unit variance. We denote the LMMSE detection by "MMSE", the BP-based detection named graph-based iterative Gaussian detector (GIGD) [64] by "GIGD," the detection with a quadratic programming [58] by "Quad-min", and the massive overloaded MIMO signal detection proposed in [7] by "ERTS". The parameters of ERTS are the same as those in [7], e.g., the maximum number of RTSs is $N_{\mathrm{RTS}} = 500$ and the maximum number of iterations in RTS is $N_{\mathrm{itr}} = 300$. "IW-SOAV" denotes the convex optimization-based IW-SOAV in Algorithm 2.2. The parameters of the Douglas-Rachford algorithm are set as $\boldsymbol{z}_0 = \mathbf{0}$, $\varepsilon = 0.1, \gamma = 1$, and $\theta_t = 1.9$ $(t = 0, 1, \ldots, T_{\mathrm{itr}})$, which give fast convergence. The number of iterations in the Douglas-Rachford algorithm is fixed to $T_{\mathrm{itr}} = 50$, which is sufficiently large for the convergence of the algorithm. The parameter λ in (2.6) is selected as shown in TABLE 2.1, which is determined from simulation results. In the figures, T_{itr} is the number of iterative W-SOAV optimizations in IW-SOAV. In Fig. 2.2, where $N_{\mathrm{t}} = 50$, the performance of IW-SOAV is inferior to that of ERTS. In Fig. 2.3 with $N_{\mathrm{t}} = 100$, however, the performance of ERTS has degraded and IW-SOAV has better performance in high SNR. The reason for the performance degradation of ERTS is that, if the number of transmit antennas is large, RTS often fails to find the true transmitted signal vector due to the huge number of candidates of the transmitted vector. Although we may get better performance with ERTS by increasing the number of RTSs, the computational complexity could be prohibitive to achieve comparable performance as IW-SOAV. Specifically, the computational complexity of ERTS is given by $O(N_{\mathrm{t}}^3) + O(N_{\mathrm{RTS}} N_{\mathrm{t}}^2)$ in the worst case, and since the number of all candidates of the transmit signal vector increases exponentially with the number of transmit antennas, the required N_{RTS} to keep good performance will increase more rapidly than N_{t}. On the other hand, the computational complexity of IW-SOAV is $O(N_{\mathrm{t}}^3)$.

Figure 2.4 shows the BER performance versus the number of receive antennas N_{r} for $N_{\mathrm{t}} = 150$ and the SNR per receive antenna of 20 dB. We can observe that IW-SOAV with $L = 5$ requires less antennas than other schemes to achieve a certain BER performance. For BER $= 10^{-4}$, IW-SOAV can reduce more than ten receive antennas compared to ERTS.

In Fig. 2.5, we also show the BER performance for spatially correlated MIMO channels with $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (100, 64)$. We assume a linear array with equally spaced antennas in both the receiver and the transmitter, and set to $d_{\mathrm{r}} = d_{\mathrm{t}} = 0.5\lambda_{\mathrm{w}}$ in the simulations, where d_{r} and d_{t} are the antenna spacing at the receiver and the transmitter, respectively, and λ_{w} is the wavelength. From Fig. 2.5, we can see that the proposed scheme can achieve better performance compared to the conventional schemes even in

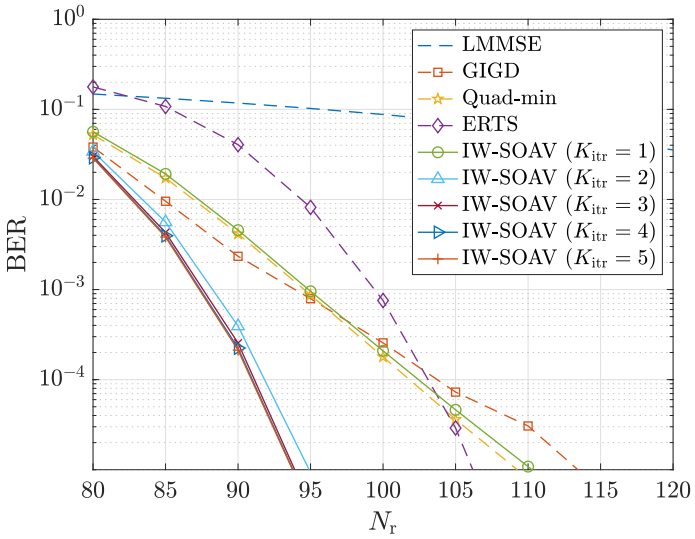


Figure 2.4: BER performance versus N_r for MIMO with $N_t = 150$ and the SNR per receive antenna of 20 dB

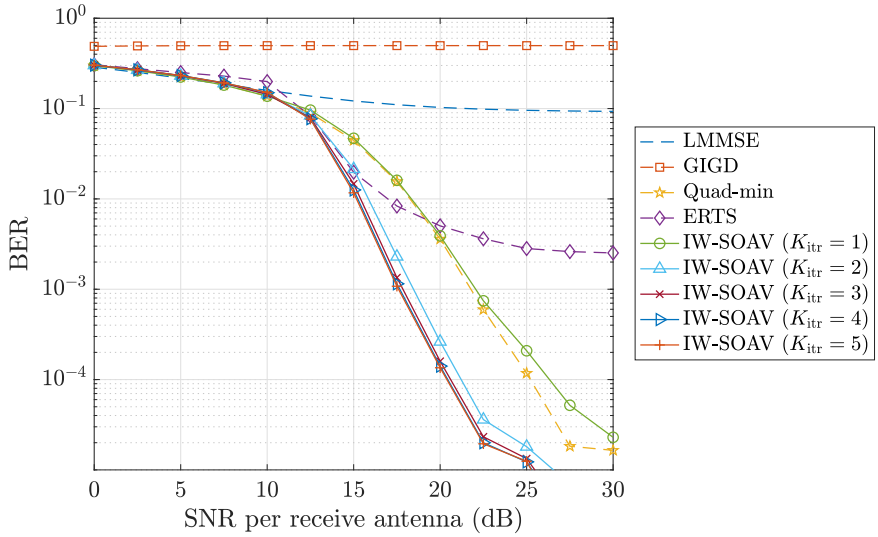


Figure 2.5: BER performance for spatially correlated MIMO with $(N_t, N_r) = (100, 64)$

the spatially correlated MIMO channels, while the performance of GIGD and ERTS is degraded significantly.

To compare the computational complexity, we evaluate the average computation time to detect a transmitted symbol vector versus N_t and the corresponding BER performance for the fixed ratio $N_r/N_t = 2/3$ and the SNR per receive antenna of 17.5 dB in Figs. 2.6 and 2.7, respectively. The simulation is conducted by using a computer with 2 GHz

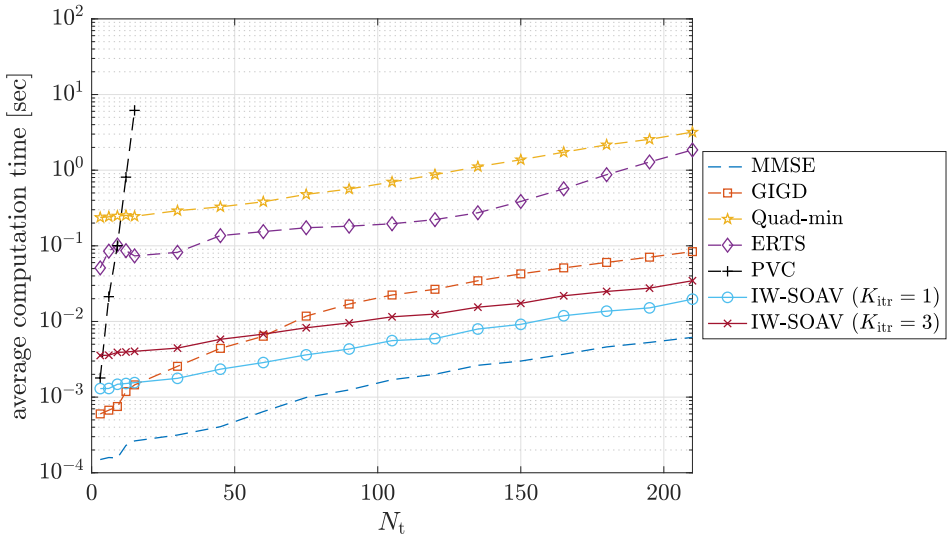


Figure 2.6: Average computation time versus N_t for uncoded MIMO with $N_r/N_t = 2/3$ and the SNR per receive antenna of 17.5 dB

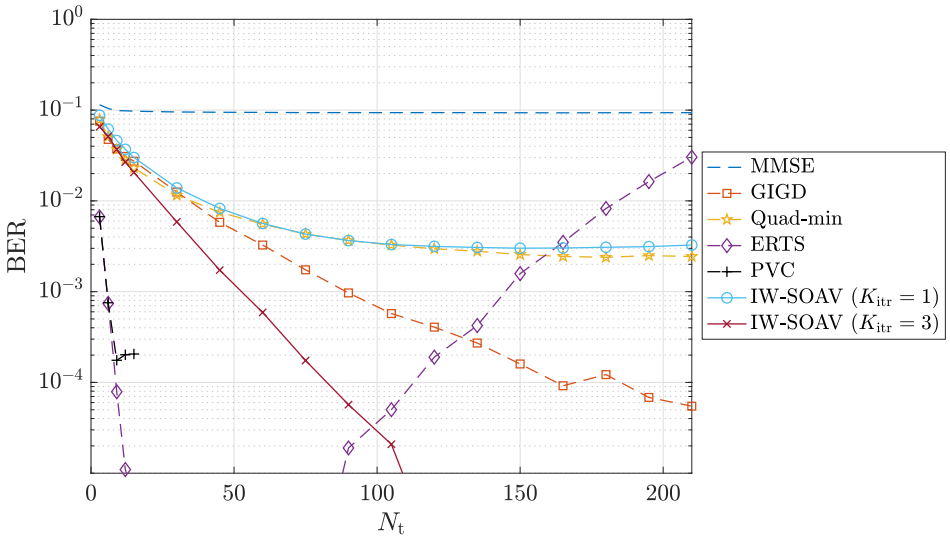


Figure 2.7: BER performance versus N_t for uncoded MIMO with $N_r/N_t = 2/3$ and the SNR per receive antenna of 17.5 dB

Intel Core i7-3667U and 8 GB memory. The channel matrix is composed of i.i.d. Gaussian variables. In the figures, "PVC" represents the signal detection scheme called pre-voting cancellation (PVC) [6], which is intended for small-scale overloaded MIMO systems. Although PVC can achieve a comparable BER performance to ML detection for small-scale MIMO systems, its average computation time rapidly increases along

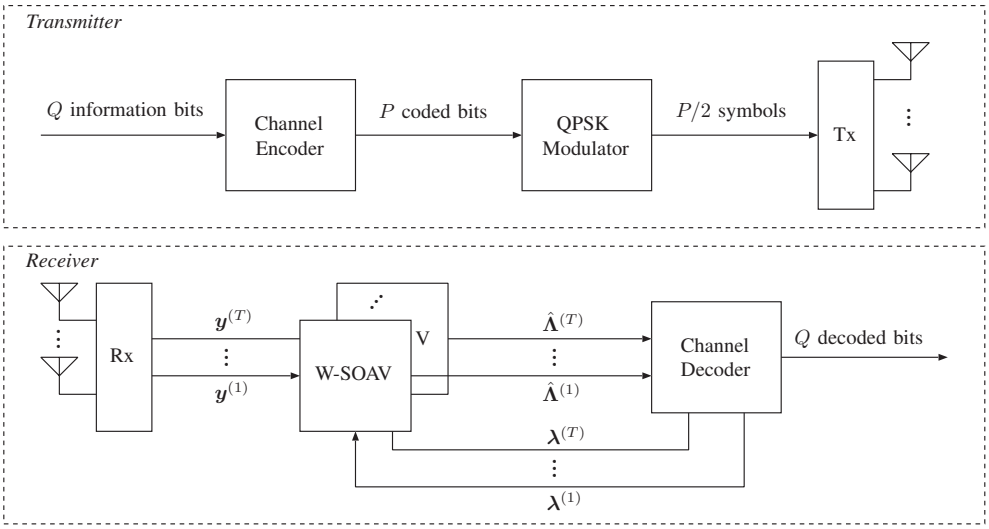


Figure 2.8: Model of coded MIMO systems

with N_t. In Fig. 2.7, the BER performance of ERTS severely degrades for large N_t. This is because the maximum number of RTSs is limited as $N_{RTS} = 500$ to avoid the prohibitive computational complexity, while the number of candidates of the transmitted signal vector exponentially increases along with N_t. Compared to the conventional detection schemes, the proposed IW-SOAV can achieve better BER performance with lower complexity in large-scale overloaded MIMO systems.

2.3.2 Signal Detection in LDPC-coded Overloaded MIMO Systems

Next, we evaluate the performance of the proposed IW-SOAV in MIMO systems with low density parity check (LDPC) codes [66, 67]. Since the proposed W-SOAV optimization can use the posterior LLRs of transmitted symbols as the prior information, we can integrate it with soft channel decoding schemes, e.g., LDPC codes or turbo codes. We thus consider a joint detection and decoding scheme using the W-SOAV optimization for coded massive overloaded MIMO systems.

Figure 2.8 shows the system model of the coded MIMO with N_t transmit antennas and N_r ($< N_t$) receive antennas. In the transmitter, Q information bits are encoded into P coded bits by a channel encoder with the code rate $R = Q/P$. For simplicity, P is assumed to be a multiple of $2N_t$. P coded bits are then modulated into $P/2$ QPSK symbols and sent from N_t transmit antennas over $T = P/2N_t$ symbols time.

The received signal vector at time $t \in \{1, \ldots, T\}$ is given by

$$\tilde{\boldsymbol{y}}^{(t)} = \tilde{\boldsymbol{A}}^{(t)}\tilde{\boldsymbol{x}}^{(t)} + \tilde{\boldsymbol{v}}^{(t)}, \tag{2.31}$$

where $\tilde{\boldsymbol{A}}^{(t)}, \tilde{\boldsymbol{x}}^{(t)}$, and $\tilde{\boldsymbol{v}}^{(t)}$ are the channel matrix, the transmitted signal vector, and the noise vector at time t, respectively. We can convert (2.31) into the real signal model

$$\boldsymbol{y}^{(t)} = \boldsymbol{A}^{(t)}\boldsymbol{x}^{(t)} + \boldsymbol{v}^{(t)}. \tag{2.32}$$

In the proposed detection and decoding, we iteratively perform the detection with the W-SOAV optimization and the channel decoding to update LLRs of transmitted symbols. The detector obtains the estimate $\hat{\boldsymbol{x}}^{(t)}$ of $\boldsymbol{x}^{(t)}$ with the W-SOAV optimization using the information from the channel decoder except for the first iteration. Specifically, by using the posterior LLR obtained at the output of the channel decoder as

$$\lambda_n^{(t)} = \log \frac{p\left(x_n^{(t)} = +1 \mid \boldsymbol{y}^{(t)}\right)}{p\left(x_n^{(t)} = -1 \mid \boldsymbol{y}^{(t)}\right)}, \tag{2.33}$$

the weight parameters of the W-SOAV optimization are given by

$$q_n^{+(t)} = \frac{e^{\lambda_n^{(t)}}}{1 + e^{\lambda_n^{(t)}}}, \; q_n^{-(t)} = \frac{1}{1 + e^{\lambda_n^{(t)}}} \tag{2.34}$$

as in (2.24). After the detection via the W-SOAV optimization with the above $q_n^{+(t)}$ and $q_n^{-(t)}$, we calculate the posterior LLRs $\hat{\boldsymbol{\Lambda}}^{(t)} = \left[\hat{\Lambda}_1^{(t)} \cdots \hat{\Lambda}_{2N_\mathrm{t}}^{(t)}\right]^\mathsf{T}$, where $\hat{\Lambda}_n^{(t)}$ is given by (2.29). Using all LLRs $\hat{\boldsymbol{\Lambda}}^{(1)}, \ldots, \hat{\boldsymbol{\Lambda}}^{(T)}$ as the input, the decoder performs the soft channel decoding and outputs new posterior LLRs $\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(T)}$ to the MIMO detector, where $\boldsymbol{\lambda}^{(t)} = \left[\lambda_1^{(t)} \cdots \lambda_{2N_\mathrm{t}}^{(t)}\right]^\mathsf{T}$. After a certain number of the iterations of the detection and decoding, the decoder outputs the decoded bits as the final estimate of the transmitted information bits.

Figures 2.9 and 2.10 show the BER performance of the proposed signal detection and decoding for LDPC coded MIMO with $(N_\mathrm{t}, N_\mathrm{r}) = (100, 64)$. The parameters of the algorithm are set as $T_\mathrm{itr} = 30$ and $\lambda = 0.01$. The code rate is $R = 1/2$, and the column and row weights of the parity check matrix are three and six, respectively. In the figures, the code length are $N_\mathrm{c} = 4000$ and 8000, respectively. We represent the proposed joint detection and decoding by "Joint det./dec.", where K_max indicates the maximum number of iterative W-SOAV optimizations. Even before the K_maxth iteration, the LDPC decoder outputs the final estimate of the information bits if the decoded bits satisfy all parity check constraints. From the figures, we can see that, as the iteration proceeds, the performance of the joint detection and decoding is considerably improved via LLR update between the W-SOAV optimization and the LDPC decoding. For comparison, we also plot the performance of the independent detection and LDPC decoding ("Independent det./dec."), where LW-SOAV with $K_\mathrm{itr} = 5$ is used as the detection scheme. Moreover,

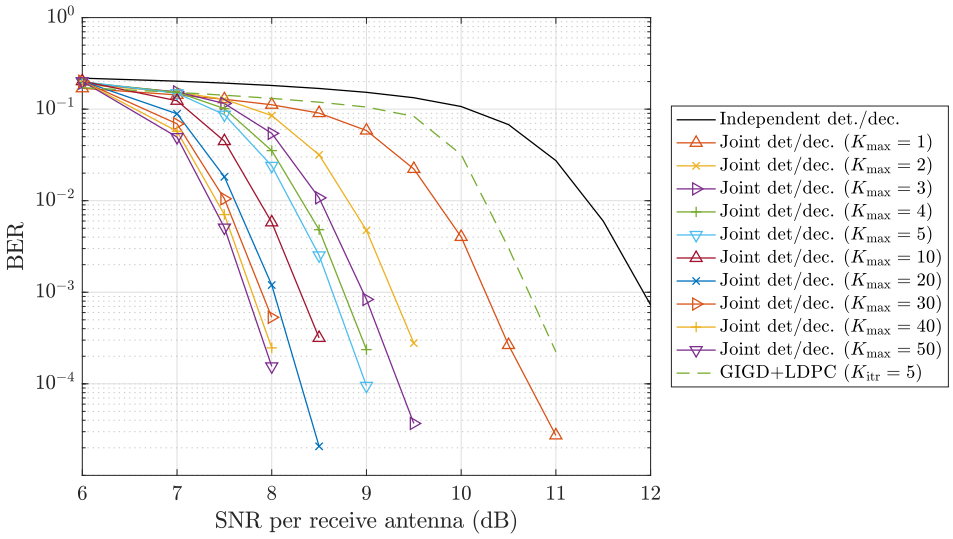


Figure 2.9: BER performance for LDPC coded MIMO with $(N_t, N_r) = (100, 64)$, $R = 1/2$ and $N = 4000$

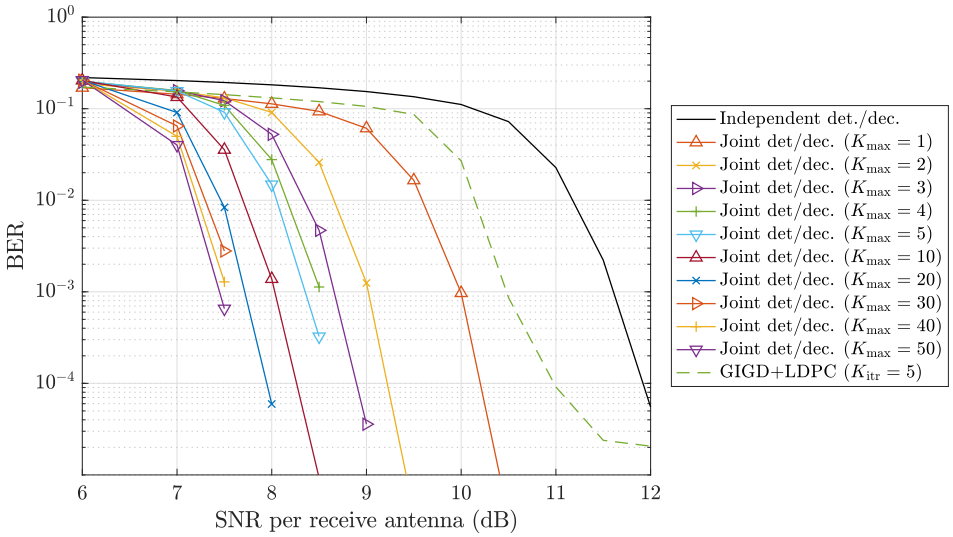


Figure 2.10: BER performance for LDPC coded MIMO with $(N_t, N_r) = (100, 64)$, $R = 1/2$ and $N = 8000$

"GIGD+LDPC" shows the performance of the joint detection and decoding with GIGD and LDPC decoding, which are integrated in the same manner as in Fig. 2.8. The number of outer iterations between the detector and the decoder is set to 5. We can see that the proposed joint detection and decoding achieves much better performance than the scheme with GIGD and the independent approach. Each element of the estimate

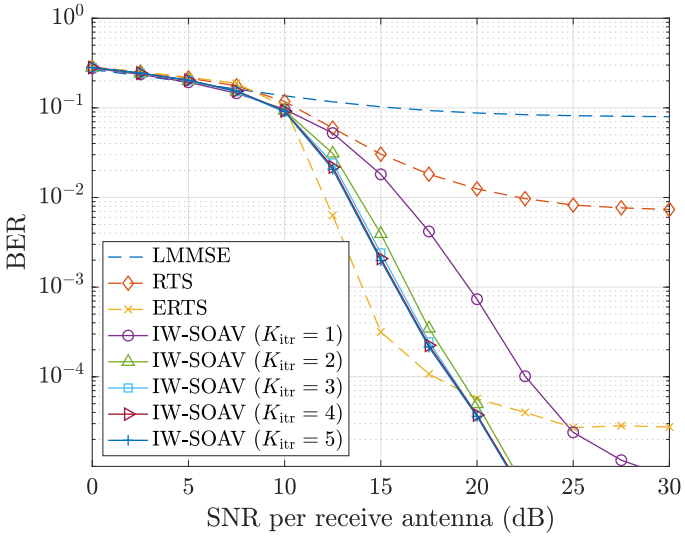


Figure 2.11: BER performance in i.i.d. channels with NO-STBC ($N_{\mathrm{t}} = 9, N_{\mathrm{r}} = 6$)

$\hat{\boldsymbol{x}}$ obtained by IW-SOAV with $K_{\mathrm{itr}} = 5$ is almost hard decision, i.e., close to 1 or -1, and hence LDPC decoding in the independent approach has poor performance even compared to the case with $K_{\mathrm{itr}} = 1$. The figures also show that the performance is improved as the code length increases.

2.3.3 Decoding of NO-STBC

Next, we show some numerical results for the decoding of NO-STBC discussed in Section 1.2. We use QPSK and NO-STBC given by (1.17) with $\delta = e^{\sqrt{5}j}$ and $\rho = e^{j}$. The parameters in IW-SOAV are $\varepsilon = 0.1$, $\gamma = 1$, $\boldsymbol{z}_0 = \boldsymbol{0}$, $T_{\mathrm{itr}} = 50$, and $\theta_t = 1.9$ $(t = 0, \ldots, T_{\mathrm{itr}})$.

Figures 2.11 and 2.12 show the BER performance for overloaded NO-STBCs. In the figures, we assume i.i.d. channels as $\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{H}}_{\mathrm{i.i.d.}}$, where the elements of $\tilde{\boldsymbol{H}}_{\mathrm{i.i.d.}}$ are i.i.d. circular complex Gaussian variables with zero mean and unit variance. We denote the LMMSE decoding by "LMMSE", the RTS-based decoding [22] by "RTS", and the proposed IW-SOAV scheme by "Proposed". We also plot the performance of ERTS [7], which has been proposed for signal detection in uncoded overloaded MIMO systems with tens of antennas. The parameters of RTS and ERTS are the same as those in [22] and [7], respectively. For the proposed IW-SOAV, the parameter λ is determined as in TABLE 2.1. Figure 2.11 shows the performance for more number of antennas $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (9, 6)$, where the size of the measurement matrix $\boldsymbol{A}$ in the resultant real-valued model is 108×162. In this case, we estimate the transmitted symbol vector in $\{1, -1\}^{162}$, which is equivalent to the signal detection for uncoded massive overloaded MIMO with

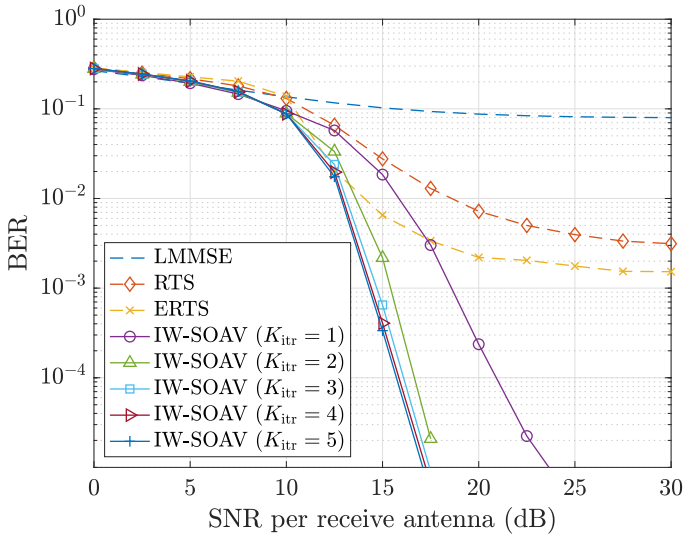


Figure 2.12: BER performance in i.i.d. channels with NO-STBC ($N_{\text{t}} = 12, N_{\text{r}} = 8$)

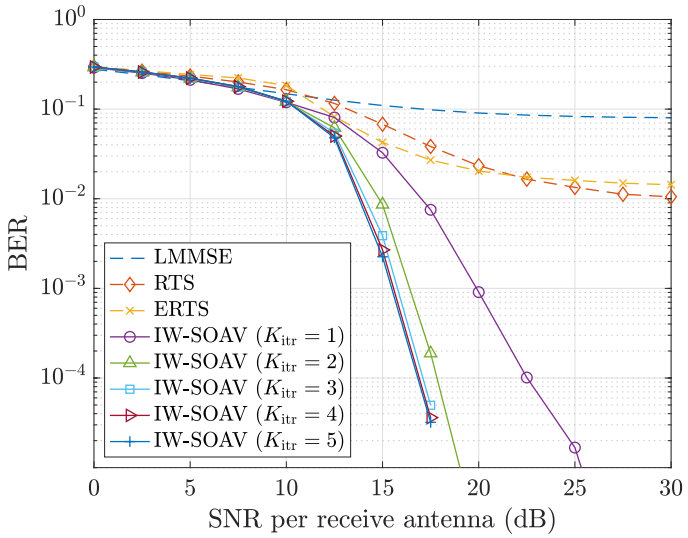


Figure 2.13: BER performance in spatially correlated channels with NO-STBC ($N_{\text{t}} = 12, N_{\text{r}} = 8$)

$(N_{\text{t}}, N_{\text{r}}) = (81, 54)$. Although ERTS achieves better performance than the proposed IW-SOAV for SNRs around 15 dB, its complexity is much larger than the proposed IW-SOAV with $K_{\text{itr}} = 3$ according to Fig. 2.6. In Fig. 2.12, where $(N_{\text{t}}, N_{\text{r}}) = (12, 8)$ with $\boldsymbol{A}$ of size 192×288, the proposed IW-SOAV outperforms the conventional schemes for all SNRs.

Figure 2.13 shows the BER performance in spatially correlated channels. We assume

$(N_t, N_r) = (12, 8)$ and a linear array with equally spaced antennas at the receiver and transmitter. We denote the antenna spacing at the receiver and the transmitter by d_r and d_t, respectively, and set $d_r = d_t = 0.5\lambda_w$ in the simulations, where λ_w is the wavelength. Figure 2.13 shows that the performance of the proposed IW-SOAV in spatially correlated channels is comparable to that in i.i.d. channels, while the BER of ERTS significantly degrades.

2.4 Conclusion

In this chapter, we have proposed the binary vector reconstruction method named IW-SOAV, which iteratively solves the convex W-SOAV optimization problem with updating weights in the objective function. The W-SOAV optimization can be efficiently solved with the proximal splitting methods. A similar approach can be applied to the reconstruction of non-binary vectors as in [65]. Simulation results show that IW-SOAV can achieve better performance than conventional methods in large-scale overloaded MIMO signal detection and the decoding of NO-STBCs.

Chapter 3

Reconstruction of Complex Discrete-Valued Vector via Convex Optimization with Sparse Regularizers

3.1 Introduction

As shown in Section 1.2, we need to perform the discrete-valued vector reconstruction in the complex-valued domain in many applications of communication systems. Several message passing-based methods [42, 50, 53] can be used for complex-valued vectors and its asymptotic performance in the large system limit can be theoretically predicted. For arbitrary measurement matrices, however, the performance has not been verified theoretically. Moreover, the assumption of large system limit remains in the derivation and the analysis, and hence the performance may degrade for finite-scale problems. On the other hand, as described in Section 1.3, several convex optimization-based methods have been proposed for the discrete-valued vector reconstruction [54, 57, 60]. However, they reconstruct the discrete-valued vector in the real-valued domain and cannot be directly used for the reconstruction of complex discrete-valued vectors in general. When the real part and the imaginary part of the unknown vector are independent each other, we can use the reconstruction methods by converting the original model in the complex-valued domain into the equivalent model in the real-valued domain as shown in (1.3). When they are not independent, however, this approach is not appropriate because we cannot take advantage of the dependency between them. In such cases, we should directly use the discrete nature of the unknown vector in the complex-valued domain.

In this chapter, we extend the SOAV optimization for the reconstruction of complex discrete-valued vectors. This extension enables us to directly reconstruct the complex discrete-valued vector even when the real part and the imaginary part are not independent. We provide an optimization algorithm for the proposed SCSR optimization on the

basis of ADMM [61, 68–71]. To obtain better reconstruction performance, we further extend the SCSR optimization to the weighted sum of complex sparse regularizers (W-SCSR) optimization and propose IW-SCSR, which iterates the W-SCSR optimization with updating parameters in the objective function in each iteration. We also discuss the selection of the parameter in the W-SCSR optimization and the computational complexity reduction scheme for the proposed algorithm. Moreover, we prove that the sequence obtained by the proposed algorithm converges to the optimal solution of the W-SCSR optimization problem without any explicit assumptions on the measurement matrix. Simulation results show that the proposed IW-SCSR can achieve better performance than AMP- and EP-based algorithms for overloaded MIMO signal detection with around tens of antennas. For sparse discrete-valued vector, IW-SCSR outperforms the ℓ_1 optimization, which uses only the sparsity of the unknown vector. The proposed IW-SCSR also achieves good performance for channel equalization in the single carrier block transmission using cyclic prefix, where the measurement matrix becomes a block circulant matrix. These results suggest that the proposed IW-SCSR has wider range of applicability than some existing message passing-based methods, especially for communications applications.

The remainder of this chapter is organized as follows. We present the proposed SCSR optimization and IW-SCSR in Section 3.2. In Section 3.3, we show some simulation results to demonstrate the validity of the proposed approach. Section 3.4 gives some conclusions.

3.2 Proposed Method

In this section, we present the SCSR optimization and the proposed algorithm based on ADMM. We also propose IW-SCSR and discuss the convergence of the proposed algorithm for the W-SCSR optimization.

3.2.1 SCSR Optimization

A straightforward approach to reconstruct the discrete-valued vector $\tilde{\boldsymbol{x}}$ in (1.2) is the ML method in (1.32) under the additive Gaussian noise. The problem (1.32) is a combinatorial optimization problem and hence the required computational complexity can be prohibitive when the problem size (N, M) is large. We thus require a low-complexity method for the large-scale discrete-valued vector reconstruction.

We extend the SOAV optimization [60] in (1.36), which reconstructs the discrete-valued vector in the real-valued domain, to the reconstruction of the complex discrete-

valued vector. The proposed SCSR optimization is given by

$$\underset{\boldsymbol{s}\in\mathbb{C}^N}{\text{minimize}} \left\{ \sum_{\ell=1}^{L} q_\ell \tilde{g}_\ell\left(\boldsymbol{s} - c_\ell \mathbf{1}\right) + \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\boldsymbol{s} \right\|_2^2 \right\}, \tag{3.1}$$

where λ and $q_\ell \geq 0$ $(\ell = 1, \ldots, L)$ are the parameters. The function $\tilde{g}_\ell : \mathbb{C}^N \to \mathbb{R}$ is a sparse regularizer and thus the first term $\sum_{\ell=1}^{L} q_\ell \tilde{g}_\ell\left(\boldsymbol{s} - c_\ell \mathbf{1}\right)$ can be considered as a regularizer for $\boldsymbol{x} \in C^N$, which uses the fact that $\boldsymbol{x} - c_\ell \mathbf{1}$ has some zero elements. When $\tilde{g}_1(\cdot), \ldots, \tilde{g}_L(\cdot)$ are all convex, the SCSR optimization can be regarded as a convex relaxation of the ML method in (1.32).

The SCSR optimization (3.1) is an optimization problem in the complex-valued domain $\mathbb{C}^N$. As described in Section 1.1, the conventional SOAV optimization in the real-valued domain might be inappropriate for the complex discrete-valued vectors in general. On the other hand, the SCSR optimization problem can directly consider the reconstruction in the complex-valued domain.

3.2.2 Choice of Sparse Regularizers

In this paper, we consider ℓ_1 regularization-based two convex sparse regularizers $h_\star^{(1)}(\cdot)$ and $h_{\star\star}^{(1)}(\cdot)$ given by

$$h_\star^{(1)}(\boldsymbol{u}) = \|\boldsymbol{u}\|_1 \tag{3.2}$$

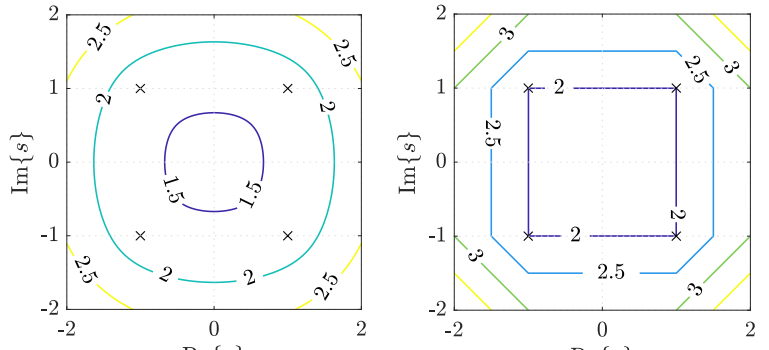
$$= \sum_{n=1}^{N} \sqrt{\mathrm{Re}\{u_n\}^2 + \mathrm{Im}\{u_n\}^2}, \tag{3.3}$$

$$h_{\star\star}^{(1)}(\boldsymbol{u}) = \|\mathrm{Re}\{\boldsymbol{u}\}\|_1 + \|\mathrm{Im}\{\boldsymbol{u}\}\|_1 \tag{3.4}$$

$$= \sum_{n=1}^{N} \left(|\mathrm{Re}\{u_n\}| + |\mathrm{Im}\{u_n\}|\right) \tag{3.5}$$

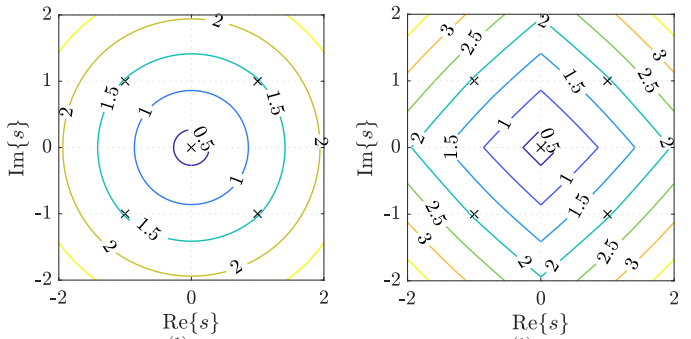
as the candidates of $\tilde{g}_\ell(\cdot)$. The first regularizer $h_\star^{(1)}(\cdot)$ is based on the modulus for complex numbers, whereas $h_{\star\star}^{(1)}(\cdot)$ handles the real part and the imaginary part separately. When the real part and the imaginary part are independent on C, the SCSR optimization with $h_{\star\star}^{(1)}(\cdot)$ is equivalent to the corresponding SOAV optimization in the real-valued domain for (1.3).

We need to choose the regularizers $h_\star^{(1)}(\cdot)$ and $h_{\star\star}^{(1)}(\cdot)$ appropriately for C. For example, in Fig. 3.1, we show the contour plot of $\sum_{\ell=1}^{L} q_\ell \tilde{g}_\ell\left(s - c_\ell\right)$ in the SCSR optimization (3.1) for $(c_1, c_2, c_3, c_4) = (1 + j, -1 + j, -1 - j, 1 - j)$ and $(q_1, q_2, q_3, q_4) = (0.25, 0.25, 0.25, 0.25)$. Figs. 3.1(a) and 3.1(b) show the contours for $\tilde{g}_\ell(\cdot) = h_\star^{(1)}(\cdot)$ and $\tilde{g}_\ell(\cdot) = h_{\star\star}^{(1)}(\cdot)$ $(\ell = 1, \ldots, 4)$, respectively. We can see that the contours are quite different.



(a) $\tilde{g}_\ell(\cdot) = h^{(1)}_\star(\cdot)$ $(\ell = 1, \ldots, 4)$ (b) $\tilde{g}_\ell(\cdot) = h^{(1)}_{\star\star}(\cdot)$ $(\ell = 1, \ldots, 4)$

Figure 3.1: Contour plot of the function $\sum_{\ell=1}^{4} q_\ell \tilde{g}_\ell (s - c_\ell)$: $(c_1, c_2, c_3, c_4) = (1 + j, -1 + j, -1 - j, 1 - j)$ and $(q_1, q_2, q_3, q_4) = (0.25, 0.25, 0.25, 0.25)$. The crosses indicate c_ℓ.



(a) $\tilde{g}_1(\cdot) = h^{(1)}_\star(\cdot)$ and $\tilde{g}_\ell(\cdot) = h^{(1)}_{\star\star}(\cdot)$ $(\ell = 2, \ldots, 5)$ (b) $\tilde{g}_\ell(\cdot) = h^{(1)}_{\star\star}(\cdot)$ $(\ell = 1, \ldots, 5)$

Figure 3.2: Contour plot of the function $\sum_{\ell=1}^{5} q_\ell \tilde{g}_\ell (s - c_\ell)$: $(c_1, c_2, c_3, c_4, c_5) = (0, 1 + j, -1 + j, -1 - j, 1 - j)$ and $(q_1, q_2, q_3, q_4, q_5) = (0.8, 0.05, 0.05, 0.05, 0.05)$. The crosses indicate c_ℓ.

The function $\sum_{\ell=1}^{4} q_\ell \tilde{g}_\ell (s - c_\ell)$ has the minimum value only at $s = 0$ when $\tilde{g}_\ell(\cdot) = h^{(1)}_\star(\cdot)$, whereas it has the minimum value on $\{s \mid \mathrm{Re}\{s\} \in [-1, 1] \text{ and } \mathrm{Im}\{s\} \in [-1, 1]\}$ when $\tilde{g}_\ell(\cdot) = h^{(1)}_{\star\star}(\cdot)$. From the perspective of using the discreteness of $\boldsymbol{x} \in \mathcal{C}^N$, the function $\sum_{\ell=1}^{4} q_\ell \tilde{g}_\ell (s - c_\ell)$ should have the minimum value at least on $\mathcal{C} = \{1 + j, -1 + j, -1 - j, 1 - j\}$ and hence $h^{(1)}_{\star\star}(\cdot)$ is preferable in this case. As another example, we also show the contour for $(c_1, c_2, c_3, c_4, c_5) = (0, 1 + j, -1 + j, -1 - j, 1 - j)$ and $(q_1, q_2, q_3, q_4, q_5) = (0.8, 0.05, 0.05, 0.05, 0.05)$ in Fig. 3.2. The regularizers are selected as $g_1(\cdot) = h^{(1)}_\star(\cdot)$ and $\tilde{g}_\ell(\cdot) = h^{(1)}_{\star\star}(\cdot)$ $(\ell = 2, \ldots, 5)$ in Fig. 3.2(a), and $\tilde{g}_\ell(\cdot) = h^{(1)}_{\star\star}(\cdot)$ $(\ell = 1, \ldots, 5)$ in Fig. 3.2(b). When we use the regularizer in Fig. 3.2(b), either the real part or the imaginary part of s can be zero because the regularizer treats them independently. This property is not suitable for $\mathcal{C} = \{0, 1 + j, -1 + j, -1 - j, 1 - j\}$, where the real part be-

comes zero only when the imaginary part is zero. We thus should use the regularization with $g_1(\cdot) = h_{\star}^{(1)}(\cdot)$ and $\tilde{g}_\ell(\cdot) = h_{\star\star}^{(1)}(\cdot)$ $(\ell = 2, \ldots, 5)$ in Fig. 3.2(a) for discrete-valued vectors in $\{0, 1+j, -1+j, -1-j, 1-j\}^N$.

3.2.3 Proposed Algorithm for SCSR Optimization

We propose an algorithm for the SCSR optimization (3.1) on the basis of ADMM. The optimization problem (3.1) can be rewritten with new variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \in \mathbb{C}^N$ as

$$\begin{aligned} &\underset{\boldsymbol{s}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \in \mathbb{C}^N}{\text{minimize}} \left\{ \sum_{\ell=1}^{L} q_\ell \tilde{g}_\ell \left(\boldsymbol{z}_\ell - c_\ell \mathbf{1} \right) + \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}} \boldsymbol{s} \right\|_2^2 \right\} \\ &\text{subject to } \ \boldsymbol{s} = \boldsymbol{z}_\ell \ \ (\ell = 1, \ldots, L) \,. \end{aligned} \tag{3.6}$$

The problem (3.6) is further rewritten as the standard form of ADMM, i.e.,

$$\begin{aligned} &\underset{\boldsymbol{s} \in \mathbb{C}^N, \ \boldsymbol{z} \in \mathbb{C}^{LN}}{\text{minimize}} \ \{ f(\boldsymbol{s}) + g(\boldsymbol{z}) \} \\ &\text{subject to } \ \boldsymbol{\Phi} \boldsymbol{s} = \boldsymbol{z}, \end{aligned} \tag{3.7}$$

where $\boldsymbol{z} = \left[\boldsymbol{z}_1^{\mathsf{T}} \ \cdots \ \boldsymbol{z}_L^{\mathsf{T}} \right]^{\mathsf{T}} \in \mathbb{C}^{LN}$, $f(\boldsymbol{s}) = \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}} \boldsymbol{s} \right\|_2^2$, $g(\boldsymbol{z}) = \sum_{\ell=1}^{L} q_\ell \tilde{g}_\ell \left(\boldsymbol{z}_\ell - c_\ell \mathbf{1} \right)$, and $\boldsymbol{\Phi} = \left[\boldsymbol{I}_N \ \cdots \ \boldsymbol{I}_N \right]^{\mathsf{T}} \in \mathbb{R}^{LN \times N}$.

We derive the update equations of the proposed algorithm for the optimization problem (3.7). The augmented Lagrangian function for (3.7) is given by

$$\mathcal{L}_\rho(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\theta}) = f(\boldsymbol{s}) + g(\boldsymbol{z}) + 2\mathrm{Re}\{\boldsymbol{\theta}^{\mathsf{H}}(\boldsymbol{\Phi}\boldsymbol{s} - \boldsymbol{z})\} + \rho \|\boldsymbol{\Phi}\boldsymbol{s} - \boldsymbol{z}\|_2^2, \tag{3.8}$$

where $\boldsymbol{\theta} \in \mathbb{C}^{LN}$ and $\rho > 0$. The update equations of ADMM are given by

$$\boldsymbol{s}^{t+1} = \underset{\boldsymbol{s} \in \mathbb{C}^N}{\arg\min} \ \mathcal{L}_\rho(\boldsymbol{s}, \boldsymbol{z}^t, \boldsymbol{\theta}^t), \tag{3.9}$$

$$\boldsymbol{z}^{t+1} = \underset{\boldsymbol{z} \in \mathbb{C}^{LN}}{\arg\min} \ \mathcal{L}_\rho(\boldsymbol{s}^{t+1}, \boldsymbol{z}, \boldsymbol{\theta}^t), \tag{3.10}$$

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \rho(\boldsymbol{\Phi}\boldsymbol{s}^{t+1} - \boldsymbol{z}^{t+1}), \tag{3.11}$$

where t is the iteration index. From the identity $2\mathrm{Re}\{\boldsymbol{\theta}^{\mathsf{H}}\boldsymbol{u}\} + \rho\|\boldsymbol{u}\|_2^2 = \rho\|\boldsymbol{u}+\boldsymbol{w}\|_2^2 - \rho\|\boldsymbol{w}\|_2^2$ $(\boldsymbol{u} \in \mathbb{C}^{LN}$ and $\boldsymbol{w} = \boldsymbol{\theta}/\rho)$, we have

$$\boldsymbol{s}^{t+1} = \underset{\boldsymbol{s} \in \mathbb{C}^N}{\arg\min} \left\{ f(\boldsymbol{s}) + \rho \left\| \boldsymbol{\Phi}\boldsymbol{s} - \boldsymbol{z}^t + \boldsymbol{w}^t \right\|_2^2 \right\}, \tag{3.12}$$

$$\boldsymbol{z}^{t+1} = \underset{\boldsymbol{z} \in \mathbb{C}^{LN}}{\arg\min} \left\{ g(\boldsymbol{z}) + \rho \left\| \boldsymbol{\Phi}\boldsymbol{s}^{t+1} - \boldsymbol{z} + \boldsymbol{w}^t \right\|_2^2 \right\}, \tag{3.13}$$

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \boldsymbol{\Phi}\boldsymbol{s}^{t+1} - \boldsymbol{z}^{t+1}, \tag{3.14}$$

where $\boldsymbol{w}^t = \boldsymbol{\theta}^t/\rho \in \mathbb{C}^{LN}$.

The update of $\boldsymbol{s}^t$ in (3.12) can be written as

$$\boldsymbol{s}^{t+1} = \arg\min_{\boldsymbol{s}\in\mathbb{C}^N} \left\{ \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\boldsymbol{s} \right\|_2^2 + \rho \left\| \boldsymbol{\Phi}\boldsymbol{s} - \boldsymbol{z}^t + \boldsymbol{w}^t \right\|_2^2 \right\}. \tag{3.15}$$

The Wirtinger derivative [72] of the objective function in (3.15) is given by

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{s}^{\mathsf{H}}} \left\{ \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\boldsymbol{s} \right\|_2^2 + \rho \left\| \boldsymbol{\Phi}\boldsymbol{s} - \boldsymbol{z}^t + \boldsymbol{w}^t \right\|_2^2 \right\} \\ &= \left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}} \right) \boldsymbol{s} - \left(\rho \sum_{\ell=1}^{L} \left(\boldsymbol{z}_\ell^t - \boldsymbol{w}_\ell^t \right) + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{y}} \right), \end{aligned} \tag{3.16}$$

where $\boldsymbol{w}_\ell^t \in \mathbb{C}^N$ $(\ell = 1, \ldots, L)$ are subvectors of $\boldsymbol{w}^t$ defined as $\boldsymbol{w}^t = [(\boldsymbol{w}_1^t)^{\mathsf{T}} \cdots (\boldsymbol{w}_L^t)^{\mathsf{T}}]^{\mathsf{T}}$. We can thus rewrite (3.12) as

$$\boldsymbol{s}^{t+1} = \left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}} \right)^{-1} \left(\rho \sum_{\ell=1}^{L} \left(\boldsymbol{z}_\ell^t - \boldsymbol{w}_\ell^t \right) + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{y}} \right). \tag{3.17}$$

The update of $\boldsymbol{z}^t$ in (3.13) can be written with the proximity operator of $\frac{1}{2\rho} g(\cdot)$ as

$$\boldsymbol{z}^{t+1} = \operatorname{prox}_{\frac{1}{2\rho} g} \left(\boldsymbol{\Phi}\boldsymbol{s}^{t+1} + \boldsymbol{w}^t \right). \tag{3.18}$$

Let $\bar{g}_\ell(\boldsymbol{z}_\ell) \coloneqq \tilde{g}_\ell(\boldsymbol{z}_\ell - c_\ell \boldsymbol{1})$ and $\boldsymbol{u} = \left[\boldsymbol{u}_1^{\mathsf{T}} \cdots \boldsymbol{u}_L^{\mathsf{T}} \right]^{\mathsf{T}} \in \mathbb{C}^{LN}$ $(\boldsymbol{u}_\ell \in \mathbb{C}^N)$. The proximity operator of $\frac{1}{2\rho} g(\cdot)$ can be written as

$$\operatorname{prox}_{\frac{1}{2\rho} g}(\boldsymbol{u}) = \begin{bmatrix} \operatorname{prox}_{\frac{q_1}{2\rho} \bar{g}_1}(\boldsymbol{u}_1) \\ \vdots \\ \operatorname{prox}_{\frac{q_L}{2\rho} \bar{g}_L}(\boldsymbol{u}_L) \end{bmatrix} \tag{3.19}$$

$$= \begin{bmatrix} c_1 \boldsymbol{1} + \operatorname{prox}_{\frac{q_1}{2\rho} g_1}(\boldsymbol{u}_1 - c_1 \boldsymbol{1}) \\ \vdots \\ c_L \boldsymbol{1} + \operatorname{prox}_{\frac{q_L}{2\rho} g_L}(\boldsymbol{u}_L - c_L \boldsymbol{1}) \end{bmatrix} \tag{3.20}$$

because the function $g(\cdot)$ is separable as $g(\boldsymbol{z}) = \sum_{\ell=1}^{L} q_\ell \bar{g}_\ell(\boldsymbol{z}_\ell)$. From (3.19) to (3.20), we have used $\bar{g}_\ell(\boldsymbol{z}_\ell) = \tilde{g}_\ell(\boldsymbol{z}_\ell - c_\ell \boldsymbol{1})$ and the property of proximity operator for translation [61].

We thus need to calculate the proximity operator about $h_\star^{(1)}(\cdot)$ and $h_{\star\star}^{(1)}(\cdot)$ in (3.3) and (3.5), respectively, which are candidates of $\tilde{g}_\ell(\cdot)$. From the result in [71], the

proximity operator of $\gamma h_{\star}^{(1)}(\cdot)$ ($\gamma > 0$) is given by

$$\left[\mathrm{prox}_{\gamma h_{\star}^{(1)}}(\boldsymbol{u})\right]_n = \begin{cases} (|[\boldsymbol{u}]_n| - \gamma)\dfrac{[\boldsymbol{u}]_n}{|[\boldsymbol{u}]_n|} & (|[\boldsymbol{u}]_n| \geq \gamma) \\ 0 & (|[\boldsymbol{u}]_n| < \gamma) \end{cases}, \tag{3.21}$$

where $\boldsymbol{u} \in \mathbb{C}^N$. We can transform the proximity operator of $\gamma h_{\star\star}^{(1)}(\cdot)$ as

$$\begin{aligned} &\mathrm{prox}_{\gamma h_{\star\star}^{(1)}}(\boldsymbol{u}) \\ &= \underset{\boldsymbol{s}\in\mathbb{C}^N}{\arg\min} \left\{ \gamma h_{\star\star}^{(1)}(\boldsymbol{s}) + \frac{1}{2}\|\boldsymbol{s}-\boldsymbol{u}\|_2^2 \right\} \qquad (3.22) \\ &= \underset{\substack{\boldsymbol{s}=\boldsymbol{s}_{\mathrm{R}}+j\boldsymbol{s}_{\mathrm{I}}\in\mathbb{C}^N \\ (\boldsymbol{s}_{\mathrm{R}},\boldsymbol{s}_{\mathrm{I}}\in\mathbb{R}^N)}}{\arg\min} \left\{ \left(\gamma\|\boldsymbol{s}_{\mathrm{R}}\|_1 + \frac{1}{2}\|\boldsymbol{s}_{\mathrm{R}}-\boldsymbol{u}_{\mathrm{R}}\|_2^2\right) + \left(\gamma\|\boldsymbol{s}_{\mathrm{I}}\|_1 + \frac{1}{2}\|\boldsymbol{s}_{\mathrm{I}}-\boldsymbol{u}_{\mathrm{I}}\|_2^2\right) \right\}, \qquad (3.23) \end{aligned}$$

where $\boldsymbol{u}_{\mathrm{R}} \coloneqq \mathrm{Re}\{\boldsymbol{u}\} \in \mathbb{R}^N$ and $\boldsymbol{u}_{\mathrm{I}} \coloneqq \mathrm{Im}\{\boldsymbol{u}\} \in \mathbb{R}^N$ are the real and the imaginary parts of $\boldsymbol{u} \in \mathbb{C}^N$, respectively. The minimization with respect to $\boldsymbol{s} \in \mathbb{C}^N$ in (3.23) can be divided into the minimization with respect to $\boldsymbol{s}_{\mathrm{R}} \in \mathbb{R}^N$ and $\boldsymbol{s}_{\mathrm{I}} \in \mathbb{R}^N$. We can thus write $\mathrm{prox}_{\gamma h_{\star\star}^{(1)}}(\boldsymbol{u})$ with the proximity operator of the ℓ_1 norm in the real-valued domain as

$$\begin{aligned} &\left[\mathrm{prox}_{\gamma h_{\star\star}^{(1)}}(\boldsymbol{u})\right]_n \\ &= \left[\mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{u}_{\mathrm{R}})\right]_n + j\cdot\left[\mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{u}_{\mathrm{I}})\right]_n \qquad (3.24) \\ &= \mathrm{sign}\left([\boldsymbol{u}_{\mathrm{R}}]_n\right)\max\left(|[\boldsymbol{u}_{\mathrm{R}}]_n| - \gamma, 0\right) + j\cdot\mathrm{sign}\left([\boldsymbol{u}_{\mathrm{I}}]_n\right)\max\left(|[\boldsymbol{u}_{\mathrm{I}}]_n| - \gamma, 0\right), \qquad (3.25) \end{aligned}$$

where $[\boldsymbol{u}_{\mathrm{R}}]_n$ and $[\boldsymbol{u}_{\mathrm{I}}]_n$ are the nth element of $\boldsymbol{u}_{\mathrm{R}}$ and $\boldsymbol{u}_{\mathrm{I}}$, respectively. By using (3.21) or (3.25), we can compute the proximity operator of $\frac{1}{2\rho}g(\cdot)$ in (3.20).

We summarize the proposed algorithm for the SCSR optimization (3.7) in Algorithm 3.1. The order of the computational complexity is $O\left(N^3\right)$ because it is dominated by the inverse matrix $\left(\rho L\boldsymbol{I}_N + \lambda\tilde{\boldsymbol{A}}^{\mathsf{H}}\tilde{\boldsymbol{A}}\right)^{-1}$. It should be noted that the computation is required only once in the algorithm. Once we obtain the inverse matrix, the update equations of the proposed algorithm can be performed with $O(N^2)$. Note that the proposed algorithm does not require the proximity operator of $\sum_{\ell=1}^{L} q_\ell\tilde{g}_\ell\left(\boldsymbol{s} - c_\ell\boldsymbol{1}\right)$, which depends on the selection of $\tilde{g}_\ell(\cdot)$. We can implement the proposed algorithm only with $\mathrm{prox}_{\frac{q_\ell}{2\rho}\tilde{g}_\ell}(\cdot)$ given by $\mathrm{prox}_{\gamma h_{\star}^{(1)}}(\cdot)$ in (3.21) or $\mathrm{prox}_{\gamma h_{\star\star}^{(1)}}(\cdot)$ in (3.25).

3.2.4 IW-SCSR

The SOAV optimization has been extended to IW-SOAV optimization to use the prior information about the unknown vector in Chapter 2. In Chapter 2, an iterative approach

Algorithm 3.1 Proposed Algorithm for SCSR Optimization (3.7)

Input: $\tilde{\boldsymbol{y}} \in \mathbb{C}^M$, $\tilde{\boldsymbol{A}} \in \mathbb{C}^{M \times N}$
Output: $\hat{\boldsymbol{x}} \in \mathbb{C}^N$

1: Fix $\rho > 0$, $\boldsymbol{z}^0 \in \mathbb{C}^{LN}$, and $\boldsymbol{w}^0 \in \mathbb{C}^{LN}$
2: **for** $t = 0$ to $T_{\mathrm{itr}} - 1$ **do**
3: $\quad \boldsymbol{s}^{t+1} = \left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}}\right)^{-1} \left(\rho \sum_{\ell=1}^{L} \left(\boldsymbol{z}_\ell^t - \boldsymbol{w}_\ell^t\right) + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{y}}\right)$
4: $\quad \boldsymbol{z}_\ell^{t+1} = c_\ell \boldsymbol{1} + \mathrm{prox}_{\frac{q_\ell}{2\rho} \tilde{g}_\ell} \left(\boldsymbol{s}^{t+1} + \boldsymbol{w}_\ell^t - c_\ell \boldsymbol{1}\right) \ (\ell = 1, \ldots, L)$
5: $\quad \boldsymbol{w}_\ell^{t+1} = \boldsymbol{w}_\ell^t + \boldsymbol{s}^{t+1} - \boldsymbol{z}_\ell^{t+1} \ (\ell = 1, \ldots, L)$
6: **end for**
7: $\hat{\boldsymbol{x}} = \boldsymbol{s}^{T_{\mathrm{itr}}}$

named IW-SOAV has also been proposed to obtain better performance. The IW-SOAV iterates the W-SOAV optimization with updating parameters in the objective function.

Assuming that the sparse regularizer $\tilde{g}_\ell(\cdot)$ is element-wise as $h_\star^{(1)}(\cdot)$ or $h_{\star\star}^{(1)}(\cdot)$, we here extend the SCSR optimization problem (3.1) to the W-SCSR optimization given by

$$\underset{\boldsymbol{s} \in \mathbb{C}^N}{\text{minimize}} \ \left\{ \sum_{\ell=1}^{L} \sum_{n=1}^{N} q_{n,\ell} \tilde{g}_\ell \left(s_n - c_\ell\right) + \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}} \boldsymbol{s} \right\|_2^2 \right\}, \tag{3.26}$$

which is equivalent to

$$\begin{aligned} &\underset{\boldsymbol{s}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \in \mathbb{C}^N}{\text{minimize}} \ \left\{ \sum_{\ell=1}^{L} \sum_{n=1}^{N} q_{n,\ell} \tilde{g}_\ell \left(z_{n,\ell} - c_\ell\right) + \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}} \boldsymbol{s} \right\|_2^2 \right\} \\ &\text{subject to} \ \ \boldsymbol{s} = \boldsymbol{z}_\ell \ \ (\ell = 1, \ldots, L). \end{aligned} \tag{3.27}$$

Here, $q_{n,\ell}$ is the parameter and $z_{n,\ell}$ is the nth element of $\boldsymbol{z}_\ell$ ($n = 1, \ldots, N$ and $\ell = 1, \ldots, L$). Note that we can use different parameters $q_{n,\ell}$ for each element s_n of $\boldsymbol{s}$, whereas a common parameter q_ℓ is used for all s_n in (3.1). The optimization problem (3.27) can be further rewritten as

$$\begin{aligned} &\underset{\boldsymbol{s} \in \mathbb{C}^N, \ \boldsymbol{z} \in \mathbb{C}^{LN}}{\text{minimize}} \ \left\{ f\left(\boldsymbol{s}\right) + g_{\mathrm{w}}\left(\boldsymbol{z}\right) \right\} \\ &\text{subject to} \ \ \boldsymbol{\Phi} \boldsymbol{s} = \boldsymbol{z}, \end{aligned} \tag{3.28}$$

where $g_{\mathrm{w}}\left(\boldsymbol{z}\right) = \sum_{\ell=1}^{L} \sum_{n=1}^{N} q_{n,\ell} \tilde{g}_\ell \left(z_{n,\ell} - c_\ell\right)$. The optimization algorithm for (3.28) can be obtained by replacing $\mathrm{prox}_{\frac{1}{2\rho} g}(\cdot)$ in (3.20) with $\mathrm{prox}_{\frac{1}{2\rho} g_{\mathrm{w}}}(\cdot)$. By using the same approach as in (3.20), $\mathrm{prox}_{\frac{1}{2\rho} g_{\mathrm{w}}}\left(\boldsymbol{u}\right)$ is given by

$$\left[\mathrm{prox}_{\frac{1}{2\rho} g_{\mathrm{w}}}\left(\boldsymbol{u}\right)\right]_{(\ell-1)N+n} = c_\ell + \mathrm{prox}_{\frac{q_{n,\ell}}{2\rho} \tilde{g}_\ell}\left(u_{n,\ell} - c_\ell\right), \tag{3.29}$$

Algorithm 3.2 IW-SCSR

Input: $\tilde{\boldsymbol{y}} \in \mathbb{C}^M$, $\tilde{\boldsymbol{A}} \in \mathbb{C}^{M \times N}$
Output: $\hat{\boldsymbol{x}} \in \mathbb{C}^N$

Initialize $q_{n,\ell}$ ($n = 1, \ldots, N$ and $\ell = 1, \ldots, L$).
for $k = 1$ to K_{itr} **do**
  Fix $\beta_{\mathrm{r}} > 0$, $\rho > 0$, $\boldsymbol{z}^0 \in \mathbb{C}^{LN}$, and $\boldsymbol{w}^0 \in \mathbb{C}^{LN}$
  $\lambda = \dfrac{\sum_{\ell'=1}^{L} p_{\ell'} \sum_{\ell=1}^{L} \sum_{n=1}^{N} q_{n,\ell} \tilde{g}_\ell (c_{\ell'} - c_\ell)}{\beta_{\mathrm{r}} M \sigma_{\mathrm{v}}^2}$
  for $t = 0$ to $T_{\mathrm{itr}} - 1$ **do**
    $\boldsymbol{s}^{t+1} = \left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}}\right)^{-1} \left(\rho \sum_{\ell=1}^{L} \left(\boldsymbol{z}_\ell^t - \boldsymbol{w}_\ell^t\right) + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{y}}\right)$
    $z_{n,\ell}^{t+1} = c_\ell + \mathrm{prox}_{\frac{q_{n,\ell}}{2\rho} \tilde{g}_\ell} \left(s_n^{t+1} + w_{n,\ell}^t - c_\ell\right)$ ($n = 1, \ldots, N$ and $\ell = 1, \ldots, L$)
    $\boldsymbol{w}_\ell^{t+1} = \boldsymbol{w}_\ell^t + \boldsymbol{s}^{t+1} - \boldsymbol{z}_\ell^{t+1}$ ($\ell = 1, \ldots, L$)
  end for
  $d_{n,\ell} = \left| s_n^{T_{\mathrm{itr}}} - c_\ell \right|$ ($n = 1, \ldots, N$ and $\ell = 1, \ldots, L$)
  $q_{n,\ell} = \dfrac{d_{n,\ell}^{-1}}{\sum_{\ell'=1}^{L} d_{n,\ell'}^{-1}}$ ($n = 1, \ldots, N$ and $\ell = 1, \ldots, L$)
end for
$\hat{\boldsymbol{x}} = \boldsymbol{s}^{T_{\mathrm{itr}}}$

where $u_{n,\ell}$ denotes the nth element of $\boldsymbol{u}_\ell$ ($\ell = 1, \ldots, L$ and $n = 1, \ldots, N$). The coefficient $q_{n,\ell}/2\rho$ of $\tilde{g}_\ell(\cdot)$ depends not only on ℓ but also on n in (3.29) unlike $\mathrm{prox}_{\frac{q_\ell}{2\rho} \tilde{g}_\ell}(\cdot)$ in (3.20).

We propose an iterative approach called IW-SCSR in Algorithm 3.2, where we iteratively calculate the solution of the W-SCSR optimization (3.28) with the update of the parameter $q_{n,\ell}$. In such an iterative approach, the parameter $q_{n,\ell}$ can be updated by using the estimate at the previous iteration $\hat{\boldsymbol{x}}^{\mathrm{pre}} = \left[\hat{x}_1^{\mathrm{pre}} \ \cdots \ \hat{x}_N^{\mathrm{pre}}\right]^{\mathsf{T}}$. In this paper, we propose a parameter update given by

$$q_{n,\ell} = \frac{d_{n,\ell}^{-1}}{\sum_{\ell'=1}^{L} d_{n,\ell'}^{-1}}, \tag{3.30}$$

where $d_{n,\ell} = \left| \hat{x}_n^{\mathrm{pre}} - c_\ell \right|$ is the distance between $\hat{x}_n^{\mathrm{pre}}$ and c_ℓ. The denominator of (3.30) has a role for the normalization of $q_{n,\ell}$, i.e., $\sum_{\ell=1}^{L} q_{n,\ell} = 1$ ($n = 1, \ldots, N$). If $d_{n,\ell}$ is small, then the corresponding $q_{n,\ell}$ becomes large and the estimate of x_n will be close to c_ℓ.

3.2.5 Selection of Parameter λ

The performance of the W-SCSR optimization (3.27), (3.28) depends on the selection of the parameter λ, which controls the balance between $f(\boldsymbol{s})$ and $\tilde{g}(\boldsymbol{z}) = \tilde{g}(\boldsymbol{\Phi}\boldsymbol{s})$. The

value of $f(\boldsymbol{s})$ for the true vector $\tilde{\boldsymbol{x}}$ is given by $f(\tilde{\boldsymbol{x}}) = \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} \right\|_2^2 = \lambda \left\| \tilde{\boldsymbol{v}} \right\|_2^2$ and hence the optimal value of λ depends on the noise variance σ_{v}^2 in general. We thus need to choose a good value of λ depending on the noise variance. To tackle this problem, we propose an adaptive parameter selection method taking the noise variance into account. Specifically, we determine λ so that the ratio $\mathrm{E}\left[\tilde{g}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{x}}\right)\right] / \mathrm{E}\left[f\left(\tilde{\boldsymbol{x}}\right)\right]$ becomes a constant β_{r} (> 0), i.e.,

$$\frac{\mathrm{E}\left[\tilde{g}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{x}}\right)\right]}{\mathrm{E}\left[f\left(\tilde{\boldsymbol{x}}\right)\right]} = \beta_{\mathrm{r}}, \tag{3.31}$$

where $\mathrm{E}\left[\cdot\right]$ represents the expectation with respect to the distributions of $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{v}}$. Note that we use the expectation $\mathrm{E}\left[f\left(\tilde{\boldsymbol{x}}\right)\right]$ and $\mathrm{E}\left[\tilde{g}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{x}}\right)\right]$ instead of $f\left(\tilde{\boldsymbol{x}}\right)$ and $\tilde{g}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{x}}\right)$ because the true vector $\tilde{\boldsymbol{x}}$ and the noise vector $\tilde{\boldsymbol{v}}$ are unknown. Since we can calculate the left side of (3.31) as

$$\frac{\mathrm{E}\left[\tilde{g}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{x}}\right)\right]}{\mathrm{E}\left[f\left(\tilde{\boldsymbol{x}}\right)\right]} = \frac{\sum_{\ell'=1}^{L} p_{\ell'} \sum_{\ell=1}^{L} \sum_{n=1}^{N} q_{n,\ell} \tilde{g}_{\ell}(c_{\ell'} - c_{\ell})}{\lambda M \sigma_{\mathrm{v}}^2}, \tag{3.32}$$

the proposed λ is given by

$$\lambda = \frac{\sum_{\ell'=1}^{L} p_{\ell'} \sum_{\ell=1}^{L} \sum_{n=1}^{N} q_{n,\ell} \tilde{g}_{\ell}(c_{\ell'} - c_{\ell})}{\beta_{\mathrm{r}} M \sigma_{\mathrm{v}}^2}. \tag{3.33}$$

Once we fix the value of β_{r}, the proposed λ in (3.33) adaptively changes in accordance with the noise variance σ_{v}^2. The proposed λ becomes large when the noise variance σ_{v}^2 is small, and vice versa.

3.2.6 Computational Complexity Reduction

The order of the computational complexity of LW-SCSR is dominated by the inverse matrix $\left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}}\right)^{-1}$, which requires the complexity of $O(N^3)$. If we update the parameter λ at each outer iteration k, we need to compute K_{itr} inverse matrices as a whole. However, the calculation of these inverse matrices can be eliminated by computing the singular value decomposition (SVD) of $\tilde{\boldsymbol{A}}$ before executing the algorithm. In the underdetermined case with $M < N$, the SVD of $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{V}^{\mathsf{H}}$, where $\boldsymbol{\Xi} = \left[\mathrm{diag}(\xi_1, \ldots, \xi_M) \ \boldsymbol{0}_{M \times (N-M)}\right] \in \mathbb{R}^{M \times N}$ is a rectangular diagonal matrix with the singular values $\xi_1, \ldots, \xi_M$ of $\boldsymbol{A}$. $\boldsymbol{U} \in \mathbb{C}^{M \times M}$ and $\boldsymbol{V} \in \mathbb{C}^{N \times N}$ are unitary matrices composed of the left and right singular vectors, respectively. With the SVD of $\tilde{\boldsymbol{A}}$, we can rewrite the inverse matrix as

$$\begin{aligned} &\left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}}\right)^{-1} \\ &= \boldsymbol{V} \mathrm{diag}\left(\frac{1}{\rho L + \lambda \xi_1^2}, \ldots, \frac{1}{\rho L + \lambda \xi_M^2}, \frac{1}{\rho L}, \ldots, \frac{1}{\rho L}\right) \boldsymbol{V}^{\mathsf{H}}. \end{aligned} \tag{3.34}$$

Once we obtain the SVD of $\tilde{\boldsymbol{A}}$, we do not need to directly compute the inverse matrix $\left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}}\right)^{-1}$ at each outer iteration t even if the parameter λ is updated in the algorithm. It should be noted that the order of the overall computational complexity of IW-SCSR is still $O(MN \min(M, N))$ because we require the SVD of the measurement matrix $\tilde{\boldsymbol{A}}$.

When the measurement matrix $\tilde{\boldsymbol{A}}$ has some special structure, we can compute the inverse matrix more efficiently. As an example, we consider the channel equalization in Section 1.2. Taking advantage of the block circulant structure of $\tilde{\boldsymbol{A}}$ in (1.14), we can efficiently compute the inverse matrix $\left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}}\right)^{-1}$. The matrix $\tilde{\boldsymbol{A}}$ in (1.14) can be decomposed as

$$\tilde{\boldsymbol{A}} = \left(\boldsymbol{D}^{\mathsf{H}} \otimes \boldsymbol{I}_{N_{\mathrm{r}}}\right) \boldsymbol{B} \left(\boldsymbol{D} \otimes \boldsymbol{I}_{N_{\mathrm{t}}}\right), \tag{3.35}$$

where $\boldsymbol{D} \in \mathbb{C}^{Q_{\mathrm{b}} \times Q_{\mathrm{b}}}$ denotes the normalized DFT matrix. The matrix $\boldsymbol{B}$ is block diagonal given by

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{B}_{Q_{\mathrm{b}}} \end{bmatrix} \in \mathbb{C}^{N_{\mathrm{r}} Q_{\mathrm{b}} \times N_{\mathrm{t}} Q_{\mathrm{b}}}, \tag{3.36}$$

where

$$\boldsymbol{B}_q = \sum_{i=0}^{L_{\mathrm{p}}-1} \boldsymbol{\Gamma}^{(i)} \omega^{i(q-1)} \in \mathbb{C}^{N_{\mathrm{r}} \times N_{\mathrm{t}}} \tag{3.37}$$

and $\omega = e^{2\pi j / Q_{\mathrm{b}}}$ $(q = 1, \ldots, Q_{\mathrm{b}})$. From (3.35), the inverse matrix in IW-SCSR can be rewritten as

$$\begin{aligned} &\left(\rho L \boldsymbol{I}_N + \lambda \tilde{\boldsymbol{A}}^{\mathsf{H}} \tilde{\boldsymbol{A}}\right)^{-1} \\ &= \left(\rho L \boldsymbol{I}_N + \lambda \left(\boldsymbol{D}^{\mathsf{H}} \otimes \boldsymbol{I}_{N_{\mathrm{t}}}\right) \boldsymbol{B}^{\mathsf{H}} \boldsymbol{B} \left(\boldsymbol{D} \otimes \boldsymbol{I}_{N_{\mathrm{t}}}\right)\right)^{-1} && (3.38) \\ &= \left(\boldsymbol{D}^{\mathsf{H}} \otimes \boldsymbol{I}_{N_{\mathrm{t}}}\right) \left(\rho L \boldsymbol{I}_N + \lambda \boldsymbol{B}^{\mathsf{H}} \boldsymbol{B}\right)^{-1} \left(\boldsymbol{D} \otimes \boldsymbol{I}_{N_{\mathrm{t}}}\right) && (3.39) \\ &= \left(\boldsymbol{D}^{\mathsf{H}} \otimes \boldsymbol{I}_{N_{\mathrm{t}}}\right) \begin{bmatrix} \boldsymbol{R}_1^{-1} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{R}_{Q_{\mathrm{b}}}^{-1} \end{bmatrix} \left(\boldsymbol{D} \otimes \boldsymbol{I}_{N_{\mathrm{t}}}\right), && (3.40) \end{aligned}$$

where $\boldsymbol{R}_q = \rho L \boldsymbol{I}_{N_{\mathrm{t}}} + \lambda \boldsymbol{B}_q^{\mathsf{H}} \boldsymbol{B}_q \in \mathbb{C}^{N_{\mathrm{t}} \times N_{\mathrm{t}}}$. The size of the inverse matrices is reduced to $N_{\mathrm{t}} \times N_{\mathrm{t}}$ in (3.40) and the required computational complexity becomes $O(N_{\mathrm{t}}^3 Q_{\mathrm{b}})$, which is significantly smaller than the original direct calculation with $O(N^3) = O(N_{\mathrm{t}}^3 Q_{\mathrm{b}}^3)$. Note that we can compute the inverse matrix $\boldsymbol{R}_q^{-1}$ in the same way as (3.34) with the SVD of $\boldsymbol{B}_q$. The property in (3.35) is also used in [18] to propose an equalization method in the frequency domain.

3.2.7 Convergence Property

Here, we investigate the convergence of the proposed algorithm for the W-SCSR optimization problem (3.28). By extending the result for ADMM in the real-valued domain [70], residual convergence, objective convergence, and dual variable convergence have been proved for ADMM in the complex-valued domain [71]. However, the convergence of $\{\boldsymbol{s}^t\}$ to the optimizer of the problem has not been discussed in [71]. We thus prove the following theorem about the convergence of $\{\boldsymbol{s}^t\}$ for (3.28).

Theorem 3.2.1. Assume that the Lagrangian function $\tilde{\mathcal{L}}(\boldsymbol{s},\boldsymbol{z},\boldsymbol{\theta}) \coloneqq f(\boldsymbol{s}) + \tilde{g}(\boldsymbol{z}) + 2\mathrm{Re}\{\boldsymbol{\theta}^{\mathsf{H}}(\boldsymbol{\Phi}\boldsymbol{s}-\boldsymbol{z})\}$ for (3.28) has a saddle point, i.e., there is $(\boldsymbol{s}^*,\boldsymbol{z}^*,\boldsymbol{\theta}^*)$ such that

$$\tilde{\mathcal{L}}(\boldsymbol{s}^*,\boldsymbol{z}^*,\boldsymbol{\theta}) \le \tilde{\mathcal{L}}(\boldsymbol{s}^*,\boldsymbol{z}^*,\boldsymbol{\theta}^*) \le \tilde{\mathcal{L}}(\boldsymbol{s},\boldsymbol{z},\boldsymbol{\theta}^*) \tag{3.41}$$

holds for any $\boldsymbol{s},\boldsymbol{z},\boldsymbol{\theta}$. Also, assume that the sparse regularizer $\tilde{g}_\ell(\cdot)$ is $h_\star^{(1)}(\cdot)$ or $h_{\star\star}^{(1)}(\cdot)$. Then, the sequence $\{\boldsymbol{s}^t\}$ $(t=1,2,\dots)$ obtained by the proposed algorithm for (3.28) converges to the optimal solution of (3.28).

Proof. The functions $f(\boldsymbol{s})$ and $\tilde{g}(\boldsymbol{z})$ are proper, closed, and convex. From Theorem 16 in [71], we have the residual convergence $\boldsymbol{\Phi}\boldsymbol{s}^t - \boldsymbol{z}^t \to \boldsymbol{0}$ and the objective convergence $f(\boldsymbol{s}^t) + \tilde{g}(\boldsymbol{z}^t) \to f(\boldsymbol{s}^*) + \tilde{g}(\boldsymbol{z}^*)$ $(t\to\infty)$. Note that $\boldsymbol{s}^*$ and $\boldsymbol{z}^*$ are the optimal values of $\boldsymbol{s}$ and $\boldsymbol{z}$ in (3.28), respectively, and satisfy $\boldsymbol{\Phi}\boldsymbol{s}^* = \boldsymbol{z}^*$.

In order to see the convergence of $\{\boldsymbol{s}^t\}$ to one of the optimizers of (3.28), we evaluate $\left|\tilde{g}(\boldsymbol{\Phi}\boldsymbol{s}^t) - \tilde{g}(\boldsymbol{z}^t)\right|$, which is upper bounded as

$$\left|\tilde{g}(\boldsymbol{\Phi}\boldsymbol{s}^t) - \tilde{g}(\boldsymbol{z}^t)\right| \le \sum_{\ell=1}^{L}\sum_{n=1}^{N} q_{n,\ell}\left|\tilde{g}_\ell(s_n^t - c_\ell) - \tilde{g}_\ell\left(z_{n,\ell}^t - c_\ell\right)\right|. \tag{3.42}$$

When $\tilde{g}_\ell(\cdot) = h_\star^{(1)}(\cdot)$, we have

$$\begin{aligned}
&\left|\tilde{g}_\ell(s_n^t - c_\ell) - \tilde{g}_\ell\left(z_{n,\ell}^t - c_\ell\right)\right| \\
&= \left|\left|s_n^t - c_\ell\right| - \left|z_{n,\ell}^t - c_\ell\right|\right| && (3.43)\\
&\le \left|s_n^t - z_{n,\ell}^t\right| && (3.44)\\
&\to 0 \quad (t\to\infty) && (3.45)
\end{aligned}$$

because $\boldsymbol{s}^t - \boldsymbol{z}_\ell^t \to \boldsymbol{0}$. When $\tilde{g}_\ell(\cdot) = h_{\star\star}^{(1)}(\cdot)$, we can also obtain

$$\begin{aligned}
&\left|\tilde{g}_\ell(s_n^t - c_\ell) - \tilde{g}_\ell\left(z_{n,\ell}^t - c_\ell\right)\right| \\
&= \left|\left(\left|\mathrm{Re}\{s_n^t - c_\ell\}\right| + \left|\mathrm{Im}\{s_n^t - c_\ell\}\right|\right) - \left(\left|\mathrm{Re}\{z_{n,\ell}^t - c_\ell\}\right| + \left|\mathrm{Im}\{z_{n,\ell}^t - c_\ell\}\right|\right)\right| && (3.46)\\
&\le \left|\mathrm{Re}\{s_n^t - z_{n,\ell}^t\}\right| + \left|\mathrm{Im}\{s_n^t - z_{n,\ell}^t\}\right| && (3.47)\\
&\to 0 \quad (t\to\infty). && (3.48)
\end{aligned}$$

From (3.42), (3.45), and (3.48), we have $\left|\tilde{g}\left(\boldsymbol{\Phi}\boldsymbol{s}^t\right) - \tilde{g}\left(\boldsymbol{z}^t\right)\right| \to 0$ $(k \to \infty)$ and

$$\begin{aligned}
&\left|f\left(\boldsymbol{s}^t\right) + \tilde{g}\left(\boldsymbol{\Phi}\boldsymbol{s}^t\right) - \left(f\left(\boldsymbol{s}^*\right) + \tilde{g}\left(\boldsymbol{\Phi}\boldsymbol{s}^*\right)\right)\right| \\
&\leq \left|f\left(\boldsymbol{s}^t\right) + \tilde{g}\left(\boldsymbol{z}^t\right) - \left(f\left(\boldsymbol{s}^*\right) + \tilde{g}\left(\boldsymbol{z}^*\right)\right)\right| + \left|\tilde{g}\left(\boldsymbol{\Phi}\boldsymbol{s}^t\right) - \tilde{g}\left(\boldsymbol{z}^t\right)\right| \qquad (3.49) \\
&\to 0 \quad (t \to \infty)\,. \qquad (3.50)
\end{aligned}$$

Hence, $f\left(\boldsymbol{s}^t\right) + \tilde{g}\left(\boldsymbol{\Phi}\boldsymbol{s}^t\right)$ converges to the optimal value of the objective function. Since the objective function is continuous, we conclude that $\left\{\boldsymbol{s}^t\right\}$ converges to one of optimizers of (3.28). □

3.3 Simulation Results

In this section, we evaluate the performance of the proposed method by computer simulations. We consider MIMO signal detection and channel equalization described in Section 1.2. In both cases, the additive noise vector $\tilde{\boldsymbol{v}}$ is assumed to be circular complex Gaussian distributed with mean $\boldsymbol{0}$ and covariance matrix $\sigma_{\mathrm{v}}^2 \boldsymbol{I}_M$.

3.3.1 MIMO Signal Detection

We first compare the performance of the proposed SCSR optimization and the SOAV optimization in the real-valued domain. Figure 3.3 shows the average of SER defined as $\|\boldsymbol{Q}(\hat{\boldsymbol{x}}) - \tilde{\boldsymbol{x}}\|_0 / N$ for QPSK with $C = \{1 + j, -1 + j, -1 - j, 1 - j\}$, where $\hat{\boldsymbol{x}}$ is the estimate of $\tilde{\boldsymbol{x}}$ and $\boldsymbol{Q}(\hat{\boldsymbol{x}}) = \arg\min_{\boldsymbol{s} \in C^N} \|\boldsymbol{s} - \hat{\boldsymbol{x}}\|_1$. The result is obtained by averaging the SER over 1,000 independent realizations of the measurement matrix. The problem size is $(N, M) = (50, 40)$. We assume i.i.d. flat Rayleigh fading channels and hence the measurement matrix $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}}_{\text{i.i.d.}} \in \mathbb{C}^{M \times N}$ is composed of i.i.d. circular complex Gaussian variables with zero mean and unit variance. The SNR is defined as $\mathrm{E}\left[\|\tilde{\boldsymbol{x}}\|_2^2\right] / \sigma_{\mathrm{v}}^2$. In the figure, we denote the LMMSE method by 'LMMSE,' the SOAV optimization in the real-valued domain by 'SOAV,' the SCSR optimization with $\tilde{g}_\ell(\cdot) = h_\star^{(1)}(\cdot)$ $(\ell = 1, \ldots, 4)$ by 'SCSR $(h_\star^{(1)}(\cdot))$,' and the SCSR optimization with $\tilde{g}_\ell(\cdot) = h_{\star\star}^{(1)}(\cdot)$ $(\ell = 1, \ldots, 4)$ by 'SCSR $(h_{\star\star}^{(1)}(\cdot))$.' The parameter of the SCSR optimization is fixed as $q_{n,\ell} = 1/4$. The parameter λ is determined from (3.33) with $\beta_{\mathrm{r}} = 15$, which achieves good performance in the simulation. The parameter ρ in the proposed algorithm is $\rho = 0.1$ and the number of inner iterations is $T_{\mathrm{itr}} = 100$. From the figure, we can see that the SCSR optimization with $h_{\star\star}^{(1)}(\cdot)$, which treats the real and imaginary part separately, can achieve the same performance as the SOAV optimization in the real-valued domain.

We then investigate the convergence of the proposed algorithm for the SCSR optimization. In Fig. 3.4, we show the convergence curve of the algorithm for $\rho = 0.01, 0.1$, and 0.3. The problem size is $(N, M) = (50, 40)$ and $C = \{1 + j, -1 + j, -1 - j, 1 - j\}$. The SNR is 17.5 dB. The regularizer is $\tilde{g}_\ell(\cdot) = h_{\star\star}^{(1)}(\cdot)$ $(\ell = 1, \ldots, 4)$ and we fix $q_{n,\ell} = 1/4$

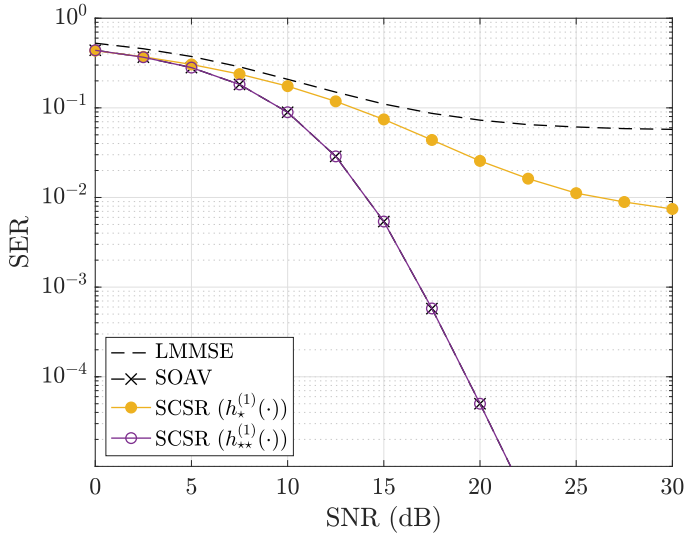


Figure 3.3: SER in i.i.d. MIMO channels (QPSK, $(N, M) = (50, 40)$, $\beta_{\mathrm{r}} = 15$)

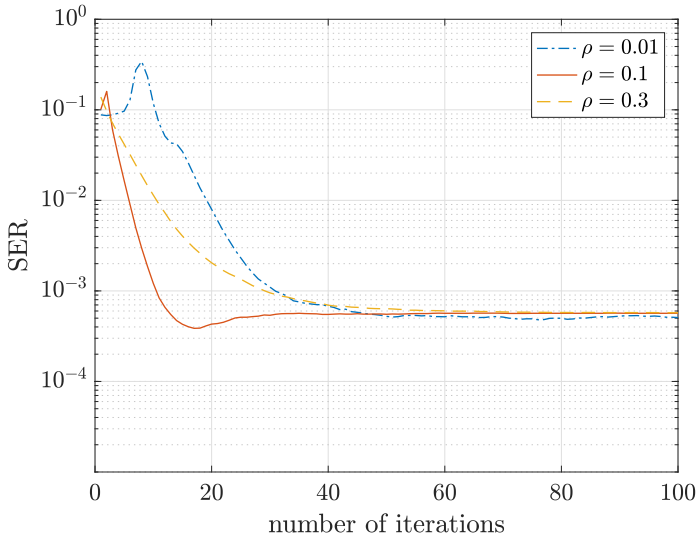


Figure 3.4: SER versus the number of iterations in i.i.d. MIMO channels (QPSK, $(N, M) = (50, 40)$, SNR = 17.5 dB, $\beta_{\mathrm{r}} = 15$)

and $\beta_{\mathrm{r}} = 15$. We can see that these three curves converge to almost the same SER if the number of iterations is large enough. Since $\rho = 0.1$ achieves the fastest convergence of the three and 100 iterations are enough to convergent in the figure, we use these values hereafter.

In Figs. 3.5 and 3.6, we compare the SER performance of the proposed IW-SCSR and some conventional methods for QPSK with $(N, M) = (50, 40)$ in i.i.d. and correlated

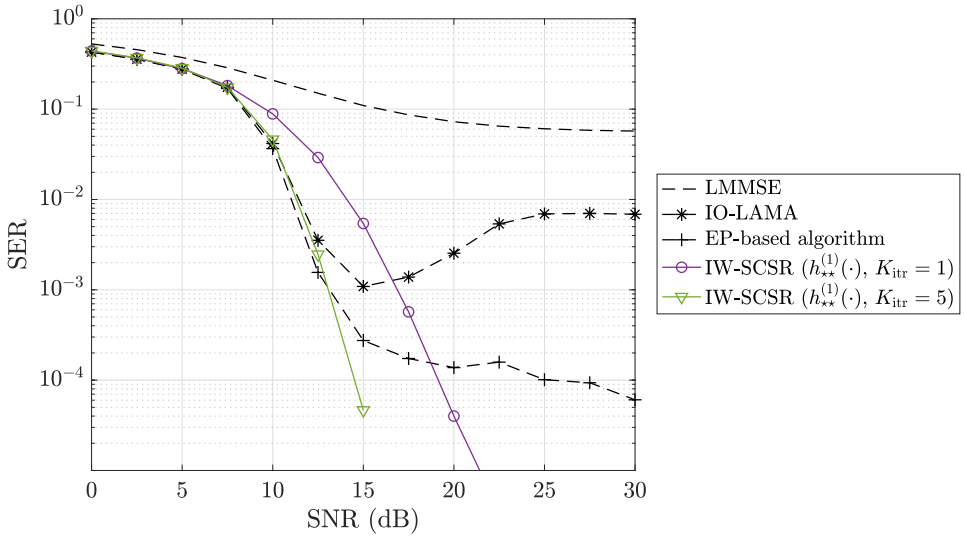


Figure 3.5: SER in i.i.d. MIMO channels (QPSK, $(N, M) = (50, 40)$, $\beta_{\rm r} = 15$)

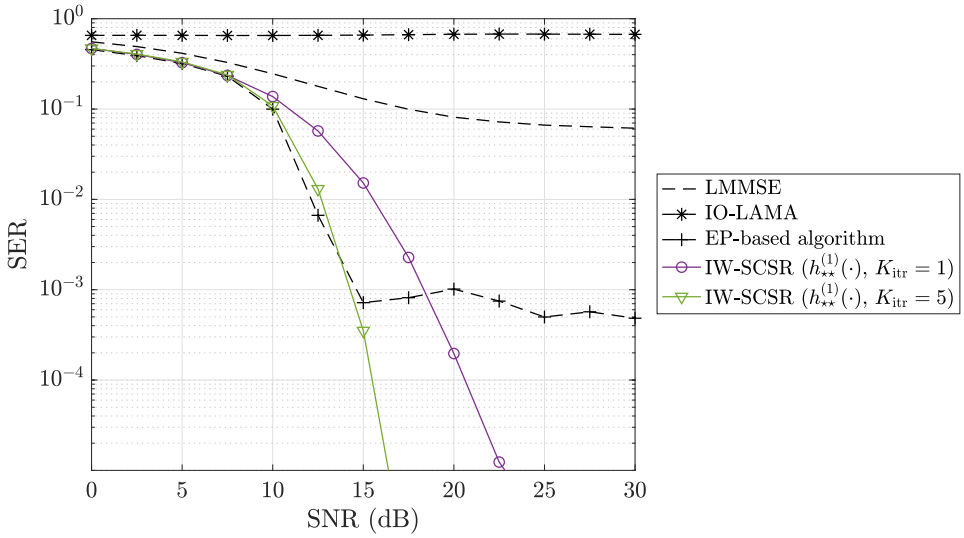


Figure 3.6: SER in correlated MIMO channels (QPSK, $(N, M) = (50, 40)$, $\beta_{\rm r} = 15$)

MIMO channels, respectively. The result is obtained by averaging the SER over more than 3, 000 independent realizations of the measurement matrix hereafter. In the figures, 'IO-LAMA' represents individually-optimal large MIMO AMP (IO-LAMA) [42], which is MIMO signal detection method based on AMP. 'EP-based algorithm' denotes the EP-based method [53] for discrete-valued vector reconstruction. 'IW-SCSR' indicates the proposed method in Algorithm 3.2 and $K_{\rm itr}$ denotes the number of iterations of the outer loop. For IW-SCSR, the parameter is initialized as $q_{n,\ell} = 1/4$, the regularizer

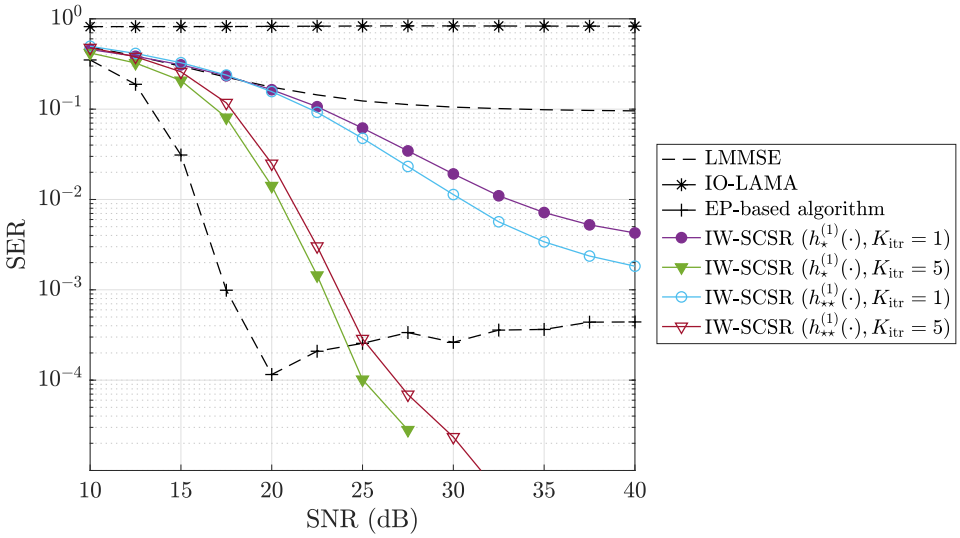


Figure 3.7: SER in correlated MIMO channels (8PSK, $(N, M) = (50, 48)$, $\beta_{\mathrm{r}} = 15$)

is $\tilde{g}_\ell(\cdot) = h_{\star\star}^{(1)}(\cdot)$ $(\ell = 1, \ldots, 4)$, and the parameter is determined from (3.33) with $\beta_{\mathrm{r}} = 15$. From Fig. 3.5, we can see that the performance of IW-SCSR is improved by the weight update and IW-SCSR with $K_{\mathrm{itr}} = 5$ outperforms the other methods in high SNR region. The message passing-based methods, i.e., IO-LAMA and the EP-based algorithm, assume large-scale problems and hence they have severe error floors in this case. Fig. 3.6 shows the SER performance in correlated MIMO channels $\tilde{\boldsymbol{A}} = \boldsymbol{\Psi}_{\mathrm{r}}^{\frac{1}{2}} \tilde{\boldsymbol{A}}_{\mathrm{i.i.d.}} \boldsymbol{\Psi}_{\mathrm{t}}^{\frac{1}{2}}$ described in Section I.2. We set $d_{\mathrm{r}}/\lambda_{\mathrm{w}} = d_{\mathrm{t}}/\lambda_{\mathrm{w}} = 1/2$ in the simulation. In Fig. 3.6, the SER performance of IO-LAMA severely degrades because the algorithm assumes i.i.d. measurement matrix. On the other hand, the assumption is not required for convex optimization-based IW-SCSR and hence the performance does not severely degrade compared to IO-LAMA. Although the EP-based algorithm also works well in the low SNR region, it has the error floor in the high SNR region. We can see that the proposed IW-SCSR can achieve good performance even in correlated channels.

Figure 3.7 shows the SER performance in correlated channels for 8PSK with $C = \left\{ e^{j(\ell-1)\pi/4} \mid \ell = 1, \ldots, 8 \right\}$. Note that the SOAV optimization in the real-valued domain is not appropriate in this case because the real part and the imaginary part are dependent on C. The problem size is $(N, M) = (50, 48)$. The parameter $q_{n,\ell}$ of IW-SCSR is initialized as $q_{n,\ell} = 1/8$. We use (3.33) with $\beta_{\mathrm{r}} = 15$ for the parameter λ in IW-SCSR. In Fig. 3.7, the EP-based algorithm outperforms the other methods in the low SNR region. In the high SNR region, however, the EP-based algorithm has the error floor and IW-SCSR with $K_{\mathrm{itr}} = 5$ can achieve better performance than the EP-based algorithm. In Figs. 3.6 and 3.7, we observe that the proposed IW-SCSR for uniformly distributed unknown

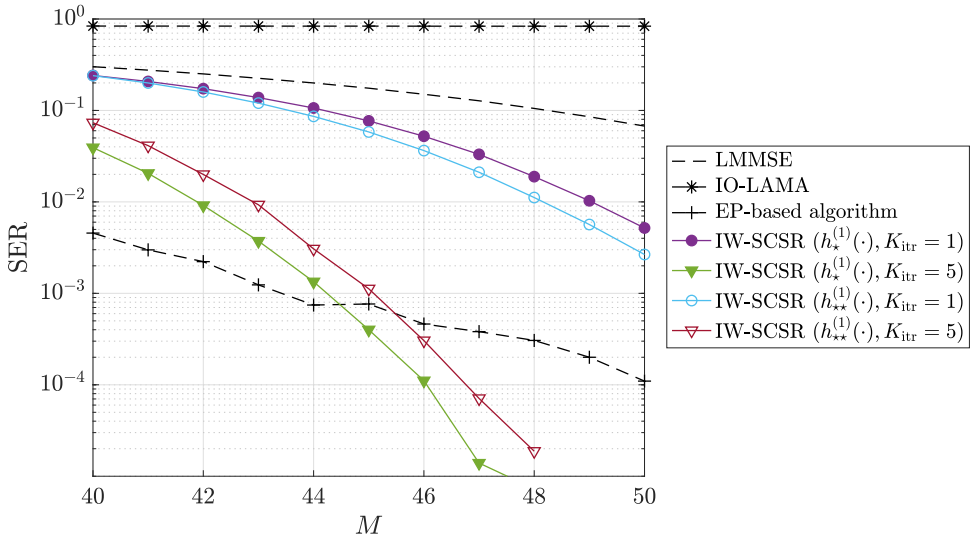


Figure 3.8: SER versus M in correlated MIMO channels (8PSK, $N = 50$, SNR $= 30$ dB, $\beta_{\rm r} = 15$)

vectors is more effective when the cardinality of C is smaller.

In Fig. 3.8, we show the SER performance versus M in correlated MIMO channels with 8PSK, $N = 50$, and SNR $= 30$ dB. The parameters of IW-SCSR are the same as those in Fig. 3.7. From the figure, we can see that the performance of IW-SCSR improves as the number of measurements M increases. When M is greater than 44, IW-SCSR with $h_{\star}^{(1)}(\cdot)$ achieves better performance than the EP-based algorithm.

Figures. 3.9 and 3.10 show the SER performance for $C = \{0, 1+j, -1+j, -1-j, 1-j\}$. The SOAV optimization in the real-valued domain is not suitable in this case as well as in the case of Fig. 3.7. The problem size is $(N, M) = (50, 30)$ in Fig. 3.9 and $(N, M) = (200, 120)$ in Fig. 3.10, and the measurement matrix is correlated as in Figs. 3.6 and 3.7. In the simulation, we assume that $\tilde{\boldsymbol{x}}$ is a discrete-valued sparse vector with $\|\tilde{\boldsymbol{x}}\|_0 = 0.2N$ and the nonzero elements are uniformly distributed on $\{1 + j, -1 + j, -1 - j, 1 - j\}$. The parameter of IW-SCSR is initialized as $q_{n,1} = 0.8$ and $q_{n,2} = \cdots = q_{n,5} = 0.05$. The sparse regularizer $\tilde{g}_\ell(\cdot)$ is set as $g_1(\cdot) = h_{\star}^{(1)}(\cdot)$ and $\tilde{g}_\ell(\cdot) = h_{\star\star}^{(1)}(\cdot)$ $(\ell = 2, \ldots, 5)$ as in Fig. 3.2(a). We denote the ℓ_1 optimization by 'ℓ_1,' which uses only the sparsity and solves

$$\underset{\boldsymbol{s}\in\mathbb{C}^N}{\text{minimize}} \quad \|\boldsymbol{s}\|_1 + \lambda \left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\boldsymbol{s}\right\|_2^2. \tag{3.51}$$

The parameter λ in (3.51) is fixed as the same value in IW-SCSR with $K_{\rm itr} = 1$, which is determined from (3.33) with $\beta_{\rm r} = 10$. In the figures, IW-SCSR with $K_{\rm itr} = 1$ can achieve a bit better performance than the ℓ_1 optimization. We also observe that the performance of IW-SCSR is further improved when $K_{\rm itr} = 5$. Although the EP-based algorithm has

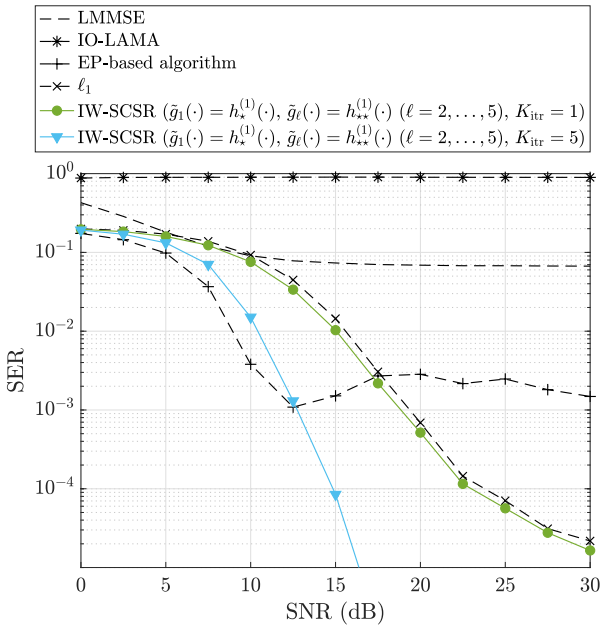


Figure 3.9: SER in correlated MIMO channels ($C = \{0, 1 + j, -1 + j, -1 - j, 1 - j\}$, $(N, M) = (50, 30)$, $\|\tilde{\boldsymbol{x}}\|_0 = 10$, $\beta_{\mathrm{r}} = 10$)

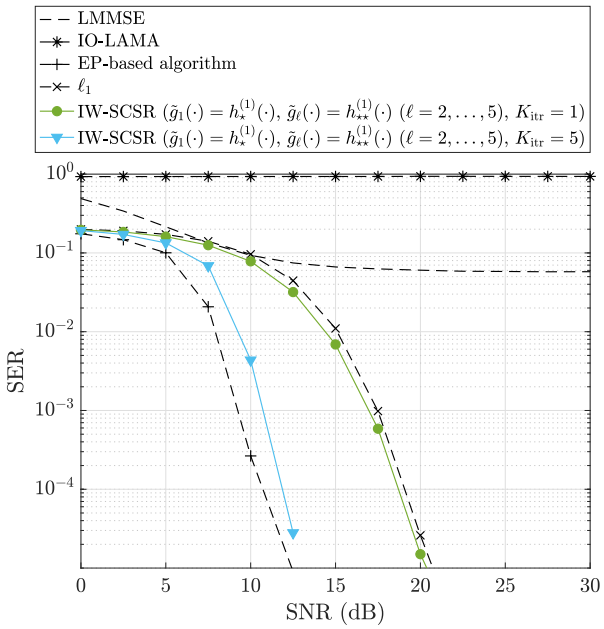


Figure 3.10: SER in correlated MIMO channels ($C = \{0, 1 + j, -1 + j, -1 - j, 1 - j\}$, $(N, M) = (200, 120)$, $\|\tilde{\boldsymbol{x}}\|_0 = 40$, $\beta_{\mathrm{r}} = 10$)

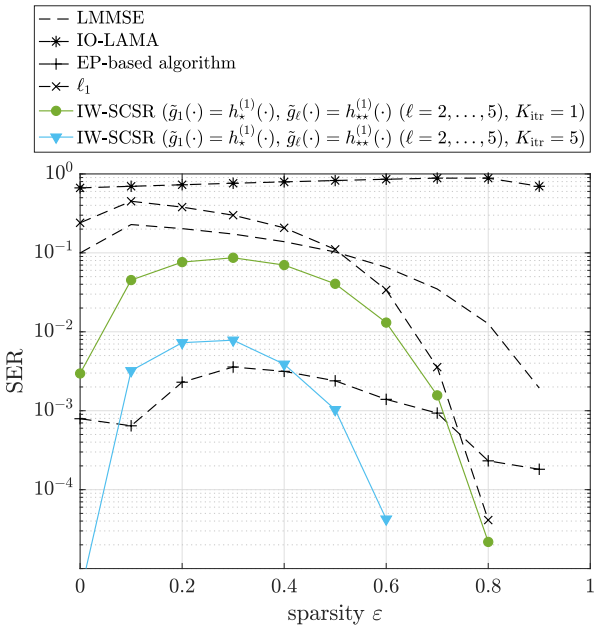


Figure 3.11: SER versus the sparsity ε in correlated MIMO channels ($C = \{0, 1+j, -1+j, -1-j, 1-j\}$, $(N, M) = (50, 40)$, SNR = 17.5 dB, $\beta_{\mathrm{r}} = 10$)

the error floor in Fig. 3.9, it achieves the best performance for the larger-scale problem in Fig. 3.10.

We then show the SER performance versus the sparsity of the unknown discrete-valued vector in Fig. 3.11. We assume $C = \{0, 1+j, -1+j, -1-j, 1-j\}$ and represent the sparsity of the unknown vector by $\varepsilon := 1 - \|\tilde{\boldsymbol{x}}\|_0 / N$. The nonzero elements are assumed to be uniformly distributed on $\{1+j, -1+j, -1-j, 1-j\}$ as in Figs. 3.9 and 3.10. The problem size is $(N, M) = (50, 40)$, the SNR is 17.5 dB, and we define $\beta_{\mathrm{r}} = 10$. The parameter $q_{n,\ell}$ is initialized as $q_{n,1} = \varepsilon$ and $q_{n,2} = \cdots = q_{n,5} = (1-\varepsilon)/4$. In Fig. 3.11, the ℓ_1 optimization has a poor performance for non-sparse vector with small ε, whereas IW-SCSR and the EP-based algorithm have better performance because they use the discreteness of the unknown vector $\boldsymbol{x}$. We can also see that the proposed IW-SCSR outperforms the EP-based algorithm for $\varepsilon \geq 0.5$.

3.3.2 Channel Equalization

In Figs. 3.12 and 3.13, we evaluate the SER performance for channel equalization described in Section 1.2. Unlike MIMO signal detection in flat fading channel, the measurement matrix becomes a block circulant matrix in this problem. In Figs. 3.12 and 3.13, we assume QPSK with $C = \{1+j, -1+j, -1-j, 1-j\}$, $L_{\mathrm{p}} = 5$, and $Q_{\mathrm{b}} = 32$. We also assume $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (4, 3)$ in Fig. 3.12 and $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (8, 6)$ in Fig. 3.13.

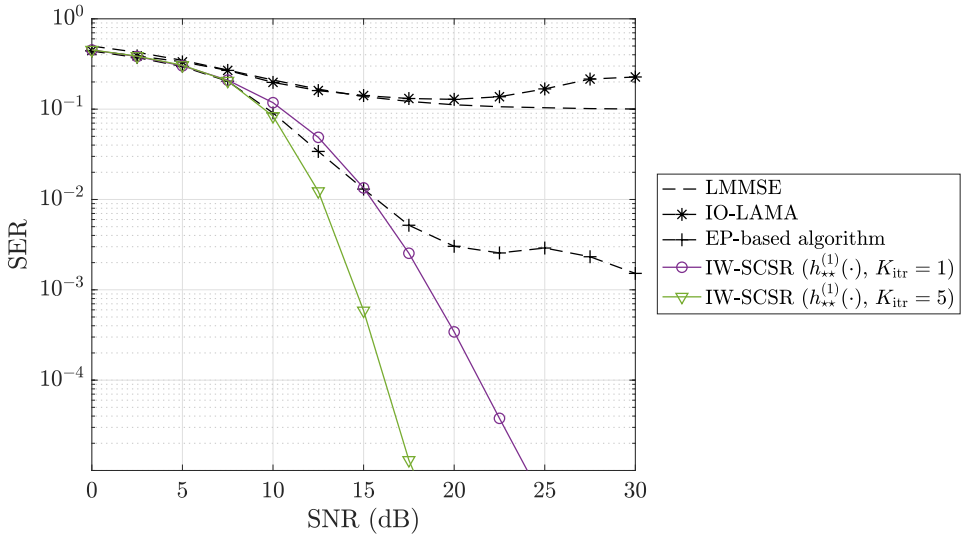


Figure 3.12: SER in channel equalization (QPSK, $(N, M) = (128, 96)$, $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (4, 3)$, $L_{\mathrm{p}} = 5$, $Q_{\mathrm{b}} = 32$, $\beta_{\mathrm{r}} = 15$)

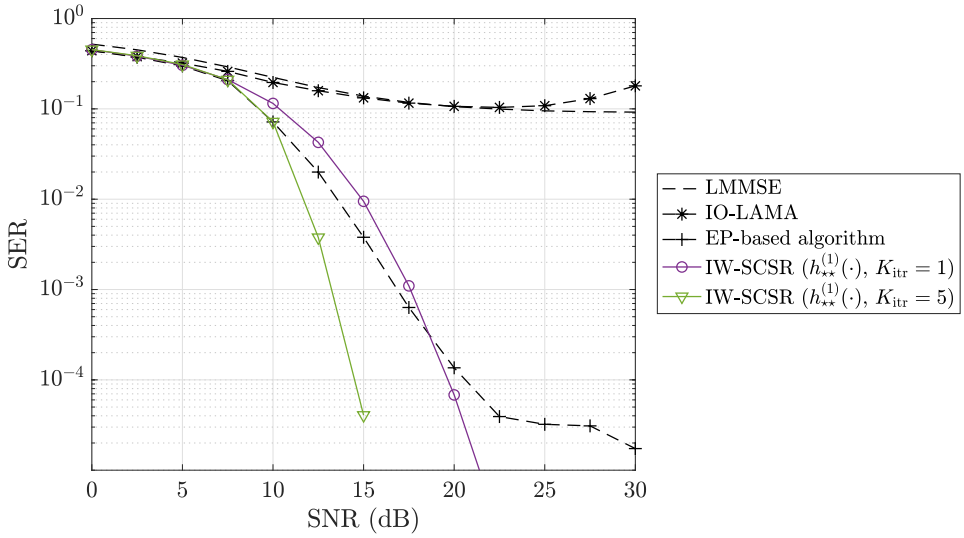


Figure 3.13: SER in channel equalization (QPSK, $(N, M) = (256, 192)$, $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (8, 6)$, $L_{\mathrm{p}} = 5$, $Q_{\mathrm{b}} = 32$, $\beta_{\mathrm{r}} = 15$)

The impulse response $\left\{\gamma^{(i)}_{n_{\mathrm{r}},n_{\mathrm{t}}}\right\}$ $(i = 0, \ldots, L_{\mathrm{p}})$ is composed of i.i.d. circular complex Gaussian variables with zero mean and unit variance. The SNR is here defined as $(L_{\mathrm{p}}N_{\mathrm{t}}/N)\mathrm{E}\left[\|\tilde{\boldsymbol{x}}\|_2^2\right]/\sigma_{\mathrm{v}}^2$. For IW-SCSR, we use the same regularizers and parameters as those in Figs. 3.5 and 3.6. In Figs. 3.12 and 3.13, we observe that the performance of IW-SCSR is better than that of the conventional methods. Unlike in the case of MIMO

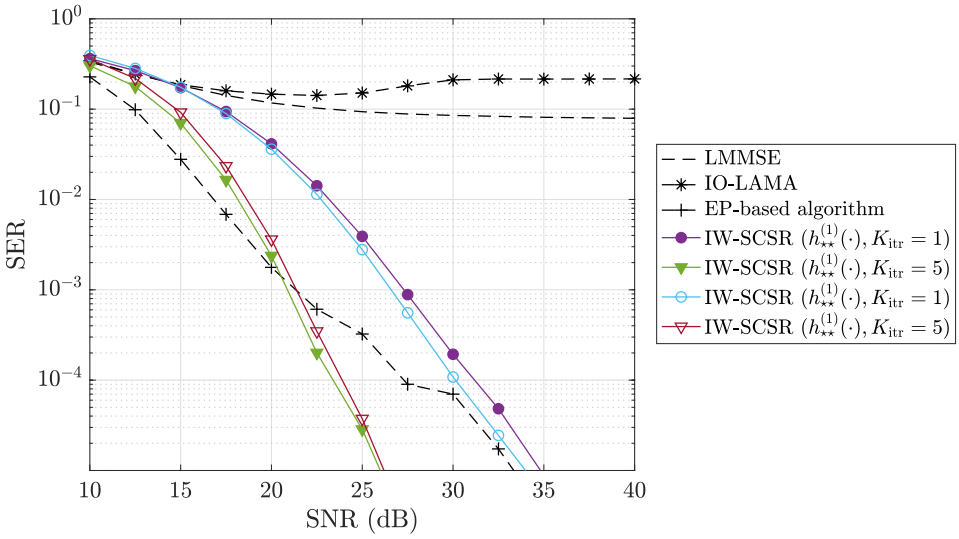


Figure 3.14: SER in channel equalization (8PSK, $(N, M) = (64, 64)$, $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (2, 2)$, $L_{\mathrm{p}} = 5$, $Q_{\mathrm{b}} = 32$, $\beta_{\mathrm{r}} = 15$)

signal detection in Fig. 3.10, the performance of the EP-based algorithm is worse than IW-SCSR even in the larger-scale problem with $(N, M) = (256, 192)$. This is possibly because the measurement matrix in channel equalization has the specific structure as in (1.14).

Figure 3.14 shows the performance for 8PSK with $\mathcal{C} = \left\{e^{j(\ell-1)\pi/4} \mid \ell = 1, \ldots, 8\right\}$, $(N_{\mathrm{t}}, N_{\mathrm{r}}) = (2, 2)$, $L_{\mathrm{p}} = 5$, and $Q_{\mathrm{b}} = 32$. The regularizers and parameters of IW-SCSR is same as those in Fig. 3.7. In the high SNR region, IW-SCSR outperforms the other methods. From Figs. 3.12 and 3.14, we can see that the proposed IW-SCSR achieves good performance in channel equalization as well as in MIMO signal detection.

3.4 Conclusion

In this chapter, we have proposed the SCSR optimization for the reconstruction of complex discrete-valued vector by extending the SOAV optimization in the real-valued domain. The SCSR optimization uses the discreteness of the unknown vector in the complex-valued domain by including the sum of sparse regularizers in the objective function. As the sparse regularizer for complex-valued vectors, we have presented two regularizers $h_{\star}^{(1)}(\cdot)$ and $h_{\star\star}^{(1)}(\cdot)$, which should be appropriately chosen in accordance with the distribution of the unknown vector. We have also proposed the iterative approach named IW-SCSR, which iterates the W-SCSR optimization with updating the parameters in the objective function. We have proved that the sequence obtained by the proposed algorithm converges to the optimal solution of the W-SCSR optimization problem.

Simulation results show that the proposed IW-SCSR works well for the underdetermined discrete-valued vector reconstruction, whereas the conventional message passing-based algorithms have error floors in the high SNR region. The proposed method can also reconstruct the discrete-valued vector even for the correlated measurement matrix, which appears in MIMO signal detection in correlated channels. For discrete-valued sparse vectors, IW-SCSR have better performance than the ℓ_1 optimization, which utilizes only the sparsity of the unknown vector. We have also shown that the proposed IW-SCSR can achieve good performance for channel equalization in frequency-selective fading channels.

Chapter 4

Discrete-Valued Vector Reconstruction by Nonconvex Optimization with Sum of Sparse Regularizers

4.1 Introduction

As introduced in Section 1.3, low-complexity approaches for the discrete-valued vector reconstruction have been proposed on the basis of convex optimization [54, 57, 60]. These optimization problems also take advantage of the discrete nature of the unknown vector as a prior knowledge. However, since all these methods consider convex optimization problems obtained by convex relaxation techniques, the discreteness has not been taken full advantage of.

In this chapter, to obtain better reconstruction performance without any explicit assumption on the measurement matrix, we propose a possibly nonconvex optimization problem named SSR optimization. By using the discreteness of the unknown vector and the idea of compressed sensing [40, 41], we utilize the sum of some sparse regularizers as a regularizer for the discrete-valued vector in the proposed SSR optimization. The SSR optimization can be considered as a generalization of the SOAV optimization, and is equivalent to the SOAV optimization when we use the convex ℓ_1 norm as the sparse regularizer. Other than the ℓ_1 norm, we can also use nonconvex regularizers such as the ℓ_p norm ($0 < p < 1$) [73–76], the ℓ_0 norm, and the $\ell_1 - \ell_2$ difference [77, 78]. For the SSR optimization, we propose an algorithm on the basis of ADMM [61, 68–70], which is known to achieve fast convergence in general, regardless of the convexity of the cost function. However, the ADMM-based algorithm involves the computation of an inverse matrix, which may require prohibitive computational complexity in very large-scale problems. We thus also propose a PDS [79, 80]-based algorithm, which can avoid the computation of the inverse matrix. Moreover, we extend the proposed

approach to the reconstruction of discrete-valued vectors in the complex-valued domain, which commonly emerges in the filed of communications. Simulation results show that the proposed algorithms with nonconvex regularizers can achieve better performance than that with the convex ℓ_1 regularizer, which corresponds to the conventional SOAV optimization.

The rest of this chapter is organized as follows. We propose the SSR optimization problem in Section 4.2 and derive two optimization algorithms in Section 4.3. In Section 4.4, we extend the proposed approach to the reconstruction of complex discrete-valued vector. Section 4.5 gives some simulation results. Finally, we present some conclusions in Section 4.6.

4.2 Proposed SSR Optimization Problem

For the reconstruction of $\boldsymbol{x}$ from $\boldsymbol{y}$ and $\boldsymbol{A}$ in (1.1), we propose the SSR optimization problem

$$\underset{\boldsymbol{s}\in\mathbb{R}^N}{\text{minimize}} \left\{ \sum_{\ell=1}^{L} q_\ell g_\ell\left(\boldsymbol{s} - r_\ell \boldsymbol{1}\right) + \frac{\lambda}{2} \left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\right\|_2^2 \right\}, \tag{4.1}$$

where λ (> 0) is the parameter and $\sum_{\ell=1}^{L} q_\ell = 1$ ($q_\ell \geq 0$). The function $g_\ell(\cdot)$ is a sparse regularizer and we assume that its proximity operator can be computed efficiently. The employment of sparse regularizers in the SSR optimization is based on the fact that the vector $\boldsymbol{x} - r_\ell \boldsymbol{1}$ has some zero elements, which has been utilized in the SOAV optimization [60]. We can thus consider $\sum_{\ell=1}^{L} q_\ell g_\ell\left(\boldsymbol{s} - r_\ell \boldsymbol{1}\right)$ in the objective function as a regularizer for discrete-valued vectors in $\mathcal{R}^N$.

We show some examples of the sparse regularizer $g_\ell(\cdot)$ and the corresponding proximity operators, which are required for the proposed algorithms in Section 4.3. Note that we consider both convex and nonconvex regularizers in this paper, and we can use any sparse regularizer as far as its proximity operator can be computed. Although the proximity operator is usually defined for proper closed convex functions, the minimizer in the definition of the proximity operator can also be obtained formally for the following nonconvex regularizers. We thus use the term 'proximity operator' for both convex and nonconvex functions henceforth.

Example 4.2.1 (ℓ_1 Norm)**.** For the ℓ_1 norm-based convex regularizer $h^{(1)}(\boldsymbol{u}) = \|\boldsymbol{u}\|_1 = \sum_{n=1}^{N} |u_n|$ ($\boldsymbol{u} = [u_1 \ \cdots \ u_N]^{\mathsf{T}} \in \mathbb{R}^N$), the proximity operator $\mathrm{prox}_{\gamma h^{(1)}}(\cdot)$ is given by

$$\left[\mathrm{prox}_{\gamma h^{(1)}}(\boldsymbol{u})\right]_n = \mathrm{sign}(u_n) \max\left(|u_n| - \gamma, 0\right). \tag{4.2}$$

The SSR optimization with the ℓ_1 regularizer is equivalent to the SOAV optimization [60].

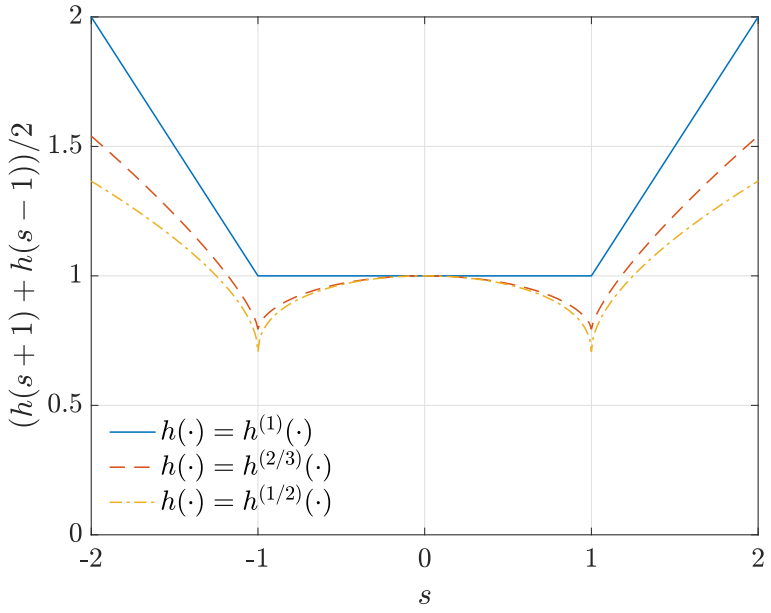


Figure 4.1: $(h(s+1)+h(s-1))/2$

Example 4.2.2 (ℓ_0 Norm)**.** The nonconvex regularizer $h^{(0)}(\boldsymbol{u}) = \|\boldsymbol{u}\|_0$ based on the ℓ_0 norm, i.e., the number of nonzero elements of $\boldsymbol{u}$, has the proximity operator given by

$$\left[\mathrm{prox}_{\gamma h^{(0)}}(\boldsymbol{u})\right]_n = \begin{cases} \{0\} & \left(|u_n| < \sqrt{2\gamma}\right) \\ \{0, u_n\} & \left(|u_n| = \sqrt{2\gamma}\right) \\ \{u_n\} & \left(|u_n| > \sqrt{2\gamma}\right) \end{cases}. \tag{4.3}$$

Example 4.2.3 (ℓ_p Norm $(0 < p < 1)$)**.** We also consider the nonconvex regularizer $h^{(p)}(\boldsymbol{u}) = \|\boldsymbol{u}\|_p^p = \sum_{n=1}^N |u_n|^p$ with the ℓ_p norm $(0 < p < 1)$. In Fig. 4.1, we compare the regularizer $(h^{(p)}(s+1) + h^{(p)}(s-1))/2$ in the binary case with $\mathcal{R} = \{-1, 1\}$ for different values of p. From the figure, we can see that the sums of nonconvex regularizers with $h^{(1/2)}(\cdot)$ and $h^{(2/3)}(\cdot)$ can promote the discrete nature more effectively compared to the convex one with $h^{(1)}(\cdot)$, because the sums of nonconvex regularizers do not have their minimum values for $s \in (-1, 1)$ but only for $s = \pm 1$. The proximity operator of the ℓ_p norm-based regularizers has been discussed in [74–76]. For arbitrary $p \in (0, 1)$, we can numerically compute the proximity operator, while the proximity operator for $p = 1/2, 2/3$ can be written explicitly. Figure 4.2 shows the proximity operators of $\gamma h^{(1)}(\cdot)$, $\gamma h^{(2/3)}(\cdot)$, $\gamma h^{(1/2)}(\cdot)$, and $\gamma h^{(0)}(\cdot)$ $(\gamma = 0.5)$. As we can see from the figure, the proximity operators of the nonconvex regularizers are not continuous.

Example 4.2.4 ($\ell_1 - \ell_2$ Difference)**.** The nonconvex regularizer $h^{(1-2)}(\boldsymbol{u}) = \|\boldsymbol{u}\|_1 - \|\boldsymbol{u}\|_2$ based on the $\ell_1 - \ell_2$ difference has been proposed for compressed sensing [77, 78]. The

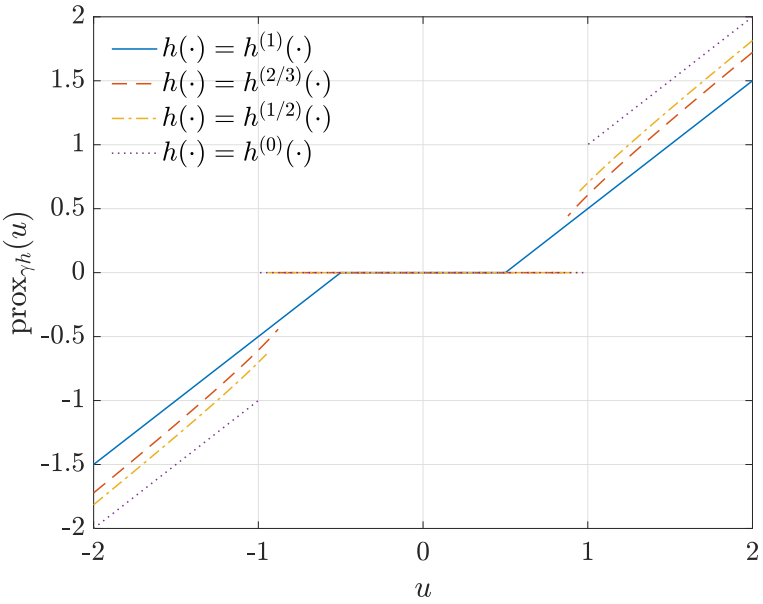


Figure 4.2: $\mathrm{prox}_{\gamma h}(u)$ ($\gamma = 0.5$)

proximity operator of $h^{(1-2)}(\cdot)$ can be computed with Lemma 1 in [78] or Proposition 7.1 in [81].

4.3 Proximal Splitting Algorithms for SSR Optimization

In this section, we propose two algorithms for the SSR optimization. The first one is based on ADMM and the second one is based on PDS.

4.3.1 ADMM-Based Algorithm

We can rewrite the optimization problem of the SSR optimization (4.1) with new variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \in \mathbb{R}^N$ as

$$\begin{aligned} &\underset{\boldsymbol{s},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_L \in \mathbb{R}^N}{\text{minimize}} \left\{ \sum_{\ell=1}^{L} q_\ell g_\ell \left(\boldsymbol{z}_\ell - r_\ell \mathbf{1} \right) + \frac{\lambda}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{s} \right\|_2^2 \right\} \\ &\text{subject to} \quad \boldsymbol{s} = \boldsymbol{z}_\ell \ (\ell = 1, \ldots, L), \end{aligned} \tag{4.4}$$

which is further rewritten as

$$\underset{\boldsymbol{s} \in \mathbb{R}^N, \boldsymbol{z} \in \mathbb{R}^{LN}}{\text{minimize}} \ \{ f(\boldsymbol{s}) + g(\boldsymbol{z}) \} \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{s} = \boldsymbol{z}. \tag{4.5}$$

Here, $\boldsymbol{z} = \left[\boldsymbol{z}_1^\mathsf{T} \cdots \boldsymbol{z}_L^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{LN}$, $\boldsymbol{\Phi} = [\boldsymbol{I}_N \cdots \boldsymbol{I}_N]^\mathsf{T}$, $f(\boldsymbol{s}) = \frac{\lambda}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2$, and $g(\boldsymbol{z}) = \sum_{\ell=1}^{L} q_\ell g_\ell(\boldsymbol{z}_\ell - r_\ell \boldsymbol{1})$.

We derive the proposed algorithm based on ADMM. The update equations of ADMM for (4.5) are given by

$$\boldsymbol{s}^{t+1} = \underset{\boldsymbol{s}\in\mathbb{C}^N}{\arg\min} \left\{ f(\boldsymbol{s}) + \frac{\rho}{2}\left\|\boldsymbol{\Phi}\boldsymbol{s} - \boldsymbol{z}^t + \boldsymbol{w}^t\right\|_2^2 \right\}, \tag{4.6}$$

$$\boldsymbol{z}^{t+1} = \underset{\boldsymbol{z}\in\mathbb{C}^{LN}}{\arg\min} \left\{ g(\boldsymbol{z}) + \frac{\rho}{2}\left\|\boldsymbol{\Phi}\boldsymbol{s}^{t+1} - \boldsymbol{z} + \boldsymbol{w}^t\right\|_2^2 \right\}, \tag{4.7}$$

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \boldsymbol{\Phi}\boldsymbol{s}^{t+1} - \boldsymbol{z}^{t+1}, \tag{4.8}$$

where t is the iteration index, ρ (> 0) is a parameter, and $\boldsymbol{w}^t \in \mathbb{R}^{LN}$. From the equation $\frac{\partial}{\partial \boldsymbol{s}^\mathsf{T}}\left\{f(\boldsymbol{s}) + \frac{\rho}{2}\left\|\boldsymbol{\Phi}\boldsymbol{s} - \boldsymbol{z}^t + \boldsymbol{w}^t\right\|_2^2\right\} = \boldsymbol{0}$, the update of $\boldsymbol{s}^t$ in (4.6) can be written as $\boldsymbol{s}^{t+1} = \left(\rho L \boldsymbol{I}_N + \lambda \boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)^{-1}\left(\rho \sum_{\ell=1}^{L}\left(\boldsymbol{z}_\ell^t - \boldsymbol{w}_\ell^t\right) + \lambda \boldsymbol{A}^\mathsf{T}\boldsymbol{y}\right)$, where $\boldsymbol{z}_\ell^t \in \mathbb{R}^N$ and $\boldsymbol{w}_\ell^t \in \mathbb{R}^N$ ($\ell = 1, \ldots, L$) are subvectors of $\boldsymbol{z}^t = \left[\boldsymbol{z}_1^{t\,\mathsf{T}} \cdots \boldsymbol{z}_L^{t\,\mathsf{T}}\right]^\mathsf{T}$ and $\boldsymbol{w}^t = \left[\boldsymbol{w}_1^{t\,\mathsf{T}} \cdots \boldsymbol{w}_L^{t\,\mathsf{T}}\right]^\mathsf{T}$, respectively. The update of $\boldsymbol{z}^t$ in (4.7) can be written as

$$\boldsymbol{z}^{t+1} = \mathrm{prox}_{\frac{1}{\rho}g}\left(\boldsymbol{\Phi}\boldsymbol{s}^{t+1} + \boldsymbol{w}^t\right) \tag{4.9}$$

$$= \begin{bmatrix} r_1\boldsymbol{1} + \mathrm{prox}_{\frac{q_1}{\rho}g_1}\left(\boldsymbol{s}^{t+1} + \boldsymbol{w}_1^t - r_1\boldsymbol{1}\right) \\ \vdots \\ r_L\boldsymbol{1} + \mathrm{prox}_{\frac{q_L}{\rho}g_L}\left(\boldsymbol{s}^{t+1} + \boldsymbol{w}_L^t - r_L\boldsymbol{1}\right) \end{bmatrix}, \tag{4.10}$$

because the function $g(\cdot)$ is separable as $g(\boldsymbol{z}) = \sum_{\ell=1}^{L} q_\ell g_\ell(\boldsymbol{z}_\ell - r_\ell\boldsymbol{1})$. We also use the property of proximity operator for translation [61] in (4.10).

We summarize the ADMM-based algorithm for the SSR optimization (4.1) as ADMM-SSR in Algorithm 4.1, where $Q(\cdot)$ denotes the element-wise quantization operator which maps the input to its nearest value in $\mathcal{R}$. One of the advantages of ADMM-SSR is that we do not require the proximity operator of the whole regularizer $\sum_{\ell=1}^{L} q_\ell g_\ell(\boldsymbol{s} - r_\ell\boldsymbol{1})$ and we can implement ADMM-SSR as long as the proximity operator of $g_\ell(\cdot)$ can be calculated as in Examples 4.2.1–4.2.4. The computational complexity is dominated by the inverse matrix $\left(\rho L\boldsymbol{I}_N + \lambda\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)^{-1}$, which usually requires $O(N^3)$ complexity [82, Ch. 11].

4.3.2 PDS-Based Algorithm

As we have mentioned in the previous subsection, ADMM-SSR requires the computation of the inverse matrix, which may require prohibitive computational complexity for very

Algorithm 4.1 ADMM-SSR: ADMM-Based Algorithm for (4.1)

Input: $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$
Output: $\hat{\boldsymbol{x}} \in \mathcal{R}^N$

1: Fix $\rho > 0$, $\boldsymbol{z}^0 \in \mathbb{R}^{NL}$, and $\boldsymbol{w}^0 \in \mathbb{R}^{NL}$
2: **for** $t = 0$ to $T_{\text{itr}} - 1$ **do**
3: $\quad \boldsymbol{s}^{t+1} = \left(\rho L \boldsymbol{I}_N + \lambda \boldsymbol{A}^\mathsf{T} \boldsymbol{A}\right)^{-1} \left(\rho \sum_{\ell=1}^{L} \left(\boldsymbol{z}_\ell^t - \boldsymbol{w}_\ell^t\right) + \lambda \boldsymbol{A}^\mathsf{T} \boldsymbol{y}\right)$
4: $\quad \boldsymbol{z}_\ell^{t+1} = r_\ell \mathbf{1} + \text{prox}_{\frac{q_\ell}{\rho} g_\ell}\left(\boldsymbol{s}^{t+1} + \boldsymbol{w}_\ell^t - r_\ell \mathbf{1}\right) \ (\ell = 1, \ldots, L)$
5: $\quad \boldsymbol{w}_\ell^{t+1} = \boldsymbol{w}_\ell^t + \boldsymbol{s}^{t+1} - \boldsymbol{z}_\ell^{t+1} \ (\ell = 1, \ldots, L)$
6: **end for**
7: $\hat{\boldsymbol{x}} = \boldsymbol{Q}(\boldsymbol{s}^{T_{\text{itr}}})$

large-scale problems. To overcome this problem, we also propose an algorithm based on primal-dual splitting [80], which can avoid the computation of the inverse matrix.

We first rewrite the SSR optimization problem (4.1) as

$$\underset{\boldsymbol{s} \in \mathbb{R}^N}{\text{minimize}} \ \{f(\boldsymbol{s}) + g(\boldsymbol{\Phi}\boldsymbol{s})\}, \tag{4.11}$$

where $f(\cdot)$ and $g(\cdot)$ are defined below (4.5). PDS is applicable to the problem of the form (4.11) and is given by

$$\boldsymbol{s}^{t+1} = \boldsymbol{s}^t - \rho_1 \left(\nabla f\left(\boldsymbol{s}^t\right) + \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{w}^t\right), \tag{4.12}$$

$$\boldsymbol{z}^{t+1} = \boldsymbol{w}^t + \rho_2 \boldsymbol{\Phi} \left(2\boldsymbol{s}^{t+1} - \boldsymbol{s}^t\right), \tag{4.13}$$

$$\boldsymbol{w}^{t+1} = \text{prox}_{\rho_2 g^*}\left(\boldsymbol{z}^{t+1}\right), \tag{4.14}$$

where ρ_1, ρ_2 (> 0) are the parameters, $\nabla f(\cdot)$ denotes the gradient of the function $f(\cdot)$, and $g^*(\cdot)$ represents the convex conjugate of $g(\cdot)$. The update of $\boldsymbol{s}^t$ in (4.12) can be written as

$$\boldsymbol{s}^{t+1} = \boldsymbol{s}^t - \rho_1 \left(\lambda \boldsymbol{A}^\mathsf{T} \left(\boldsymbol{A}\boldsymbol{s}^t - \boldsymbol{y}\right) + \sum_{\ell=1}^{L} \boldsymbol{w}_\ell^t\right). \tag{4.15}$$

because $\nabla f(\boldsymbol{s}) = \lambda \boldsymbol{A}^\mathsf{T} (\boldsymbol{A}\boldsymbol{s} - \boldsymbol{y})$. The proximity operator $\text{prox}_{\rho_2 g^*}(\cdot)$ in (4.14) is expressed as $\text{prox}_{\rho_2 g^*}(\boldsymbol{u}) = \boldsymbol{u} - \rho_2 \text{prox}_{g/\rho_2}(\boldsymbol{u}/\rho_2)$. Hence, from (4.10) and (4.14), we can update $\boldsymbol{w}_\ell^{t+1}$ as

$$\boldsymbol{w}_\ell^{t+1} = \boldsymbol{z}_\ell^{t+1} - \rho_2 \left(r_\ell \mathbf{1} + \text{prox}_{\frac{q_\ell}{\rho_2} g_\ell}\left(\frac{\boldsymbol{z}_\ell^{t+1}}{\rho_2} - r_\ell \mathbf{1}\right)\right). \tag{4.16}$$

Algorithm 4.2 PDS-SSR: PDS-Based Algorithm for (4.1)

Input: $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$

Output: $\hat{\boldsymbol{x}} \in \mathcal{R}^N$

1: Fix $\rho_1 > 0$, $\rho_2 > 0$, $\boldsymbol{s}^0 \in \mathbb{R}^N$, and $\boldsymbol{w}^0 \in \mathbb{R}^{NL}$

2: **for** $t = 0$ to $T_{\mathrm{itr}} - 1$ **do**

3: $\quad \boldsymbol{s}^{t+1} = \boldsymbol{s}^t - \rho_1 \left(\lambda \boldsymbol{A}^{\mathsf{T}} \left(\boldsymbol{A}\boldsymbol{s}^t - \boldsymbol{y} \right) + \sum_{\ell=1}^{L} \boldsymbol{w}_\ell^t \right)$

4: $\quad \boldsymbol{z}_\ell^{t+1} = \boldsymbol{w}_\ell^t + \rho_2 \left(2\boldsymbol{s}^{t+1} - \boldsymbol{s}^t \right) \ (\ell = 1, \ldots, L)$

5: $\quad \boldsymbol{w}_\ell^{t+1} = \boldsymbol{z}_\ell^{t+1} - \rho_2 \left(r_\ell \boldsymbol{1} + \mathrm{prox}_{\frac{q_\ell}{\rho_2} g_\ell} \left(\frac{\boldsymbol{z}_\ell^{t+1}}{\rho_2} - r_\ell \boldsymbol{1} \right) \right) \ (\ell = 1, \ldots, L)$

6: **end for**

7: $\hat{\boldsymbol{x}} = \boldsymbol{Q}(\boldsymbol{s}^{T_{\mathrm{itr}}})$

We summarize the PDS-based algorithm named PDS-SSR in Algorithm 4.2. As is the case with ADMM-SSR, PDS-SSR also requires only the proximity operator of $g_\ell(\cdot)$. Since PDS-SSR computes only the addition of vectors and the multiplication of a matrix and a vector, it requires $O(MN)$ complexity [82, Ch. 6], which is lower than that of ADMM-SSR.

4.3.3 Convergence of Proposed Algorithms

The convergence of the proposed algorithms depends on the convexity of the sparse regularizer $g_\ell(\cdot)$. When $g_1(\cdot), \ldots, g_L(\cdot)$ are all convex, the objective function of the SSR optimization is also convex. In this case, the sequence $\{\boldsymbol{s}^t\}$ obtained by ADMM-SSR converges to the optimizer of the problem from the general result for ADMM [69]. From Theorem 3.1 in [80], the sequence $\{\boldsymbol{s}^t\}$ obtained by PDS-SSR also converges if the parameters ρ_1 and ρ_2 satisfy $1/\rho_1 - \rho_2 L \geq \lambda \left\| \boldsymbol{A}^{\mathsf{T}} \boldsymbol{A} \right\|_2 / 2$. When $g_\ell(\cdot)$ is nonconvex, however, the convergence to the global optimizer is not guaranteed in general. Although some convergence property have been proved under several assumptions [83–87], their results cannot be directly used for the proposed algorithms.

4.4 Extension to Complex-Valued Case

In this section, we extend the proposed method to the reconstruction of the complex-valued vector $\tilde{\boldsymbol{x}} \in C^N \subset \mathbb{C}^N$ as described in Section 1.1.2.

For the reconstruction of the complex discrete-valued vector, we extend the SSR

optimization (4.1) to the problem

$$\underset{\boldsymbol{s}\in\mathbb{C}^N}{\text{minimize}} \left\{ \sum_{\ell=1}^{L} q_\ell \tilde{g}_\ell \left(\boldsymbol{s} - c_\ell \boldsymbol{1} \right) + \lambda \left\| \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}} \boldsymbol{s} \right\|_2^2 \right\}, \tag{4.17}$$

which is referred to as the SCSR optimization hereafter. The function $\tilde{g}_\ell(\cdot)$ is a sparse regularizer for the complex-valued sparse vector. The SCSR optimization with the ℓ_1 regularizer has been proposed in Chapter 3, whereas we consider nonconvex regularizers as well in this paper. As discussed in Chapter 3, the optimization in the complex-valued domain is more suitable than that in the real-valued domain when the real part and the imaginary part of the unknown vector are not independent.

For the SCSR optimization (4.17), we newly consider two kinds of sparse regularizers as the candidates of $\tilde{g}_\ell(\cdot)$. For example, as the regularizers based on the ℓ_p norm, we define $\tilde{h}_\star^{(p)}(\tilde{\boldsymbol{u}}) = \sum_{n=1}^N |\tilde{u}_n|^p$ and $\tilde{h}_{\star\star}^{(p)}(\tilde{\boldsymbol{u}}) = \sum_{n=1}^N \left(|\mathrm{Re}\{\tilde{u}_n\}|^p + |\mathrm{Im}\{\tilde{u}_n\}|^p \right)$, where $\tilde{\boldsymbol{u}} = [\tilde{u}_1 \ \cdots \ \tilde{u}_N]^\mathsf{T} \in \mathbb{C}^N$. The first regularizer $\tilde{h}_\star^{(p)}(\cdot)$ is based on the modulus for complex numbers, whereas the second one $\tilde{h}_{\star\star}^{(p)}(\cdot)$ treats the real part and the imaginary part independently. We also define $h_\star^{(1)}(\cdot)$, $h_\star^{(0)}(\cdot)$, $h_\star^{(1-2)}(\cdot)$, $h_{\star\star}^{(1)}(\cdot)$, $h_{\star\star}^{(0)}(\cdot)$, and $h_{\star\star}^{(1-2)}(\cdot)$ in the same manner. The proximity operator of $\gamma \tilde{h}_\star(\cdot)$ ($\tilde{h}_\star(\cdot) = \tilde{h}_\star^{(1)}(\cdot)$, $\tilde{h}_\star^{(0)}(\cdot)$, $\tilde{h}_\star^{(p)}(\cdot)$, $\tilde{h}_\star^{(1-2)}(\cdot)$) in the complex-valued domain can be written with that of the corresponding regularizer $\gamma h(\cdot)$ ($h(\cdot) = h^{(1)}(\cdot)$, $h^{(0)}(\cdot)$, $h^{(p)}(\cdot)$, $h^{(1-2)}(\cdot)$, respectively) in the real-valued domain. Note that $\tilde{h}_\star(\cdot)$ satisfy $\tilde{h}_\star(\tilde{\boldsymbol{u}}) = h(|\tilde{\boldsymbol{u}}|)$, where we define $|\tilde{\boldsymbol{u}}| = [|\tilde{u}_1| \ \cdots \ |\tilde{u}_N|]^\mathsf{T}$. By using this property, the proximity operator of $\gamma \tilde{h}_\star(\cdot)$ can be derived as

$$\left[\mathrm{prox}_{\gamma \tilde{h}_\star}(\tilde{\boldsymbol{u}}) \right]_n = \left[\mathrm{prox}_{\gamma h}(|\tilde{\boldsymbol{u}}|) \right]_n \frac{\tilde{u}_n}{|\tilde{u}_n|} \tag{4.18}$$

with a simple manipulation. The proximity operator of $\gamma \tilde{h}_{\star\star}(\cdot)$ can also be written with the corresponding proximity operator $\mathrm{prox}_{\gamma h}(\cdot)$. Since we have $\tilde{h}_{\star\star}(\tilde{\boldsymbol{u}}) = h(\boldsymbol{u}_\mathrm{R}) + h(\boldsymbol{u}_\mathrm{I})$ from the definition, the proximity operator can be written as

$$\left[\mathrm{prox}_{\gamma \tilde{h}_{\star\star}}(\tilde{\boldsymbol{u}}) \right]_n = \left[\mathrm{prox}_{\gamma h}(\boldsymbol{u}_\mathrm{R}) \right]_n + j \cdot \left[\mathrm{prox}_{\gamma h}(\boldsymbol{u}_\mathrm{I}) \right]_n, \tag{4.19}$$

by using a similar approach to (3.25), where $\boldsymbol{u}_\mathrm{R} = \mathrm{Re}\{\tilde{\boldsymbol{u}}\}$ and $\boldsymbol{u}_\mathrm{I} = \mathrm{Im}\{\tilde{\boldsymbol{u}}\}$.

Since ADMM with complex-valued variables have been discussed in [71] and Chapter 3, we propose the ADMM-based algorithm for the SCSR optimization (4.17) by using the approach in Chapter 3. The resultant algorithm is obtained by replacing $\mathbb{R}$, $(\cdot)^\mathsf{T}$, r_ℓ, and $\mathrm{prox}_{\frac{q_\ell}{\rho} g_\ell}(\cdot)$ in Algorithm 4.1 with $\mathbb{C}$, $(\cdot)^\mathsf{H}$, c_ℓ, and $\mathrm{prox}_{\frac{q_\ell}{2\rho} \tilde{g}_\ell}(\cdot)$, respectively.

4.5 Simulation Results

In this section, we evaluate the performance of the proposed algorithms. In the simulation, the measurement matrix is composed of i.i.d. Gaussian variables with zero mean

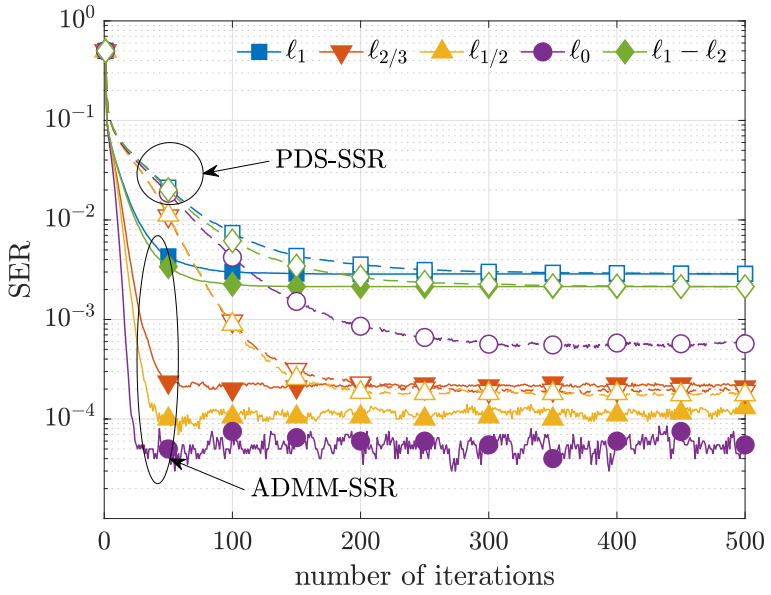


Figure 4.3: SER versus number of iterations for binary vectors ($(N, M) = (200, 160)$, SNR $= 15$ dB)

and unit variance. We also assume the zero mean additive white Gaussian noise. The initialization is given by $\boldsymbol{z}^0 = \boldsymbol{w}^0 = \boldsymbol{0}$ in ADMM-SSR and $\boldsymbol{s}^0 = \boldsymbol{0}, \boldsymbol{w}^0 = \boldsymbol{0}$ in PDS-SSR. Other parameters such as λ and ρ are chosen to achieve good performance in the simulation.

Figure 4.3 shows the SER versus number of iterations for the binary vector with $(r_1, r_2) = (-1, 1)$ and $(p_1, p_2) = (1/2, 1/2)$. The result is obtained by averaging the SER over $2,000$ independent realizations of the measurement matrix. The problem size is $(N, M) = (200, 160)$ and the SNR is 15 dB. The parameters are set as $q_1 = q_2 = 1/2$, $\lambda = 0.05$, $\rho = 3$, $\rho_1 = 2/\left(\lambda \left\|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\right\|_2 + 4\right)$, and $\rho_2 = 1/2$. In Fig. 4.3, we denote the sparse regularizers based on the ℓ_1 norm, the $\ell_{2/3}$ norm, the $\ell_{1/2}$ norm, the ℓ_0 norm, and the $\ell_1 - \ell_2$ difference by 'ℓ_1,' '$\ell_{2/3}$,' '$\ell_{1/2}$,' 'ℓ_0,' and '$\ell_1 - \ell_2$,' respectively. We can see that ADMM-SSR and PDS-SSR converge to the same SER when we use the convex ℓ_1 regularizer. The proposed algorithms with nonconvex regularizers, especially with the ℓ_p and ℓ_0 norms, can achieve much better SER performance.

In Fig. 4.4, we show the SER of ADMM-SSR versus SNR for the binary vector reconstruction with $(N, M) = (200, 150)$. For comparison, we also show the performance of the LMMSE and the box relaxation method [54] as 'LMMSE' and 'Box,' respectively. The parameters in ADMM-SSR are set as $\lambda = 0.05$, $\rho = 3$, and $K_{\text{itr}} = 300$. The nonconvex regularizers can outperform the convex ℓ_1 regularizer and the box relaxation method.

Figure 4.5 shows the SER versus SNR for the reconstruction of complex discrete-

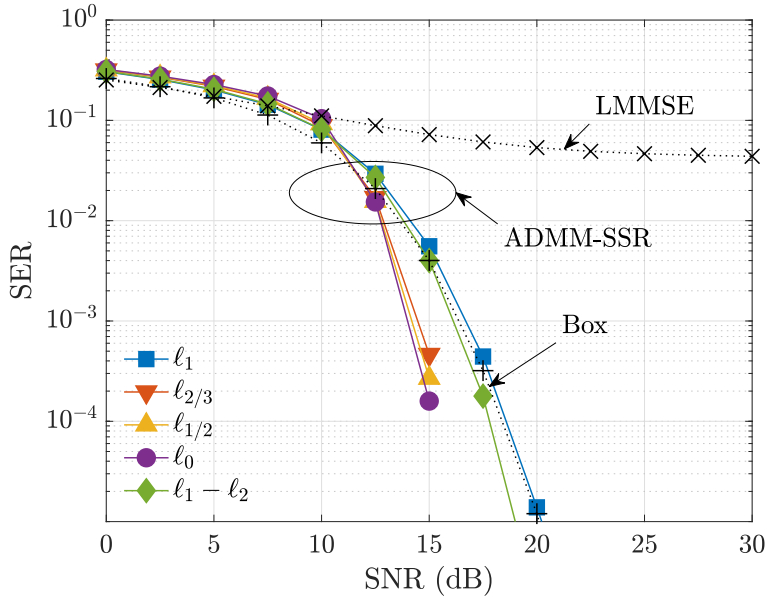


Figure 4.4: SER versus SNR for binary vectors($(N, M) = (200, 150)$)

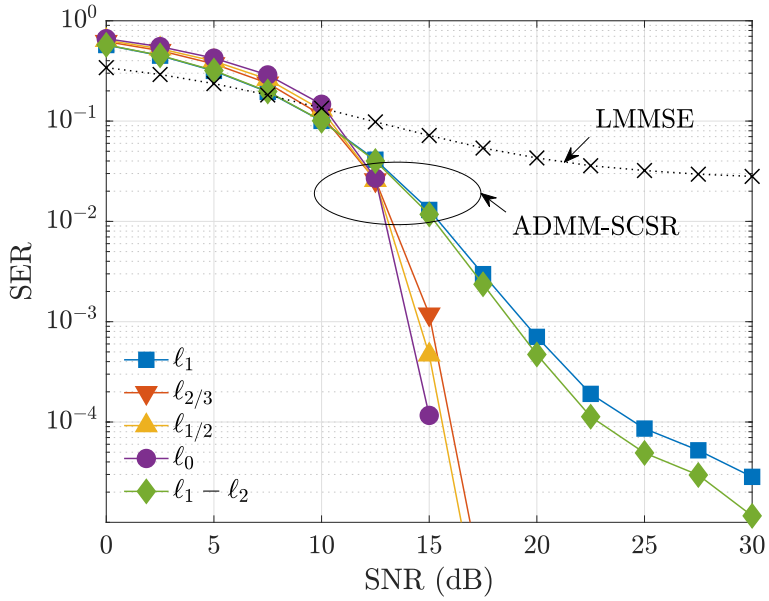


Figure 4.5: SER versus SNR for complex-valued vectors ($(N, M) = (200, 160)$)

valued vectors with $(N, M) = (200, 160)$. The distribution of the unknown complex variable is given by $(c_1, c_2, c_3, c_4, c_5) = (0, 1+j, -1+j, -1-j, 1-j)$ and $(p_1, p_2, p_3, p_4, p_5) = (0.6, 0.1, 0.1, 0.1, 0.1)$, where $p_\ell = \Pr(x_n = c_\ell)$ $(\ell = 1, \ldots, 5)$. We use the sparse regularizers given by $\tilde{g}_1(\cdot) = \tilde{h}_\star(\cdot)$ and $\tilde{g}_\ell(\cdot) = \tilde{h}_{\star\star}(\cdot)$ $(\ell = 2, \ldots, 5)$ because the real part

becomes zero only when the imaginary part is zero in this case. The parameters of the proposed algorithm are set as $q_\ell = p_\ell$, $\lambda = 0.05$, $\rho = 3$, and $K_{\mathrm{itr}} = 300$. From the figure, we can see that the proposed approach with nonconvex regularizer can achieve good performance even for the reconstruction of the complex discrete-valued vector.

4.6 Conclusion

In this chapter, we have proposed possibly nonconvex optimization problems for the discrete-valued vector reconstruction in both real- and complex-valued cases. The proposed method utilizes the sum of sparse regularizers as the regularizer for the discrete-valued vector. Simulation results show that the proposed algorithms with nonconvex regularizers can achieve better performance than that with the convex ℓ_1 regularizer.

Chapter 5

Asymptotic Performance Analysis of Discrete-Valued Vector Reconstruction with Sum of ℓ_1 Regularizers

5.1 Introduction

For large-scale discrete-valued vector reconstruction, some convex optimization-based methods have been proposed to obtain good performance with reasonable computational complexity. Box relaxation method [54, 55] considers the ML method under the hypercube including all possible discrete-valued vectors. Regularization-based method and transform-based method [57] apply the idea of compressed sensing [40, 41] to discrete-valued vector reconstruction. SOAV optimization [60] takes a similar approach and uses a weighted sum of ℓ_1 regularizers as a regularizer for the discrete-valued vector. One of advantages of the SOAV optimization over the other convex optimization-based methods is that it can take the probability distribution of unknown variables into consideration. The SOAV optimization has been applied to various practical problems [14, 26, 29, 88, 89], whereas only a few theoretical aspects of the performance are known in the literature.

In this chapter, we analyze the asymptotic performance of discrete-valued vector reconstruction based on the SOAV optimization. To make the analysis simpler, we firstly modify the conventional SOAV optimization into the Box-SOAV optimization by using the boundedness of the unknown vector. We then investigate the performance of Box-SOAV by using CGMT [90, 91], which has been used for the performance analyses of several convex optimization problems. We provide the asymptotic SER of Box-SOAV in the large system limit with a fixed measurement ratio, which is defined as the ratio of the number of unknown variables to the number of measurements. The asymptotic SER is characterized by the probability distribution of the unknown vector, the measurement ratio, the parameters of Box-SOAV, and the optimizer of a scalar

optimization problem. The result enables us to predict the performance of Box-SOAV in the large-scale discrete-valued vector reconstruction. We also derive the asymptotic distribution of the estimate obtained by the Box-SOAV optimization. By using the asymptotic distribution, we can optimize the quantizer for the hard decision of the unknown discrete-valued vector in terms of the asymptotic SER. Moreover, we propose a procedure to choose the parameter value of the Box-SOAV optimization to minimize the asymptotic SER. Simulation results show that the empirical SER performance of the Box-SOAV optimization and the conventional SOAV optimization agrees well with the theoretical result for Box-SOAV in large-scale problems. From the results, we can also see that the proposed asymptotically optimal parameters and quantizer can achieve better performance compared to the case with some fixed parameter and a naive quantizer.

The rest of this chapter is organized as follows. In Section 5.2, we describe the CGMT. We then provide the main analytical results for the Box-SOAV optimization in Section 5.3. The proof for the main theorem is given in Section 5.4. Section 5.5 gives some simulation results, which demonstrate the validity of the theoretical analysis for Box-SOAV. Finally, Section 5.6 presents some conclusions.

5.2 CGMT

CGMT is a theorem that associates the primary optimization (PO) problem with the auxiliary optimization (AO) problem given by

$$\text{(PO)}: \Phi(\boldsymbol{G}) = \min_{\boldsymbol{w}\in\mathcal{S}_{\mathrm{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\mathrm{u}}} \left\{\boldsymbol{u}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{w} + \rho(\boldsymbol{w},\boldsymbol{u})\right\}, \tag{5.1}$$

$$\text{(AO)}: \phi(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}_{\mathrm{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\mathrm{u}}} \left\{\|\boldsymbol{w}\|_2\, \boldsymbol{g}^{\mathsf{T}}\boldsymbol{u} - \|\boldsymbol{u}\|_2\, \boldsymbol{h}^{\mathsf{T}}\boldsymbol{w} + \rho(\boldsymbol{w},\boldsymbol{u})\right\}, \tag{5.2}$$

respectively, where $\boldsymbol{G} \in \mathbb{R}^{M\times N}$, $\boldsymbol{g} \in \mathbb{R}^{M}$, $\boldsymbol{h} \in \mathbb{R}^{N}$, $\mathcal{S}_{\mathrm{w}} \subset \mathbb{R}^{N}$, $\mathcal{S}_{\mathrm{u}} \subset \mathbb{R}^{M}$, and $\rho(\cdot,\cdot) : \mathbb{R}^{N}\times\mathbb{R}^{M} \to \mathbb{R}$. We assume that $\mathcal{S}_{\mathrm{w}}$ and $\mathcal{S}_{\mathrm{u}}$ are closed compact sets, and $\rho(\cdot,\cdot)$ is a continuous convex-concave function on $\mathcal{S}_{\mathrm{w}}\times\mathcal{S}_{\mathrm{u}}$. The elements of $\boldsymbol{G}$, $\boldsymbol{g}$, and $\boldsymbol{h}$ are i.i.d. standard Gaussian random variables. The following theorem relates the optimizer $\hat{\boldsymbol{w}}_{\Phi}(\boldsymbol{G})$ of (PO) with the optimal value of (AO) in the limit of $M, N \to \infty$ with a fixed ratio $\Delta = M/N$, which we simply denote $N \to \infty$ in this paper.

Theorem 5.2.1 (CGMT [56])**.** Let $\mathcal{S}$ be an open set in $\mathcal{S}_{\mathrm{w}}$ and $\mathcal{S}^{\mathrm{c}} = \mathcal{S}_{\mathrm{w}} \setminus \mathcal{S}$. Also, let $\phi_{\mathcal{S}^{\mathrm{c}}}(\boldsymbol{g},\boldsymbol{h})$ be the optimal value of (AO) with the constraint $\boldsymbol{w} \in \mathcal{S}^{\mathrm{c}}$. If there are constants $\eta > 0$ and $\bar{\phi}$ satisfying (i) $\phi(\boldsymbol{g},\boldsymbol{h}) \le \bar{\phi} + \eta$ and (ii) $\phi_{\mathcal{S}^{\mathrm{c}}}(\boldsymbol{g},\boldsymbol{h}) \ge \bar{\phi} + 2\eta$ with probability approaching 1, then we have $\lim_{N\to\infty} \Pr\left(\hat{\boldsymbol{w}}_{\Phi}(\boldsymbol{G}) \in \mathcal{S}\right) = 1$.

CGMT has been applied to the performance analyses of various optimization problems. The asymptotic normalized squared error (NSE) and mean squared error (MSE) have been analyzed for various regularized estimators [90–94]. The asymptotic SER

of the box relaxation method has been derived for BPSK signals in [95], and the result has been generalized for pulse amplitude modulation (PAM) in [56]. CGMT has also been used for the analysis of nonlinear measurement model [96]. A similar result has been obtained for MIMO signal detection with low resolution analog-to-digital converter (ADC), where the receiver has the quantized measurements [97]. Moreover, CGMT can be applied to the case when the measurement matrix is not perfectly known and includes Gaussian distributed errors [98]. In [99], the technique has been used to derive the asymptotically optimal power allocation between the pilots and the data in MIMO BPSK transmission. In [100], CGMT-based analysis has been applied to an optimization problem in the complex-valued domain under some assumptions, while above approaches consider optimization problems in the real-valued domain.

5.3 Main Results

In this section, we provide the main results of this paper. In Section 5.3.1, we modify the conventional SOAV optimization into the Box-SOAV optimization to make the analysis simpler. In Section 5.3.2, we derive the asymptotic SER of the estimate obtained by the Box-SOAV optimization. We then characterize the distribution of the estimate in Section 5.3.3. By using the results, we also derive the asymptotically optimal quantizer for the estimate in Section 5.3.4. Finally, we propose a parameter selection method for the Box-SOAV optimization in Section 5.3.5.

5.3.1 Box-SOAV Optimization

To make the analysis simpler, we newly consider the Box-SOAV optimization given by

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{s}\in[r_1,r_L]^N}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 + \sum_{\ell=1}^{L} q_\ell \|\boldsymbol{s} - r_\ell \boldsymbol{1}\|_1 \right\} \tag{5.3}$$

$$= \underset{\boldsymbol{s}\in\mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 + \sum_{\ell=1}^{L} q_\ell \|\boldsymbol{s} - r_\ell \boldsymbol{1}\|_1 + \mathcal{I}(\boldsymbol{s}) \right\}, \tag{5.4}$$

where the function $\mathcal{I}(\cdot)$ denotes the indicator function given by

$$\mathcal{I}(\boldsymbol{s}) = \begin{cases} 0 & (\boldsymbol{s} \in [r_1, r_L]^N) \\ \infty & (\boldsymbol{s} \notin [r_1, r_L]^N) \end{cases}. \tag{5.5}$$

This modification is reasonable because $\boldsymbol{x} \in [r_1, r_L]^N$ and it does not change the value of the objective function for $\boldsymbol{s} \in [r_1, r_L]^N$. Let $f(\boldsymbol{s}) = \sum_{\ell=1}^{L} q_\ell \|\boldsymbol{s} - r_\ell \boldsymbol{1}\|_1 + \mathcal{I}(\boldsymbol{s})$, where $f(\cdot)$ is an element-wise function and we use the same notation $f(\cdot)$ for the corresponding

scalar function hereafter. By modifying the result in [26], the proximity operator $\mathrm{prox}_{\gamma f}(z)$ ($\gamma \geq 0$) can be obtained as

$$\mathrm{prox}_{\gamma f}(z) = \begin{cases} r_1 & (z < r_1 + \gamma Q_2) \\ \quad \vdots & \quad \vdots \\ z - \gamma Q_k & (r_{k-1} + \gamma Q_k \leq z < r_k + \gamma Q_k) \\ r_k & (r_k + \gamma Q_k \leq z < r_k + \gamma Q_{k+1}) \\ \quad \vdots & \quad \vdots \\ r_L & (r_L + \gamma Q_L \leq z) \end{cases}, \tag{5.6}$$

where $Q_k = \left(\sum_{\ell=1}^{k-1} q_\ell\right) - \left(\sum_{\ell'=k}^{L} q_{\ell'}\right)$ ($k = 2, \ldots, L$). By using some proximal splitting algorithm [61] with the proximity operator in (5.6), we can obtain a sequence converging to the solution of the Box-SOAV optimization (5.4).

5.3.2 Asymptotic SER of Box-SOAV

To provide the asymptotic SER of Box-SOAV, we firstly show the following theorem.

Theorem 5.3.1. The measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is assumed to be composed of i.i.d. Gaussian random variables with zero mean and variance $1/N$. The distribution of the noise vector $\boldsymbol{v} \in \mathbb{R}^M$ is also assumed to be Gaussian with mean $\boldsymbol{0}$ and covariance matrix $\sigma_{\mathrm{v}}^2 \boldsymbol{I}$. We also assume that the optimization problem $\max_{\beta>0} \min_{\alpha>0} F(\alpha, \beta)$ has a unique optimizer (α^*, β^*), where

$$F(\alpha, \beta) = \frac{\alpha\beta\sqrt{\Delta}}{2} + \frac{\sigma_{\mathrm{v}}^2 \beta\sqrt{\Delta}}{2\alpha} - \frac{1}{2}\beta^2 - \frac{\alpha\beta}{2\sqrt{\Delta}} + \frac{\beta\sqrt{\Delta}}{\alpha} \mathrm{E}\left[\mathrm{env}_{\frac{\alpha}{\beta\sqrt{\Delta}} f}\left(X + \frac{\alpha}{\sqrt{\Delta}} H\right)\right]. \tag{5.7}$$

Here, X is the random variable with the same distribution as the unknown variables, i.e., $\Pr(X = r_\ell) = p_\ell$. H is the standard Gaussian random variable independent of X. We further define

$$\begin{aligned} \mathcal{L} = \{\psi(\cdot, \cdot) : [r_1 - r_L, r_L - r_1] \times \mathcal{R} \to \mathbb{R} \mid \\ \psi(\cdot, r_\ell) \text{ is Lipschitz continuous for any } r_\ell \in \mathcal{R}\}. \end{aligned} \tag{5.8}$$

For any function $\psi(\cdot, \cdot) \in \mathcal{L}$, we have

$$\operatorname*{plim}_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \psi\left(\hat{x}_n - x_n, x_n\right) = \mathrm{E}\left[\psi\left(\hat{X} - X, X\right)\right], \tag{5.9}$$

where $\hat{x}_n$ denotes the nth element of $\hat{\boldsymbol{x}}$ in (5.4) and

$$\hat{X} = \mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}\left(X + \frac{\alpha^*}{\sqrt{\Delta}} H\right). \tag{5.10}$$

Proof. See Section 5.4. □

The SER of Box-SOAV is given by $\frac{1}{N}\|Q(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|_0$, where the element-wise quantizer $Q(\cdot)$ maps each element of the vector to a value in $\mathcal{R}$. The asymptotic SER of Box-SOAV is given by the following corollary of Theorem 5.3.1.

Corollary 5.3.1. Under the assumptions in Theorem 5.3.1, the asymptotic SER of Box-SOAV is given by

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\|Q(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|_0 = 1 - \sum_{\ell=1}^{L} p_\ell \Pr\left(Q(\hat{X}) = r_\ell \mid X = r_\ell\right). \tag{5.11}$$

Proof. See Appendix 5.A. □

The function $F(\alpha, \beta)$ in (5.7) and the asymptotic SER in (5.11) can be calculated by using the PDF $p_{\mathrm{G}}(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$ and the CDF $P_{\mathrm{G}}(z) = \int_{-\infty}^{z} p_{\mathrm{G}}(z')dz'$ of the standard Gaussian distribution. For example, when we use the quantizer $Q_{\mathrm{NV}}(\cdot)$ that maps the input to the nearest value in $\mathcal{R}$, i.e.,

$$Q_{\mathrm{NV}}(\hat{x}) = \begin{cases} r_1 & \left(\hat{x} < \dfrac{r_1 + r_2}{2}\right) \\ \vdots & \vdots \\ r_\ell & \left(\dfrac{r_{\ell-1} + r_\ell}{2} \le \hat{x} < \dfrac{r_\ell + r_{\ell+1}}{2}\right) \\ \vdots & \vdots \\ r_L & \left(\dfrac{r_{L-1} + r_L}{2} \le \hat{x}\right) \end{cases}, \tag{5.12}$$

the asymptotic SER in (5.11) can be written as

$$\mathrm{SER}_{\mathrm{NV}} = 1 - \sum_{\ell=1}^{L} p_\ell \left\{ P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{2\alpha^*}(r_{\ell+1} - r_\ell) + \frac{Q_{\ell+1}}{\beta^*}\right) - P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{2\alpha^*}(r_{\ell-1} - r_\ell) + \frac{Q_\ell}{\beta^*}\right)\right\}, \tag{5.13}$$

where we define $Q_1 = -\infty$, $Q_{L+1} = \infty$, $r_0 = -\infty$, and $r_{L+1} = \infty$ for convenience.

5.3.3 Asymptotic Distribution of Estimates by Box-SOAV

Corollary 5.3.1 implies that the asymptotic distribution of the estimate $\hat{x}_n$ is characterized by the random variable $\hat{X}$ in (5.10). In fact, we can obtain the following convergence result from Theorem 5.3.1.

Theorem 5.3.2. Let $\mu_{\hat{\boldsymbol{x}}}$ be the empirical distribution corresponding to the CDF given by $P_{\hat{\boldsymbol{x}}}(\hat{x}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(\hat{x}_n \leq \hat{x})$, where $\mathbb{I}(\hat{x}_n \leq \hat{x}) = 1$ if $\hat{x}_n \leq \hat{x}$ and otherwise $\mathbb{I}(\hat{x}_n \leq \hat{x}) = 0$. Moreover, let $\mu_{\hat{X}}$ be the distribution of the random variable $\hat{X}$. The empirical distribution $\mu_{\hat{\boldsymbol{x}}}$ converges weakly in probability to $\mu_{\hat{X}}$, i.e., $\int g\, d\mu_{\hat{\boldsymbol{x}}} \xrightarrow{\mathrm{P}} \int g\, d\mu_{\hat{X}}$ holds as $N \to \infty$ for every continuous compactly supported function $g(\cdot) : [r_1, r_L] \to \mathbb{R}$.

Proof. See Appendix 5.B. □

From Theorem 5.3.2, we can evaluate the asymptotic distribution of the estimate obtained by Box-SOAV. The CDF of $\hat{X}$ is given by

$$
\begin{aligned}
P_{\hat{X}}(\hat{x}) &= \Pr\left(\hat{X} \leq \hat{x}\right) && (5.14)\\
&= \sum_{\ell=1}^{L} p_\ell \Pr\left(\hat{X} \leq \hat{x} \mid X = r_\ell\right) && (5.15)\\
&= \sum_{\ell=1}^{L} p_\ell \Pr\left(\mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}} H\right) \leq \hat{x}\right) && (5.16)\\
&= \sum_{\ell=1}^{L} p_\ell P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}\left\{\mathrm{prox}^{-1}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}(\hat{x}) - r_\ell\right\}\right) && (5.17)
\end{aligned}
$$

for $\hat{x} \in [r_1, r_L] \setminus \mathcal{R}$, where $\mathrm{prox}^{-1}_{\gamma f}(\cdot) : [r_1, r_L] \setminus \mathcal{R} \to \mathbb{R}$ is given by

$$
\mathrm{prox}^{-1}_{\gamma f}(\hat{x}) = \begin{cases} \hat{x} + \gamma Q_2 & (r_1 < \hat{x} < r_2) \\ \vdots & \vdots \\ \hat{x} + \gamma Q_{\ell+1} & (r_\ell < \hat{x} < r_{\ell+1}) \\ \vdots & \vdots \\ \hat{x} + \gamma Q_L & (r_{L-1} < \hat{x} < r_L) \end{cases}. \tag{5.18}
$$

The CDF in (5.17) is not continuous at $\hat{x} \in \mathcal{R}$ because the random variable $\hat{X}$ has a probability mass at $\hat{x} \in \mathcal{R}$. In fact, the conditional probability mass at $\hat{X} = r_\ell$ can be

written as

$$\begin{aligned}
&\Pr\left(\hat{X} = r_\ell \mid X = r_k\right) \\
&= \Pr\left(\mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}\left(r_k + \frac{\alpha^*}{\sqrt{\Delta}} H\right) = r_\ell\right) \qquad (5.19)\\
&= P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}(r_\ell - r_k) + \frac{Q_{\ell+1}}{\beta^*}\right) - P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}(r_\ell - r_k) + \frac{Q_\ell}{\beta^*}\right), \qquad (5.20)
\end{aligned}$$

whereas for $\hat{x} \notin \mathcal{R}$ the conditional density of $\hat{X}$ on the event $X = r_k$ is given by

$$p_{\hat{X}|X=r_k}(\hat{x}) = \frac{\sqrt{\Delta}}{\alpha^*} p_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}\left\{\mathrm{prox}^{-1}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}(\hat{x}) - r_k\right\}\right). \qquad (5.21)$$

Hence, the asymptotic density $p_{\hat{X}}(\hat{x})$ of $\hat{X}$ can be written as

$$p_{\hat{X}}(\hat{x}) = \begin{cases} \sum_{k=1}^{L} p_k \Pr\left(\hat{X} = r_\ell \mid X = r_k\right) \delta_{r_\ell}(\hat{x}) & (\hat{x} = r_\ell) \\ \sum_{k=1}^{L} p_k p_{\hat{X}|X=r_k}(\hat{x}) & (\hat{x} \notin \mathcal{R}) \end{cases}, \qquad (5.22)$$

where $\delta_{r_\ell}(\cdot)$ denotes the Dirac measure at r_ℓ.

In Fig. 5.1, we show the empirical histogram of the CDF of the estimates by Box-SOAV. In the simulation, we set $N = 1000$, $M = 750$, $\Delta = 0.75$, $(p_1, p_2, p_3) = (0.2, 0.6, 0.2)$, $(r_1, r_2, r_3) = (-1, 0, 1)$, and $(q_1, q_2, q_3) = (1, 0.005, 1)$. The SNR defined as $\sum_{\ell=1}^{L} p_\ell r_\ell^2 / \sigma_v^2$ is 20 dB. The empirical result is averaged over 20 independent realizations of the measurement matrix $\boldsymbol{A}$ and the unknown vector $\boldsymbol{x}$. To solve the Box-SOAV optimization, we use the Douglas-Rachford algorithm [61, 101]. For comparison, the figure also shows the asymptotic distribution in (5.17). We can see that the asymptotic distribution obtained from Theorem 5.3.2 agrees well with the empirical histogram of the CDF.

5.3.4 Asymptotically Optimal Quantizer

By using the asymptotic density in the previous subsection, we can design the quantizer $Q(\cdot)$ to obtain the asymptotically optimal SER. Although $Q_{\mathrm{NV}}(\cdot)$ in (5.12) is commonly used as the quantizer, it is not optimal in terms of the asymptotic SER in (5.11) in general. We here present the desired quantizer minimizing the asymptotic SER as the asymptotically optimal quantizer $Q_{\mathrm{AO}}(\hat{x}) = r_{\hat{\ell}}$. The index $\hat{\ell} \in \{1, \cdots, L\}$ can be obtained

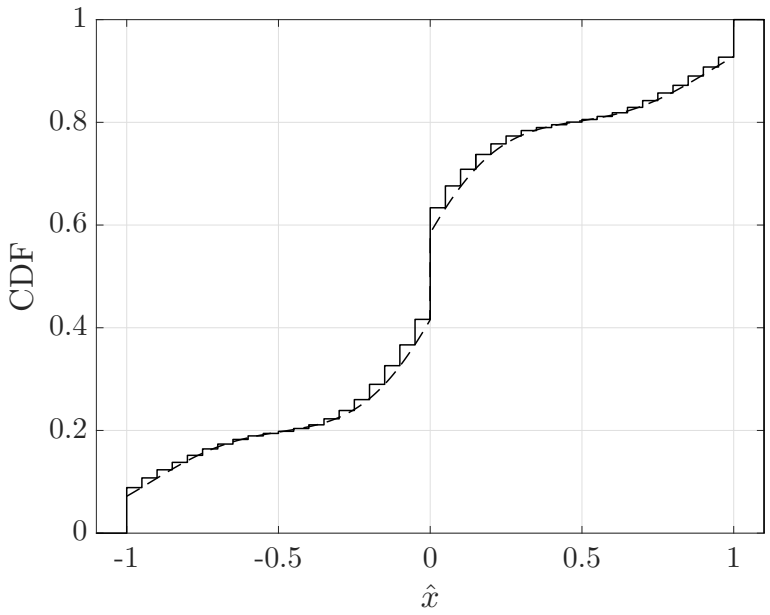


Figure 5.1: The empirical histogram of the CDF (solid line) and the asymptotic distribution (dashed line) of the estimates by Box-SOAV ($N = 1000$, $M = 750$, $\Delta = 0.75$, $(p_1, p_2, p_3) = (0.2, 0.6, 0.2)$, $(r_1, r_2, r_3) = (-1, 0, 1)$, $(q_1, q_2, q_3) = (1, 0.005, 1)$, SNR = 20 dB)

by MAP criterion as

$$\hat{\ell} = \underset{k=1,\ldots,L}{\arg\max} \Pr\left(X = r_k \mid \hat{X} = r_\ell\right) \tag{5.23}$$

$$= \underset{k=1,\ldots,L}{\arg\max} \Pr\left(X = r_k\right) \Pr\left(\hat{X} = r_\ell \mid X = r_k\right) \tag{5.24}$$

$$= \underset{k=1,\ldots,L}{\arg\max}\, p_k \left\{ P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}(r_\ell - r_k) + \frac{Q_{\ell+1}}{\beta^*}\right) - P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}(r_\ell - r_k) + \frac{Q_\ell}{\beta^*}\right)\right\} \tag{5.25}$$

when $\hat{x} = r_\ell$, and

$$\hat{\ell} = \underset{k=1,\ldots,L}{\arg\max} \Pr\left(X = r_k \mid \hat{X} = \hat{x}\right) \tag{5.26}$$

$$= \underset{k=1,\ldots,L}{\arg\max} \Pr\left(X = r_k\right) p_{\hat{X}|X=r_k}(\hat{x}) \tag{5.27}$$

$$= \underset{k=1,\ldots,L}{\arg\max}\, p_k p_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}\left\{\mathrm{prox}^{-1}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}(\hat{x}) - r_k\right\}\right) \tag{5.28}$$

when $\hat{x} \notin \mathcal{R}$. When we use the above $Q_{\mathrm{AO}}(\cdot)$ as the quantizer, the asymptotic SER in (5.11) can be written as

$$\mathrm{SER}_{\mathrm{AO}} = 1 - \sum_{\ell=1}^{L} p_\ell \Pr\left(Q_{\mathrm{AO}}(\hat{X}) = r_\ell \mid X = r_\ell\right) \tag{5.29}$$

$$= 1 - \sum_{\ell=1}^{L} p_\ell \Pr\left(\hat{X} \in Q_{\mathrm{AO}}^{-1}(r_\ell) \mid X = r_\ell\right) \tag{5.30}$$

$$= 1 - \sum_{\ell=1}^{L} p_\ell \mu_{\hat{X}|X=r_\ell}\left(Q_{\mathrm{AO}}^{-1}(r_\ell)\right) \tag{5.31}$$

in general, where we define

$$Q_{\mathrm{AO}}^{-1}(r_\ell) = \{\hat{x} \mid Q_{\mathrm{AO}}(\hat{x}) = r_\ell\} \tag{5.32}$$

and $\mu_{\hat{X}|X=r_\ell}$ denotes the distribution of $\hat{X}$ conditioned on $X = r_\ell$, i.e., the distribution of $\mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}} H\right)$. Note that the CDF corresponding to $\mu_{\hat{X}|X=r_\ell}$ is given by

$$P_{\hat{X}|X=r_\ell}(\hat{x}) = P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}\left\{\mathrm{prox}^{-1}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}(\hat{x}) - r_\ell\right\}\right) \tag{5.33}$$

as shown in (5.17). Once $Q_{\mathrm{AO}}^{-1}(r_\ell)$ is obtained, we can compute the asymptotic SER from (5.31) and (5.33).

In practice, when we use the appropriate parameters q_ℓ of Box-SOAV, the resultant asymptotically optimal quantizer is usually given by the form of

$$Q_{\mathrm{AO}}(\hat{x}) = \begin{cases} r_1 & (\hat{x} < \kappa_2^*) \\ \vdots & \vdots \\ r_\ell & (\kappa_\ell^* \le \hat{x} < \kappa_{\ell+1}^*) \\ \vdots & \vdots \\ r_L & (\kappa_L^* \le \hat{x}) \end{cases}, \tag{5.34}$$

where $-\infty = \kappa_1^* < \kappa_2^* < \cdots < \kappa_L^* < \kappa_{L+1}^* = \infty$ and $r_{\ell-1} < \kappa_\ell^* < r_\ell$ $(\ell = 2, \ldots, L)$. In this case, we have $Q_{\mathrm{AO}}^{-1}(r_\ell) = [\kappa_\ell^*, \kappa_{\ell+1}^*)$ and hence the asymptotic SER in (5.31) can be written as

$$\widetilde{\mathrm{SER}}_{\mathrm{AO}} = 1 - \sum_{\ell=1}^{L} p_\ell \left\{ P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}(\kappa_{\ell+1}^* - r_\ell) + \frac{Q_{\ell+1}}{\beta^*}\right) - P_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}(\kappa_\ell^* - r_\ell) + \frac{Q_\ell}{\beta^*}\right)\right\} \tag{5.35}$$

by using (5.18) and (5.33).

5.3.5 Proposed Parameter Selection for Box-SOAV

The parameters q_ℓ ($\ell = 1, \ldots, L$) in the Box-SOAV optimization (5.4) affects the performance of the reconstruction. From the results of the previous subsections, once the parameters q_ℓ are fixed, the asymptotic SER of Box-SOAV with the asymptotically optimal quantizer can be evaluated as follows:

1. Calculate α^* and β^* in Theorem 5.3.1.
2. Obtain the asymptotically optimal quantizer $Q_{\mathrm{AO}}(\cdot)$ based on (5.25) and (5.28).
3. Compute the asymptotic SER in (5.11) (or (5.35) in many cases).

We thus propose the approach to determine the parameters q_ℓ (or Q_ℓ in (5.6)) by numerically minimizing the resultant asymptotic SER in (5.11). In the following examples, we compare performance of the quantizers $Q_{\mathrm{NV}}(\cdot)$ and $Q_{\mathrm{AO}}(\cdot)$ with the proposed parameter selection.

Example 5.3.1 (Binary Vector)**.** We consider the reconstruction of the binary vector $\boldsymbol{x} \in \{r_1, r_2\}^N$. When the estimate $\hat{x}$ of an element of $\boldsymbol{x}$ equals r_1 or r_2, we should just quantize it on the basis of (5.25). For $\hat{x} \in (r_1, r_2)$, the output of the asymptotically optimal quantizer is determined from (5.28). We thus obtain the value of κ_2^* such that

$$p_1 p_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}\left\{\mathrm{prox}^{-1}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}}f}\left(\kappa_2^*\right) - r_1\right\}\right) = p_2 p_{\mathrm{G}}\left(\frac{\sqrt{\Delta}}{\alpha^*}\left\{\mathrm{prox}^{-1}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}}f}\left(\kappa_2^*\right) - r_2\right\}\right). \tag{5.36}$$

If the solution of (5.36) lies in (r_1, r_2), we have

$$\kappa_2^* = \mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}}f}\left(\frac{1}{2}(r_1 + r_2) + \frac{(\alpha^*)^2}{\Delta}\frac{1}{r_2 - r_1}\log\frac{p_1}{p_2}\right). \tag{5.37}$$

In this case, the asymptotically optimal quantizer can be written as

$$Q_{\mathrm{AO}}(\hat{x}) = \begin{cases} r_1 & (\hat{x} < \kappa_2^*) \\ r_2 & (\kappa_2^* \le \hat{x}) \end{cases} \tag{5.38}$$

for $\hat{x} \in (r_1, r_2)$. Figure 5.2 shows an example of the asymptotic density of the estimates by Box-SOAV given by (5.22). In the figure, we set $\Delta = 0.6$, $(p_1, p_2) = (0.3, 0.7)$, $(r_1, r_2) = (-1, 1)$, $(q_1, q_2) = (0.5, 0.5)$, and SNR of 15 dB. We can see that a certain probability mass is located at $\hat{x} = \pm 1$. The functions $p_k p_{\hat{X}|X=r_k}(\hat{x})$ $(k = 1, 2)$ are also plotted by the dotted lines in the figure. We can confirm that the two curves cross at $\hat{x} = \kappa_2^*$.

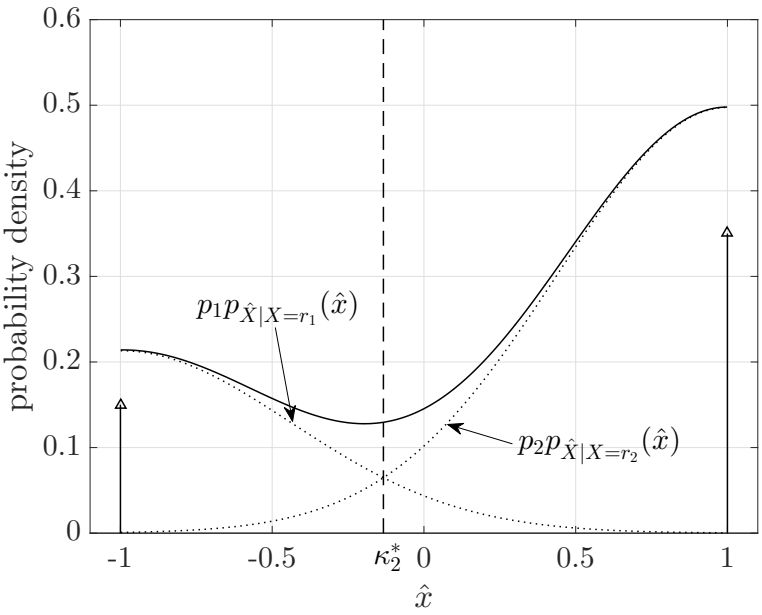


Figure 5.2: The asymptotic density (solid line) of the estimates by Box-SOAV and the threshold κ_2^* of the asymptotically optimal quantizer ($\Delta = 0.6$, $(p_1, p_2) = (0.3, 0.7)$, $(r_1, r_2) = (-1, 1)$, $(q_1, q_2) = (0.5, 0.5)$, SNR = 15 dB)

We can obtain the asymptotically optimal parameters of the Box-SOAV optimization from the theoretical result. For the reconstruction of $\boldsymbol{x} \in \{r_1, r_2\}^N$, the Box-SOAV optimization is given by

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{s}\in[r_1,r_2]^N}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 + q_1 \|\boldsymbol{s} - r_1\mathbf{1}\|_1 + q_2 \|\boldsymbol{s} - r_2\mathbf{1}\|_1 \right\} \tag{5.39}$$

$$= \underset{\boldsymbol{s}\in[r_1,r_2]^N}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 + Q_2 \sum_{n=1}^{N} s_n \right\} \tag{5.40}$$

because $q_1 \|\boldsymbol{s} - r_1\mathbf{1}\|_1 + q_2 \|\boldsymbol{s} - r_2\mathbf{1}\|_1 = Q_2 \sum_{n=1}^N s_n + (\text{const.})$ for $\boldsymbol{s} \in [r_1, r_2]^N$. Hence, the performance of Box-SOAV depends only on Q_2. By numerically computing the value of Q_2 minimizing the asymptotic SER, we can obtain the optimal $Q_{\mathrm{AO}}(\cdot)$ and the corresponding asymptotic SER. For example, Fig. 5.3 shows the asymptotic SER of $Q_{\mathrm{AO}}(\cdot)$ versus Q_2 when $\Delta = 0.7$, $(r_1, r_2) = (0, 1)$, $(p_1, p_2) = (0.8, 0.2)$, and SNR of 15 dB. From the figure, we can see that the asymptotic performance of Box-SOAV largely depends on the parameter Q_2. By using the optimal value Q_2^* of Q_2 minimizing the asymptotic SER, we can obtain the asymptotically optimal values of α^*, β^*, and κ_2^* in (5.37).

We then compare the performance of quantizer $Q_{\mathrm{NV}}(\cdot)$ in (5.12) and $Q_{\mathrm{AO}}(\cdot)$ in (5.38).

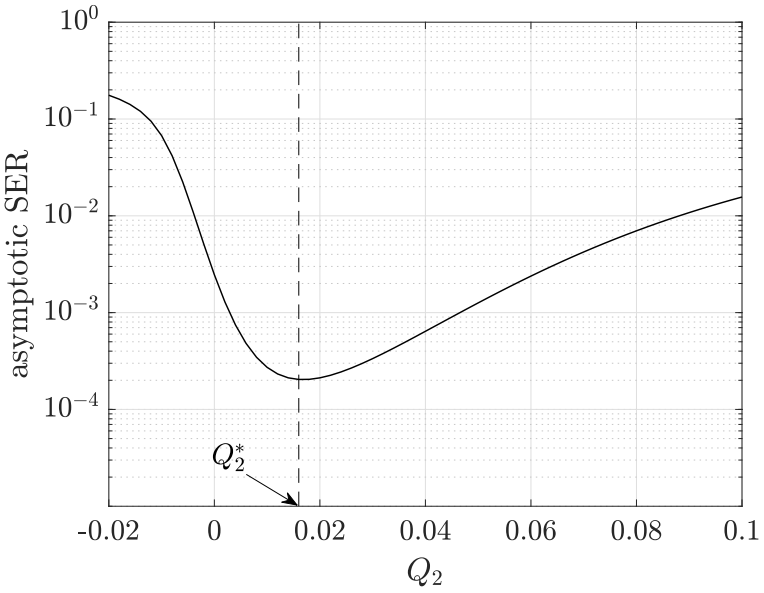


Figure 5.3: Asymptotic SER of $Q_{\mathrm{AO}}(\cdot)$ versus Q_2 ($\Delta = 0.7$, $(r_1, r_2) = (0, 1)$, $(p_1, p_2) = (0.8, 0.2)$, SNR = 15 dB)

Figure 5.4 shows the asymptotic SER of Box-SOAV with four cases: (i) $Q_{\mathrm{NV}}(\cdot)$ with $Q_2 = 0$, (ii) $Q_{\mathrm{NV}}(\cdot)$ with the optimal Q_2, (iii) $Q_{\mathrm{AO}}(\cdot)$ with $Q_2 = 0$, and (iv) $Q_{\mathrm{AO}}(\cdot)$ with the proposed optimal parameter selection. In the figure, we set $\Delta = 0.7$, $(r_1, r_2) = (-1, 1)$, and SNR of 15 dB. The figure shows that the proposed optimal parameter selection can achieve better performance than the naive selection $Q_2 = 0$. Moreover, the performance of $Q_{\mathrm{AO}}(\cdot)$ is better than that of $Q_{\mathrm{NV}}(\cdot)$, especially when the difference between p_1 and p_2 is large.

Example 5.3.2 (Discrete-Valued Sparse Vector)**.** The discrete-valued vector $\boldsymbol{x}$ with $(p_1, p_2, p_3) = ((1 - p_0)/2, p_0, (1 - p_0)/2)$ and $(r_1, r_2, r_3) = (-r, 0, r)$ $(r > 0)$ becomes sparse when p_0 is large. By a similar discussion to Example 5.3.1, the asymptotically optimal quantizer is given by

$$Q_{\mathrm{AO}}(\hat{x}) = \begin{cases} -r & (\hat{x} < \kappa_2^*) \\ 0 & (\kappa_2^* \le \hat{x} < \kappa_3^*) \\ r & (\kappa_3^* \le \hat{x}) \end{cases} \tag{5.41}$$

when $-r < \kappa_2^* < 0$, where

$$\kappa_2^* = \mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}\left(-\frac{1}{2}r + \frac{(\alpha^*)^2}{\Delta}\frac{1}{r}\log\frac{1-p_0}{2p_0}\right), \tag{5.42}$$

$$\kappa_3^* = -\kappa_2^*. \tag{5.43}$$

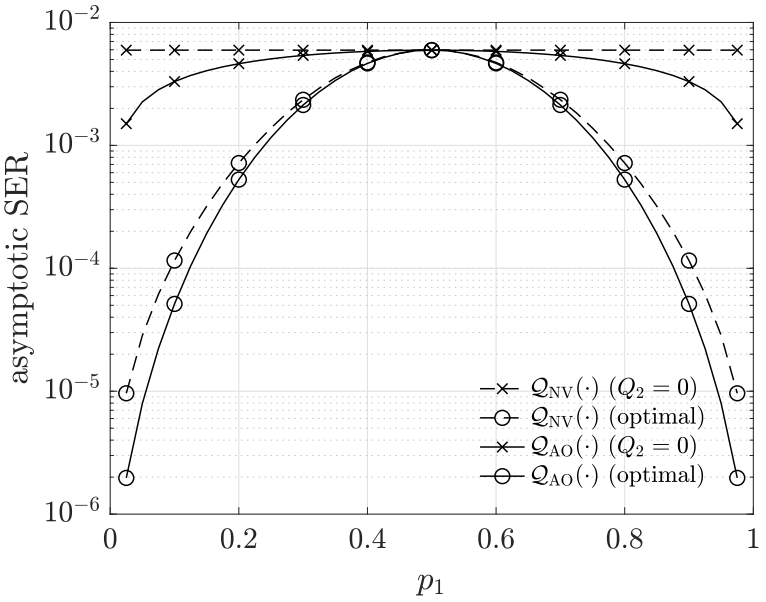


Figure 5.4: Asymptotic SER versus p_1 for binary vector ($\Delta = 0.7$, $(r_1, r_2) = (-1, 1)$, SNR = 15 dB)

For the reconstruction of $\boldsymbol{x}$ via Box-SOAV in this scenario, we can set $q_1 = q_3$ from the symmetry of the distribution. As a result, the Box-SOAV optimization problem can be written as

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{s}\in[-r,r]^N}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 + q_1 \|\boldsymbol{s} + r\boldsymbol{1}\|_1 + q_2 \|\boldsymbol{s}\|_1 + q_1 \|\boldsymbol{s} - r\boldsymbol{1}\|_1 \right\} \tag{5.44}$$

$$= \underset{\boldsymbol{s}\in[-r,r]^N}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 + q_2 \|\boldsymbol{s}\|_1 \right\} \tag{5.45}$$

because $q_1 \|\boldsymbol{s} + r\boldsymbol{1}\|_1 + q_1 \|\boldsymbol{s} - r\boldsymbol{1}\|_1 = 2q_1 rN = (\text{const.})$ for $\boldsymbol{s} \in [-r, r]^N$. Hence, only q_2 is the parameter to be optimized. Note that the Box-SOAV optimization problem results in Box-LASSO [94] in this case. Figure 5.5 shows the asymptotic SER of Box-SOAV with $Q_{\mathrm{NV}}(\cdot)$ and $Q_{\mathrm{AO}}(\cdot)$. In the figure, we set $(p_1, p_2, p_3) = (0.05, 0.9, 0.05)$, $(r_1, r_2, r_3) = (-1, 0, 1)$, and SNR of 15 dB. The parameter q_2 of Box-SOAV is numerically chosen by minimizing the asymptotic SER for each quantizer. We can observe that the asymptotically optimal quantizer $Q_{\mathrm{AO}}(\cdot)$ outperforms $Q_{\mathrm{NV}}(\cdot)$ especially for large $\Delta = M/N$.

Remark 5.3.1. In Chapter 6, a parameter selection method has been proposed for the SOAV optimization on the basis of the analysis of the DAMP algorithm. However, the

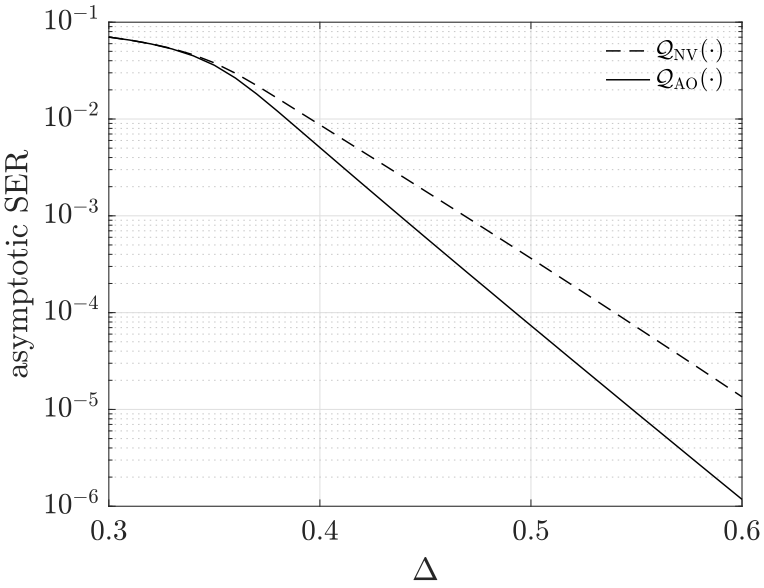


Figure 5.5: Asymptotic SER versus $\Delta = M/N$ for discrete-valued sparse vector $((p_1, p_2, p_3) = (0.05, 0.9, 0.05), (r_1, r_2, r_3) = (-1, 0, 1)$, SNR $= 15$ dB)

method does not necessarily minimize the SER of the SOAV optimization because it minimizes the required number of measurements for the perfect reconstruction in the noise-free case. Since it does not take the noise variance into consideration, it is difficult to fairly compare the result in Chapter 6 with the theoretical result in this paper.

5.4 Proof of Theorem 5.3.1

In this section, we show the proof of Theorem 5.3.1. Although the procedure of the proof roughly follows the analysis using CGMT in the literature (e.g., [56, 91]), we need to modify several parts for our problem.

5.4.1 (PO)

To obtain the (PO) problem for the proof, we firstly define the error vector as $\boldsymbol{w} := \boldsymbol{s} - \boldsymbol{x}$ and rewrite the Box-SOAV optimization (5.4) as

$$\min_{\boldsymbol{w} \in \mathcal{S}_{\mathrm{w}}} \frac{1}{N} \left\{ \frac{1}{2} \left\| \boldsymbol{A}\boldsymbol{w} - \boldsymbol{v} \right\|_2^2 + \sum_{\ell=1}^{L} q_\ell \left\| \boldsymbol{x} + \boldsymbol{w} - r_\ell \mathbf{1} \right\|_1 \right\}, \tag{5.46}$$

where $\mathcal{S}_{\mathrm{w}} = \left\{ \boldsymbol{z} \in \mathbb{R}^N \mid r_1 - x_n \le z_n \le r_L - x_n \ (n = 1, \ldots, N) \right\}$ and the objective function is normalized by N. Since the convex conjugate of the function $\xi(\boldsymbol{z}) := \frac{1}{2} \left\| \boldsymbol{z} \right\|_2^2$ is given

by

$$\xi^*(\boldsymbol{z}^*) \coloneqq \max_{\boldsymbol{u}\in\mathbb{R}^M} \left\{ \boldsymbol{u}^\mathsf{T}\boldsymbol{z}^* - \frac{1}{2}\|\boldsymbol{u}\|_2^2 \right\} \tag{5.47}$$

$$= \frac{1}{2}\|\boldsymbol{z}^*\|_2^2, \tag{5.48}$$

we have

$$\frac{1}{2}\|\boldsymbol{A}\boldsymbol{w}-\boldsymbol{v}\|_2^2 = \max_{\boldsymbol{u}\in\mathbb{R}^M} \left\{ \sqrt{N}\boldsymbol{u}^\mathsf{T}(\boldsymbol{A}\boldsymbol{w}-\boldsymbol{v}) - \frac{N}{2}\|\boldsymbol{u}\|_2^2 \right\}. \tag{5.49}$$

Hence, the optimization problem (5.46) is equivalent to

$$\min_{\boldsymbol{w}\in\mathcal{S}_\mathrm{w}} \max_{\boldsymbol{u}\in\mathbb{R}^M} \frac{1}{N}\left\{ \sqrt{N}\boldsymbol{u}^\mathsf{T}(\boldsymbol{A}\boldsymbol{w}-\boldsymbol{v}) - \frac{N}{2}\|\boldsymbol{u}\|_2^2 + \sum_{\ell=1}^{L} q_\ell \|\boldsymbol{x}+\boldsymbol{w}-r_\ell\boldsymbol{1}\|_1 \right\}. \tag{5.50}$$

Let $\boldsymbol{w}^*$ and $\boldsymbol{u}^*$ be the optimal values of $\boldsymbol{w}$ and $\boldsymbol{u}$, respectively. Since $\boldsymbol{u}^*$ satisfies $\boldsymbol{u}^* = \frac{1}{\sqrt{N}}(\boldsymbol{A}\boldsymbol{w}^*-\boldsymbol{v})$ and $\boldsymbol{w}^*$ is bounded, there exists a constant C_u such that $\|\boldsymbol{u}^*\|_2 \le C_\mathrm{u}$ with probability approaching 1. We can thus rewrite (5.50) as

$$\min_{\boldsymbol{w}\in\mathcal{S}_\mathrm{w}} \max_{\boldsymbol{u}\in\mathcal{S}_\mathrm{u}} \left\{ \frac{1}{N}\boldsymbol{u}^\mathsf{T}\left(\sqrt{N}\boldsymbol{A}\right)\boldsymbol{w} - \frac{1}{\sqrt{N}}\boldsymbol{v}^\mathsf{T}\boldsymbol{u} - \frac{1}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{N}\sum_{\ell=1}^{L} q_\ell \|\boldsymbol{x}+\boldsymbol{w}-r_\ell\boldsymbol{1}\|_1 \right\}, \tag{5.51}$$

where $\mathcal{S}_\mathrm{u} = \left\{\boldsymbol{z}\in\mathbb{R}^M \mid \|\boldsymbol{z}\|_2 \le C_\mathrm{u}\right\}$. The optimization problem (5.51) takes the form of (PO) in (5.1).

5.4.2 (AO)

We then analyze the (AO) problem corresponding to (5.51). We can confirm that $\mathcal{S}_\mathrm{w}$ and $\mathcal{S}_\mathrm{u}$ are closed compact sets and the function $-\frac{1}{\sqrt{N}}\boldsymbol{v}^\mathsf{T}\boldsymbol{u} - \frac{1}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{N}\sum_{\ell=1}^{L} q_\ell \|\boldsymbol{x}+\boldsymbol{w}-r_\ell\boldsymbol{1}\|_1$ is convex-concave on $\mathcal{S}_\mathrm{w}\times\mathcal{S}_\mathrm{u}$. Hence, the (AO) problem corresponding to (5.51) is given by

$$\min_{\boldsymbol{w}\in\mathcal{S}_\mathrm{w}} \max_{\boldsymbol{u}\in\mathcal{S}_\mathrm{u}} \left\{ \frac{1}{N}\left(\|\boldsymbol{w}\|_2 \boldsymbol{g}^\mathsf{T}\boldsymbol{u} - \|\boldsymbol{u}\|_2 \boldsymbol{h}^\mathsf{T}\boldsymbol{w}\right) - \frac{1}{\sqrt{N}}\boldsymbol{v}^\mathsf{T}\boldsymbol{u} \right.$$
$$\left. - \frac{1}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{N}\sum_{\ell=1}^{L} q_\ell \|\boldsymbol{x}+\boldsymbol{w}-r_\ell\boldsymbol{1}\|_1 \right\}, \tag{5.52}$$

which can be rewritten as

$$\min_{\boldsymbol{w}\in\mathcal{S}_{\mathrm{w}}}\max_{\boldsymbol{u}\in\mathcal{S}_{\mathrm{u}}}\left\{\frac{1}{\sqrt{N}}\left(\frac{\|\boldsymbol{w}\|_2}{\sqrt{N}}\boldsymbol{g}^{\mathsf{T}}-\boldsymbol{v}^{\mathsf{T}}\right)\boldsymbol{u}-\frac{1}{N}\|\boldsymbol{u}\|_2\,\boldsymbol{h}^{\mathsf{T}}\boldsymbol{w}\right.$$
$$\left.-\frac{1}{2}\|\boldsymbol{u}\|_2^2+\frac{1}{N}\sum_{\ell=1}^{L}q_\ell\,\|\boldsymbol{x}+\boldsymbol{w}-r_\ell\mathbf{1}\|_1\right\}. \tag{5.53}$$

Since both $\boldsymbol{g}$ and $\boldsymbol{v}$ are Gaussian, $\frac{\|\boldsymbol{w}\|_2}{\sqrt{N}}\boldsymbol{g}-\boldsymbol{v}$ is also Gaussian distributed with mean $\mathbf{0}$ and covariance matrix $\left(\frac{\|\boldsymbol{w}\|_2^2}{N}+\sigma_{\mathrm{v}}^2\right)\boldsymbol{I}$. We can thus rewrite $\left(\frac{\|\boldsymbol{w}\|_2}{\sqrt{N}}\boldsymbol{g}^{\mathsf{T}}-\boldsymbol{v}^{\mathsf{T}}\right)\boldsymbol{u}$ as $\sqrt{\frac{\|\boldsymbol{w}\|_2^2}{N}+\sigma_{\mathrm{v}}^2}\boldsymbol{g}^{\mathsf{T}}\boldsymbol{u}$ by the slight abuse of notation, where $\boldsymbol{g}$ is the random vector with i.i.d. standard Gaussian elements. By setting $\|\boldsymbol{u}\|_2=\beta$, the (AO) problem can be further rewritten as

$$\min_{\boldsymbol{w}\in\mathcal{S}_{\mathrm{w}}}\max_{0\le\beta\le C_{\mathrm{u}}}\left\{\sqrt{\frac{\|\boldsymbol{w}\|_2^2}{N}+\sigma_{\mathrm{v}}^2}\frac{\beta\,\|\boldsymbol{g}\|_2}{\sqrt{N}}-\frac{1}{N}\beta\boldsymbol{h}^{\mathsf{T}}\boldsymbol{w}-\frac{1}{2}\beta^2\right.$$
$$\left.+\frac{1}{N}\sum_{\ell=1}^{L}q_\ell\,\|\boldsymbol{x}+\boldsymbol{w}-r_\ell\mathbf{1}\|_1\right\}. \tag{5.54}$$

From the identity $\chi=\min_{\alpha>0}\left(\frac{\alpha}{2}+\frac{\chi^2}{2\alpha}\right)$ for $\chi\ (>0)$, we have

$$\sqrt{\frac{\|\boldsymbol{w}\|_2^2}{N}+\sigma_{\mathrm{v}}^2}=\min_{\alpha>0}\left(\frac{\alpha}{2}+\frac{\frac{\|\boldsymbol{w}\|_2^2}{N}+\sigma_{\mathrm{v}}^2}{2\alpha}\right) \tag{5.55}$$

and rewrite (5.54) as

$$\max_{\beta>0}\min_{\alpha>0}\left\{\frac{\alpha\beta}{2}\frac{\|\boldsymbol{g}\|_2}{\sqrt{N}}+\frac{\sigma_{\mathrm{v}}^2\beta}{2\alpha}\frac{\|\boldsymbol{g}\|_2}{\sqrt{N}}-\frac{1}{2}\beta^2-\frac{1}{N}\sum_{n=1}^{N}\frac{\alpha\beta h_n^2}{2}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\right.$$
$$\left.+\frac{\beta}{\alpha}\frac{\|\boldsymbol{g}\|_2}{\sqrt{N}}\min_{\boldsymbol{w}\in\mathcal{S}_{\mathrm{w}}}\frac{1}{N}\sum_{n=1}^{N}J_n(w_n)\right\}, \tag{5.56}$$

where

$$J_n(w_n)=\frac{1}{2}\left(w_n-\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha h_n\right)^2+\frac{\alpha}{\beta}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\sum_{\ell=1}^{L}q_\ell\,|x_n+w_n-r_\ell|\,. \tag{5.57}$$

Note that the objective function becomes separable for w_n in (5.56), and that the change in the range of β does not change the optimal value. Since the optimization with respect to $\boldsymbol{w}$ in (5.56) is given by

$$\begin{aligned}
&\min_{\boldsymbol{w}\in\mathcal{S}_{\mathrm{w}}} \frac{1}{N}\sum_{n=1}^{N} J_n(w_n) \\
&= \min_{\boldsymbol{s}\in[r_1,r_L]^N} \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{2}\left\{s_n - \left(x_n + \frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha h_n\right)\right\}^2 + \frac{\alpha}{\beta}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\sum_{\ell=1}^{L} q_\ell\,|s_n - r_\ell|\right] \qquad (5.58)\\
&= \frac{1}{N}\sum_{n=1}^{N} \mathrm{env}_{\frac{\alpha}{\beta}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2} f}\left(x_n + \frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha h_n\right), \qquad (5.59)
\end{aligned}$$

(5.56) can be rewritten as

$$\phi_N^* = \max_{\beta>0}\min_{\alpha>0} F_N(\alpha,\beta), \qquad (5.60)$$

where

$$\begin{aligned}
F_N(\alpha,\beta) = {}& \frac{\alpha\beta}{2}\frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} + \frac{\sigma_{\mathrm{v}}^2\beta}{2\alpha}\frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} - \frac{1}{2}\beta^2 - \frac{1}{N}\sum_{n=1}^{N}\frac{\alpha\beta h_n^2}{2}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2} \\
&+ \frac{\beta}{\alpha}\frac{\|\boldsymbol{g}\|_2}{\sqrt{N}}\frac{1}{N}\sum_{n=1}^{N}\mathrm{env}_{\frac{\alpha}{\beta}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2} f}\left(x_n + \frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha h_n\right). \qquad (5.61)
\end{aligned}$$

The optimal value of $\boldsymbol{w}$ is given by

$$\hat{\boldsymbol{w}}_N(\boldsymbol{h},\boldsymbol{x}) = \mathrm{prox}_{\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2} f}\left(\boldsymbol{x} + \frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha_N^*\boldsymbol{h}\right) - \boldsymbol{x}, \qquad (5.62)$$

where α_N^* and β_N^* are the optimal values of α and β corresponding to ϕ_N^*, respectively.

5.4.3 Applying CGMT

We then consider the condition (i) of Theorem 5.2.1. As $N\to\infty$, $F_N(\alpha,\beta)$ converges pointwise to $F(\alpha,\beta)$ defined in Theorem 5.3.1. Let $\phi^* = \max_{\beta>0}\min_{\alpha>0} F(\alpha,\beta)$ and denote the optimal values of α and β by α^* and β^*, respectively. By a similar discussion to the proof of [56, Lemma IV. 1], we have $\phi_N^* \xrightarrow{\mathrm{P}} \phi^*$ and $(\alpha_N^*,\beta_N^*) \xrightarrow{\mathrm{P}} (\alpha^*,\beta^*)$ as $N\to\infty$. Hence, the optimal value of (AO) satisfies the condition (i) in CGMT for $\bar{\phi} = \phi^*$ and any $\eta > 0$.

Next, we define the set $\mathcal{S}$ used in CGMT. We have the following lemma for the optimizer $\hat{\boldsymbol{w}}_N$ of (AO) in (5.62).

Lemma 5.4.1. For any function $\psi(\cdot,\cdot) \in \mathcal{L}$ (given by (5.8)), we have

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \psi\left(\hat{w}_{N,n}(h_n, x_n), x_n\right) = \mathrm{E}\left[\psi\left(\hat{X} - X, X\right)\right], \tag{5.63}$$

where $\hat{w}_{N,n}(h_n, x_n)$ denotes the nth element of $\hat{\boldsymbol{w}}_N$ in (5.62).

Proof. See Appendix 5.C. □

From Lemma 5.4.1, we can define

$$\mathcal{S} = \left\{ \boldsymbol{z} \in \mathbb{R}^N \left| \left| \frac{1}{N}\sum_{n=1}^{N} \psi(z_n, x_n) - \mathrm{E}\left[\psi\left(\hat{X} - X, X\right)\right] \right| < \varepsilon \right\} \tag{5.64}$$

and obtain $\hat{\boldsymbol{w}}_N(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{S}$ with probability approaching 1 for any ε (> 0).

Finally, we consider the condition (ii) of CGMT. From the strong convexity in $\boldsymbol{w}$ of the objective function in (5.54), we can show $\phi_{\mathcal{S}^c} \geq \phi_N^* + \tilde{\eta}$ with probability approaching 1 for a constant $\tilde{\eta}$ (> 0), where $\phi_{\mathcal{S}^c}$ denotes the optimal value of (AO) under the restriction of $\boldsymbol{w} \in \mathcal{S}^c$. Hence, by setting $\bar{\phi} = \phi^*, \eta = \tilde{\eta}/3$ in Theorem 5.2.1, we can use CGMT for $\mathcal{S}$, i.e., Lemma 5.4.1 holds not only for the optimizer $\hat{\boldsymbol{w}}_N$ of (AO) but also for that of (PO). We thus conclude the proof.

5.5 Simulation Results

In this section, we compare the theoretical results by Corollary 5.3.1 and the empirical performance obtained by computer simulations. In the simulations, the measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ and the noise vector $\boldsymbol{v} \in \mathbb{R}^M$ satisfy the assumptions in Theorem 5.3.1.

We firstly compare the empirical performance of the Box-SOAV optimization with the theoretical result. Figure 5.6 shows the SER performance for the binary vector with $(r_1, r_2) = (0, 1)$ for measurement ratios of $\Delta = 0.5, 0.6$, and 0.7. The distribution of the unknown vector is given by $(p_1, p_2) = (0.8, 0.2)$. The SNR is 15 dB. We use $Q_{\mathrm{AO}}(\cdot)$ as the quantizer and the parameter of Box-SOAV is optimized as in Example 5.3.1. In the figure, 'empirical' represents the empirical performance obtained by averaging the SER over 2000 independent realizations of the measurement matrix. We use Douglas-Rachford algorithm [61, 101] to solve the Box-SOAV optimization problem. We can see that the empirical performance agrees well with the theoretical prediction denoted by 'theoretical' for large N.

Next, we show that the proposed optimal parameters and quantizer can achieve better performance than some fixed parameter and the naive quantizer. Figure 5.7 shows the SER performance of the Box-SOAV optimization versus the SNR, where $N = 1000$, $\Delta = 0.8$, $(r_1, r_2, r_3) = (-1, 0, 1)$, and $(p_1, p_2, p_3) = (0.1, 0.8, 0.1)$. As described in

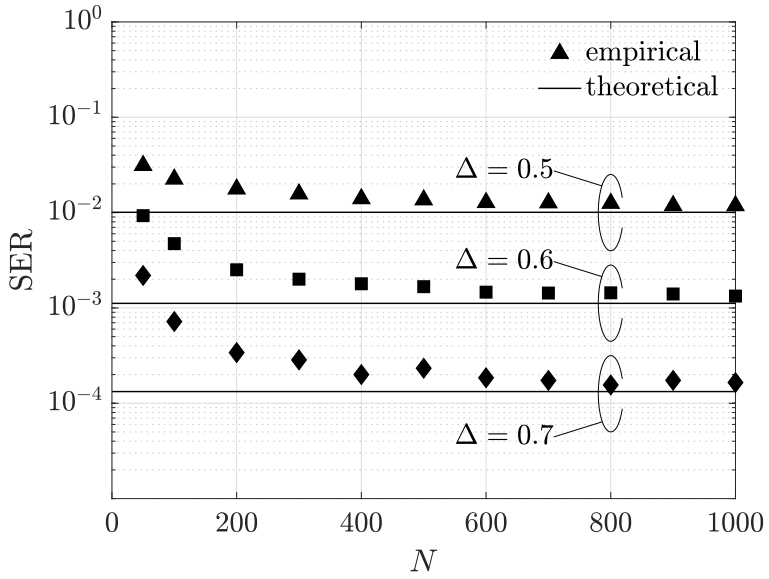


Figure 5.6: SER of Box-SOAV versus N ($(r_1, r_2) = (0, 1)$, $(p_1, p_2) = (0.8, 0.2)$, SNR = 15 dB)

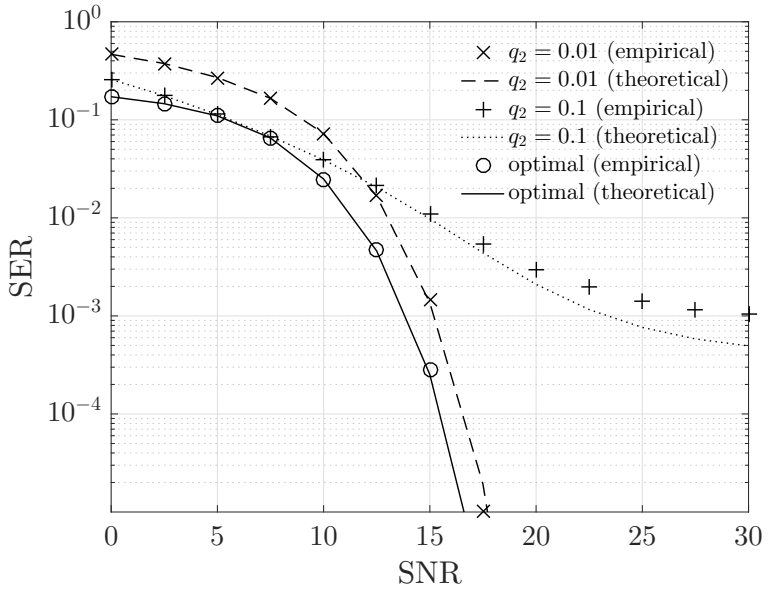


Figure 5.7: SER of Box-SOAV versus the SNR ($N = 1000$, $\Delta = 0.8$, $(r_1, r_2, r_3) = (-1, 0, 1)$, $(p_1, p_2, p_3) = (0.1, 0.8, 0.1)$)

Example 5.3.2, the parameter of the Box-SOAV optimization is only q_2 in this case. In the figure, '$q_2 = 0.01$' and '$q_2 = 0.1$' denote the performance of the Box-SOAV optimization with $q_2 = 0.01$ and $q_2 = 0.1$, respectively. We use the naive quantizer

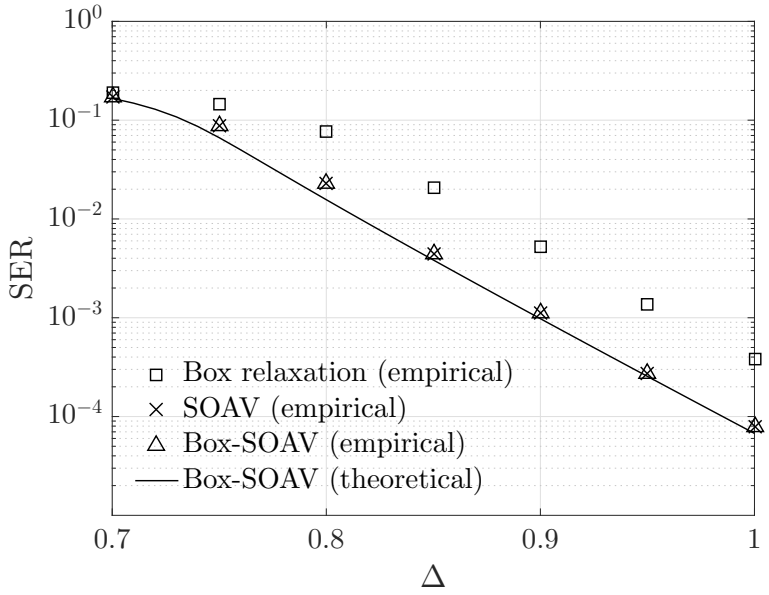


Figure 5.8: SER versus $\Delta = M/N$ ($N = 1500$, $(r_1, r_2, r_3) = (-1, 0, 1)$, $(p_1, p_2, p_3) = (0.25, 0.5, 0.25)$, SNR = 20 dB)

$Q_{\mathrm{NV}}(\cdot)$ in (5.12) for these methods. Also, 'optimal' represents the performance of the Box-SOAV optimization with the proposed asymptotically optimal parameter and quantizer $Q_{\mathrm{AO}}(\cdot)$. We can see that the proposed parameter and quantizer outperforms the fixed parameter and the naive quantizer. It should be noted that the optimal value of q_2 clearly depends on the SNR and the proposed approach can determine the value adaptively.

Finally, we compare the performance of the Box-SOAV optimization with some conventional methods. Figure 5.8 shows the SER performance versus Δ for the unknown discrete-valued vector with $(r_1, r_2, r_3) = (-1, 0, 1)$. We assume $N = 1500$, $(p_1, p_2, p_3) = (0.25, 0.5, 0.25)$, and the SNR of 20 dB. In the figure, 'SOAV' and 'Box-SOAV' represent the conventional SOAV optimization and the Box-SOAV optimization, respectively. We use $Q_{\mathrm{AO}}(\cdot)$ as the quantizer and the parameter of Box-SOAV is optimized as in Example 5.3.2. For comparison, we also evaluate the performance of the box relaxation method [54, 55] given by

$$\min_{\boldsymbol{s}\in[-1,1]^N} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 . \tag{5.65}$$

From the figure, we can see that the empirical performances of Box-SOAV and SOAV are close to the theoretical prediction of Box-SOAV. Moreover, they have better performance than the box relaxation method because they effectively use the knowledge of the distribution of the unknown vector.

5.6 Conclusion

In this chapter, we have derived the theoretical asymptotic performance of the discrete-valued vector reconstruction using the Box-SOAV optimization. By using the CGMT framework, we have shown that the asymptotic SER can be obtained with Corollary 5.3.1. Moreover, we have derived the asymptotic distribution of the estimate obtained by the Box-SOAV optimization. The asymptotic results enable us to obtain the optimal parameters of the Box-SOAV optimization and the asymptotically optimal quantizer. Simulation results show that the empirical performance is close to theoretical prediction of Corollary 5.3.1 when the problem size is sufficiently large. We have also shown that we can improve the performance of the Box-SOAV optimization by using the proposed asymptotically optimal parameters and quantizer.

Appendix 5.A Proof of Corollary 5.3.1

Let $\psi(w, x) = 1 - \chi(w + x, x)$ in Theorem 5.3.1, where the function $\chi(\cdot, \cdot) : [r_1, r_L] \times \mathcal{R} \to \{0, 1\}$ is given by

$$\chi(\hat{x}, x) = \begin{cases} 1 & (Q(\hat{x}) = x) \\ 0 & (Q(\hat{x}) \neq x) \end{cases}. \tag{5.66}$$

The left hand side of (5.9) can be written as

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} (1 - \chi(\hat{x}_n, x_n)) = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \left\| Q(\hat{\boldsymbol{x}}) - \boldsymbol{x} \right\|_0, \tag{5.67}$$

whereas the right hand side can be written as

$$\begin{aligned} \mathrm{E}\left[1 - \chi(\hat{X}, X)\right] &= 1 - \Pr\left(Q(\hat{X}) = X\right) && (5.68) \\ &= 1 - \sum_{\ell=1}^{L} p_\ell \Pr\left(Q(\hat{X}) = r_\ell \mid X = r_\ell\right), && (5.69) \end{aligned}$$

which concludes (5.11). Although $\chi(\cdot, r_\ell)$ is not Lipschitz continuous, we can approximate $\chi(\cdot, r_\ell)$ with a Lipschitz function because H is a continuous random variable and the probability measure for the discontinuity point of $\chi(\cdot, r_\ell)$ is zero (For a similar discussion, see [56, Lemma A.4]).

Appendix 5.B Proof of Theorem 5.3.2

It is sufficient to prove

$$\lim_{N\to\infty} \Pr\left(\left|\int g\, d\mu_{\hat{\boldsymbol{x}}} - \int g\, d\mu_{\hat{X}}\right| < \varepsilon\right) = 1 \tag{5.70}$$

for any $\varepsilon > 0$. From the Stone-Weierstrass theorem [102], there exists a polynomial $v(\cdot) : [r_1, r_L] \to \mathbb{R}$ such that

$$|g(x) - v(x)| < \frac{\varepsilon}{3} \tag{5.71}$$

for any $x \in [r_1, r_L]$. Hence, the absolute value in (5.70) can be upper bounded as

$$\begin{aligned}
&\left|\int g\, d\mu_{\hat{\boldsymbol{x}}} - \int g\, d\mu_{\hat{X}}\right| \\
&\leq \left|\int g\, d\mu_{\hat{\boldsymbol{x}}} - \int v\, d\mu_{\hat{\boldsymbol{x}}}\right| + \left|\int v\, d\mu_{\hat{\boldsymbol{x}}} - \int v\, d\mu_{\hat{X}}\right| + \left|\int v\, d\mu_{\hat{X}} - \int g\, d\mu_{\hat{X}}\right| && (5.72)\\
&< \left|\int v\, d\mu_{\hat{\boldsymbol{x}}} - \int v\, d\mu_{\hat{X}}\right| + \frac{2}{3}\varepsilon && (5.73)
\end{aligned}$$

Note that the polynomial $v(\cdot)$ is Lipschitz in $[r_1, r_L]$. We then define $\psi(w, x) = v(w + x)$ in Theorem 5.3.1 and obtain $\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} v(\hat{x}_n) = \mathrm{E}\left[v(\hat{X})\right]$, i.e.,

$$\lim_{N\to\infty} \Pr\left(\left|\int v\, d\mu_{\hat{\boldsymbol{x}}} - \int v\, d\mu_{\hat{X}}\right| < \frac{\varepsilon}{3}\right) = 1. \tag{5.74}$$

(5.73) and (5.74) conclude (5.70).

Appendix 5.C Proof of Lemma 5.4.1

Let $\theta(h_n, x_n) = \operatorname{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}\left(x_n + \frac{\alpha^*}{\sqrt{\Delta}} h_n\right) - x_n$. From the law of large numbers, we have

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \psi(\theta(h_n, x_n), x_n) = \mathrm{E}\left[\psi\left(\hat{X} - X, X\right)\right]. \tag{5.75}$$

Hence, it is sufficient to show

$$\operatorname*{plim}_{N\to\infty} \left|\frac{1}{N}\sum_{n=1}^{N} \left\{\psi(\hat{w}_{N,n}(h_n, x_n), x_n) - \psi(\theta(h_n, x_n), x_n)\right\}\right| = 0, \tag{5.76}$$

which is equivalent to

$$\lim_{N\to\infty} \Pr\left(\left|\frac{1}{N}\sum_{n=1}^{N}\left\{\psi(\hat{w}_{N,n}(h_n, x_n), x_n) - \psi(\theta(h_n, x_n), x_n)\right\}\right| < \varepsilon\right)$$
$$= \sum_{\ell=1}^{L} p_\ell \lim_{N\to\infty} \Pr\left(\left|\frac{1}{N}\sum_{n=1}^{N}\left\{\psi(\hat{w}_{N,n}(h_n, r_\ell), r_\ell) - \psi(\theta(h_n, r_\ell), r_\ell)\right\}\right| < \varepsilon\right) \tag{5.77}$$
$$= 1 \tag{5.78}$$

for any ε (> 0). We thus prove

$$\lim_{N\to\infty} \Pr\left(\left|\frac{1}{N}\sum_{n=1}^{N}\left\{\psi(\hat{w}_{N,n}(h_n, r_\ell), r_\ell) - \psi(\theta(h_n, r_\ell), r_\ell)\right\}\right| < \varepsilon\right) = 1 \tag{5.79}$$

for $\ell = 1, \ldots, L$ below.

If we denote the Lipschitz constant of $\psi(\cdot, r_\ell)$ by $C_{\psi,\ell}$, we have

$$\left|\psi(\hat{w}_{N,n}(h_n, r_\ell), r_\ell) - \psi(\theta(h_n, r_\ell), r_\ell)\right| \leq C_{\psi,\ell}\left|\hat{w}_{N,n}(h_n, r_\ell) - \theta(h_n, r_\ell)\right|. \tag{5.80}$$

The absolute value in the right hand side of (5.80) is upper bounded as

$$\left|\hat{w}_{N,n}(h_n, r_\ell) - \theta(h_n, r_\ell)\right|$$
$$\leq \left|\mathrm{prox}_{\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}f}\left(r_\ell + \frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha_N^* h_n\right) - \mathrm{prox}_{\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}}h_n\right)\right|$$
$$+ \left|\mathrm{prox}_{\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}}h_n\right) - \mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}}f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}}h_n\right)\right| \tag{5.81}$$
$$\leq \left|\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha_N^* h_n - \frac{\alpha^*}{\sqrt{\Delta}}h_n\right|$$
$$+ \left|\mathrm{prox}_{\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}}h_n\right) - \mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}}f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}}h_n\right)\right|. \tag{5.82}$$

In (5.82), we use the fact that $\mathrm{prox}_{\gamma f}(\cdot)$ is non-expansive. For the first term in (5.82), we have $\left|\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}\alpha_N^* h_n - \frac{\alpha^*}{\sqrt{\Delta}}h_n\right| \xrightarrow{\mathrm{P}} 0$ as $N \to \infty$. Moreover, given that $\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2}$ is sufficiently

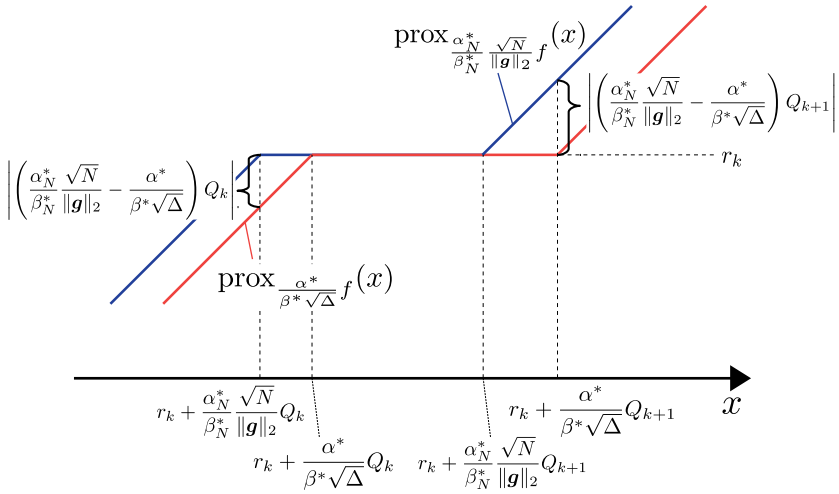


Figure 5.9: Graphical representation for (5.83).

close to $\frac{\alpha^*}{\beta^*\sqrt{\Delta}}$ when N is large, the second term is upper bounded as

$$\left|\mathrm{prox}_{\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2} f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}} h_n\right) - \mathrm{prox}_{\frac{\alpha^*}{\beta^*\sqrt{\Delta}} f}\left(r_\ell + \frac{\alpha^*}{\sqrt{\Delta}} h_n\right)\right|$$

$$\leq \max_{k=2,\ldots,L}\left\{\left|\left(\frac{\alpha_N^*}{\beta_N^*}\frac{\sqrt{N}}{\|\boldsymbol{g}\|_2} - \frac{\alpha^*}{\beta^*\sqrt{\Delta}}\right) Q_k\right|\right\} \tag{5.83}$$

$$\xrightarrow{\mathrm{P}} 0 \tag{5.84}$$

as $N \to \infty$ (See Fig. 5.9). We thus have $|\psi(\hat{w}_n(h_n, r_\ell), r_\ell) - \psi(\theta(h_n, r_\ell), r_\ell)| \xrightarrow{\mathrm{P}} 0$ and obtain (5.79), which completes the proof.

Chapter 6

Discreteness-Aware Approximate Message Passing for Discrete-Valued Vector Reconstruction

6.1 Introduction

As described in Section 1.3, some methods based on convex optimization have been proposed for the large-scale discrete-valued vector reconstruction. The regularization-based method and the transform-based method [57] borrow the idea from compressed sensing [40, 41] in the formulation to obtain convex optimization problems. As for theoretical analysis, the required number of measurements in the large system limit has been derived for the binary vector reconstruction with the regularization-based method. A more general result has been obtained for the reconstruction of uniformly distributed discrete-valued vectors via the transform-based method. For non-uniformly distributed vectors, however, no analytical result has been provided.

On the other hand, the SOAV optimization has been proposed for the reconstruction of discrete-valued vector with any discrete distribution [60]. Although the SOAV optimization is similar to the regularization-based method, it can take the probability distribution of the unknown vector into consideration. They are actually equivalent when the unknown vector is uniformly distributed. Although some theoretical analyses have been provided for the optimization problem [26, 65], the required number of measurements for the reconstruction has not been obtained for the SOAV optimization.

In this chapter, we propose an iterative algorithm based on the SOAV optimization problem and analyze its reconstruction performance. By using the idea of the AMP algorithm [38, 39] for compressed sensing, we firstly consider a probability distribution corresponding to the SOAV optimization. We then approximate the sum-product belief propagation [35, 36] for the distribution and obtain the proposed algorithm, referred

to as DAMP. For the approximation in the derivation, we assume the large system limit, where the problem size increases to infinity with a fixed ratio of the number of measurements to the number of unknown variables. The DAMP algorithm has basically the same form as that of the original AMP algorithm for compressed sensing except for their soft thresholding functions. Hence, the order of the computational complexity is the same as that of the original AMP algorithm. By using state evolution [38, 44], we analytically evaluate the asymptotic performance of the DAMP algorithm in the large system limit. We further derive the required number of measurements for the perfect reconstruction in the noise-free case. The analysis provides the optimal parameters of the soft thresholding function in terms of minimizing the required number of measurements. With the analytical result, we also propose a method to determine the parameters of the SOAV optimization. Moreover, on the basis of the state evolution, we derive Bayes optimal DAMP, which gives the minimum MSE at each iteration in the large system limit. Simulation results show that the proposed DAMP algorithms can reconstruct the discrete-valued vector from its underdetermined linear measurements. For large-scale problems, the performance agrees well with the theoretical result obtained with the state evolution. The SOAV optimization with the proposed parameters can achieve the better performance than that of the original SOAV optimization. Moreover, when the problem size is not large enough, it also outperforms some AMP-based algorithms in high SNR region. We also evaluate the performance when the measurement matrix is a partial discrete cosine transform (DCT) matrix. We compare the proposed methods with turbo compressed sensing [103, 104], which is a message passing-based algorithm designed for partial DFT measurement matrices. For small-scale problems, Bayes optimal DAMP achieves better performance than turbo compressed sensing in the high SNR region.

The rest of the chapter is organized as follows. In Section 6.2, we propose the DAMP algorithm for the discrete-valued vector reconstruction. Section 6.3 analyzes the performance of the DAMP algorithm via the state evolution framework and shows some examples of the analysis. We then apply the theoretical results to the SOAV optimization in Section 6.4 and provide Bayes optimal DAMP in Section 6.5. Section 6.6 gives some simulation results, which demonstrate the performance of the proposed algorithms and show the validity of the theoretical analysis. Finally, we present some conclusions in Section 6.7.

6.2 Proposed Discreteness-aware AMP

In this section, we briefly explain the SOAV optimization [60] and propose DAMP by taking a similar approach to that of the AMP algorithm for compressed sensing [38, 105].

6.2.1 DAMP

The derivation of DAMP begins with belief propagation with the sum-product algorithm [36] for a probability distribution corresponding to the SOAV optimization (1.36). We first consider the following joint probability distribution

$$\mu(\boldsymbol{s}) \propto \prod_{n=1}^{N} \exp\left(-\beta \sum_{\ell=1}^{L} q_\ell \left|s_n - r_\ell\right|\right) \prod_{m=1}^{M} \exp\left\{-\frac{\beta\lambda}{2}\left(y_m - \sum_{j=1}^{N} a_{m,j} s_j\right)^2\right\}, \tag{6.1}$$

where $\beta > 0$. Note that, as $\beta \to \infty$, the mass of the distribution concentrates on the solution of (1.36). Hence, we can solve (1.36) by calculating the mode of the marginal distribution of each x_n, which can be approximated via belief propagation. However, the computational complexity is prohibitive for the factor graph of (6.1) with large N.

To derive a low-complexity algorithm, we then consider the large system limit ($M, N \to \infty$ with fixed $M/N = \Delta$) and large β limit ($\beta \to \infty$), and approximate the sum-product algorithm for (6.1). As in the derivation in [39], assuming the measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ being composed of i.i.d. variables with zero mean and variance $1/M$, we have the resultant algorithm as

$$\boldsymbol{z}^t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^t + \frac{1}{\Delta}\boldsymbol{z}^{t-1}\left\langle \eta'\left(\boldsymbol{x}^{t-1} + \boldsymbol{A}^\mathsf{T}\boldsymbol{z}^{t-1}; \frac{\theta_{t-1}}{\sqrt{\Delta}}\right)\right\rangle, \tag{6.2}$$

$$\boldsymbol{x}^{t+1} = \eta\left(\boldsymbol{x}^t + \boldsymbol{A}^\mathsf{T}\boldsymbol{z}^t; \frac{\theta_t}{\sqrt{\Delta}}\right), \tag{6.3}$$

where $\boldsymbol{x}^t$ is the estimate of $\boldsymbol{x}$ at the tth iteration. The function $\eta(\cdot;\cdot)$ is given by

$$\eta(\boldsymbol{u};\gamma) = \mathrm{prox}_{\gamma J}(\boldsymbol{u}), \tag{6.4}$$

where $J(\boldsymbol{s}) = \sum_{\ell=1}^{L} q_\ell \|\boldsymbol{s} - r_\ell \mathbf{1}\|_1$ is the first term of the objective function in (1.36). By the direct calculation described in [26], the nth element of $\mathrm{prox}_{\gamma J}(\boldsymbol{u})$ is written as

$$[\mathrm{prox}_{\gamma J}(\boldsymbol{u})]_n = \begin{cases} u_n - \gamma Q_1 & (u_n < r_1 + \gamma Q_1) \\ r_1 & (r_1 + \gamma Q_1 \le u_n < r_1 + \gamma Q_2) \\ \vdots & \vdots \\ u_n - \gamma Q_k & (r_{k-1} + \gamma Q_k \le u_n < r_k + \gamma Q_k) \\ r_k & (r_k + \gamma Q_k \le u_n < r_k + \gamma Q_{k+1}) \\ \vdots & \vdots \\ u_n - \gamma Q_{L+1} & (r_L + \gamma Q_{L+1} \le u_n) \end{cases}, \tag{6.5}$$

Algorithm 6.1 DAMP algorithm

Input: $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$
Output: $\hat{\boldsymbol{x}} \in \mathbb{R}^N$

1: $\boldsymbol{x}^0 = \boldsymbol{x}^1 = \mathrm{E}[\boldsymbol{x}]$, $\boldsymbol{z}^0 = \boldsymbol{0}$, $\Delta = M/N$
2: **for** $t = 1$ to $T_{\mathrm{itr}} - 1$ **do**
3: $\quad \boldsymbol{z}^t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^t + \dfrac{1}{\Delta}\boldsymbol{z}^{t-1}\left\langle \eta'\left(\boldsymbol{x}^{t-1} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}^{t-1}; \dfrac{\hat{\theta}_{t-1}}{\sqrt{\Delta}}\right)\right\rangle$
4: $\quad \hat{\theta}_t^2 = \dfrac{\|\boldsymbol{z}^t\|_2^2}{N}$
5: $\quad \boldsymbol{x}^{t+1} = \eta\left(\boldsymbol{x}^t + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{z}^t; \dfrac{\hat{\theta}_t}{\sqrt{\Delta}}\right)$
6: **end for**
7: $\hat{\boldsymbol{x}} = \boldsymbol{x}^{T_{\mathrm{itr}}}$

where u_n is the nth element of $\boldsymbol{u}$, $Q_1 = -\sum_{\ell=1}^{L} q_\ell$, $Q_{L+1} = \sum_{\ell=1}^{L} q_\ell$, and

$$Q_k = \sum_{\ell=1}^{k-1} q_\ell - \sum_{\ell'=k}^{L} q_{\ell'} \quad (k = 2, \ldots, L). \tag{6.6}$$

Since $[\mathrm{prox}_{\gamma J}(\boldsymbol{u})]_n$ is a function of only u_n, the function $\eta(\boldsymbol{u}; \gamma)$ is a element-wise function of $\boldsymbol{u}$. The nth element of $\eta'(\boldsymbol{u}; \gamma)$ in (6.2) is the partial derivative of $\eta(\boldsymbol{u}; \gamma)$ with respect to u_n, and is given by $[\eta'(\boldsymbol{u}; \gamma)]_n = 0$ if $[\mathrm{prox}_{\gamma J}(\boldsymbol{u})]_n \in \{r_1, \ldots, r_L\}$, otherwise $[\eta'(\boldsymbol{u}; \gamma)]_n = 1$. $\theta_t^2 = \|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2/N + \Delta\sigma_{\mathrm{v}}^2$ is a scaled effective variance at the tth iteration [106]. Since the true solution $\boldsymbol{x}$ is unknown in practice, we use the alternative value for θ_t^2, e.g., $\hat{\theta}_t^2 = \|\boldsymbol{z}^t\|_2^2/N$ as in [105].

We summarize the proposed DAMP algorithm in Algorithm 6.1. It should be noted that the update equations of DAMP (6.2), (6.3) are basically the same as those of the AMP algorithm for compressed sensing [38, 39]. The only difference is the function $\eta(\boldsymbol{u}; \gamma)$, which is the soft thresholding function $[\eta(\boldsymbol{u}; \gamma)]_n = \mathrm{sign}(u_n)\max\{|u_n| - \gamma, 0\}$ in the case of the sparse vector reconstruction. Hence, the function $\eta(\boldsymbol{u}; \gamma)$ given by (6.4)–(6.6) can be considered as the soft thresholding function for the discrete-valued vector reconstruction. In TABLE 6.1, we summarize the relationship between the original AMP algorithm and the proposed DAMP algorithm.

From a Bayesian perspective, the original AMP algorithm uses the prior distribution $p_{\mathrm{pri}}(x) \propto \exp(-|x|)$ for the unknown sparse vector, whereas the DAMP algorithm uses $p_{\mathrm{pri}}(x) \propto \exp\left(-\sum_{\ell=1}^{L} q_\ell |x - r_\ell|\right)$ for the discrete-valued vector. Although the DAMP algorithm is based on the idea of the SOAV optimization, the estimate by the DAMP algorithm is not necessarily equal to that by the SOAV optimization because of the approximations in the derivation. For details of the relationship between AMP-based approaches and optimization-based approaches, see [107, 108]. Since (6.2) and (6.3)

Table 6.1: Comparison between the original AMP algorithm and the proposed DAMP algorithm

	original AMP algorithm [38, 39]	proposed DAMP algorithm
target	sparse vector	discrete-valued vector
optimization problem	ℓ_1 optimization	SOAV optimization in (1.36)
prior distribution $p_{\mathrm{pri}}(x)$	$p_{\mathrm{pri}}(x) \propto \exp(-\lvert x\rvert)$	$p_{\mathrm{pri}}(x) \propto \exp\left(-\sum_{\ell=1}^{L} q_\ell \lvert x - r_\ell\rvert\right)$
soft thresholding function	$[\eta(\boldsymbol{u};\gamma)]_n = \mathrm{sign}(u_n)\max\{\lvert u_n\rvert - \gamma, 0\}$	$\eta(\boldsymbol{u};\gamma) = \mathrm{prox}_{\gamma J}(\boldsymbol{u})$ in (6.5)

can be computed only with additions of vectors and multiplications of a matrix and a vector, the computational complexity of the algorithm is $O(MN)$ per iteration, which is lower than that of internal point methods $O(MN^2)$ used in [57].

The DAMP algorithm can also be used for complex-valued vector by rewriting the complex-valued model into the equivalent real-valued model when the real and imaginary parts are independent, e.g., $\boldsymbol{x} \in \{1+j, -1+j, -1-j, 1-j\}^N$. When they are dependent, however, the algorithm cannot be directly applied and hence some extensions are required.

6.3 Asymptotic Analysis of DAMP

In this section, we provide a theoretical analysis of DAMP with state evolution framework [38], [44]. By using state evolution, we give the required number of measurements for the perfect reconstruction and the parameter of the soft thresholding function minimizing the required number of measurements in the large system limit.

6.3.1 State Evolution

State evolution is a framework to analyze the asymptotic performance of the AMP algorithm. In the large system limit, the sample MSE $\sigma_t^2 = \|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2/N$ of $\boldsymbol{x}^t$ can be predicted via the state evolution. Similarly to the case of compressed sensing, the state evolution formula for DAMP in Algorithm 6.1 is written as

$$\sigma_{t+1}^2 = \Psi_{\mathrm{SE}}\left(\sigma_t^2 + \Delta\sigma_{\mathrm{v}}^2\right), \tag{6.7}$$

where

$$\Psi_{\mathrm{SE}}\left(\sigma^2\right) = \mathrm{E}\left[\left\{\eta\left(X + \frac{\sigma}{\sqrt{\Delta}}Z; \frac{\sigma}{\sqrt{\Delta}}\right) - X\right\}^2\right]. \tag{6.8}$$

The random variable X has the same distribution as that of the unknown discrete variable, i.e., $\Pr(X = r_\ell) = p_\ell$ $(\ell = 1, \ldots, L)$ in our problem, and Z is the standard Gaussian

random variable independent of X. In the rigorous proof for the state evolution [44], it is assumed that $\boldsymbol{A}$ is composed of i.i.d. Gaussian variables with zero mean and variance $1/M$, and $\eta(\cdot;\cdot)$ is Lipschitz continuous. In [44], however, it is expected that the state evolution is also valid for a broader class of measurement matrices $\boldsymbol{A}$, such as the matrices with i.i.d. (possibly non-Gaussian) elements with zero mean and variance $1/M$. In fact, some numerical results in [38] imply such universality of the state evolution.

6.3.2 Condition for Perfect Reconstruction by DAMP

We can analyze the performance of the DAMP algorithm by investigating the function $\Psi_{\mathrm{SE}}(\sigma^2)$, which can be analytically obtained (See Appendix 6.A). In this section, we consider the noise-free case (i.e., $\sigma_{\mathrm{v}}^2 = 0$) and investigate a sufficient condition for the perfect reconstruction defined as $\sigma_t^2 \to 0$ $(t \to \infty)$. Since we have $\Psi_{\mathrm{SE}}(0) = 0$ in the noise-free case, the sequence $\{\sigma_t^2\}_{t=0,1,\ldots}$ with the recursion (6.7) converges to zero if $\Psi_{\mathrm{SE}}(\sigma^2)$ is concave and its derivative at $\sigma^2 = 0$ is smaller than one, i.e., $\left.\dfrac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0} < 1$. In fact, the condition $\left.\dfrac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0} < 1$ results in $\Psi_{\mathrm{SE}}(\sigma^2) < \sigma^2$ and hence we have $\sigma_{t+1}^2 = \Psi_{\mathrm{SE}}(\sigma_t^2) < \sigma_t^2$. In this case, DAMP reconstructs the unknown vector $\boldsymbol{x}$ perfectly regardless of the initialization. Note that the above discussion is valid when the function $\Psi_{\mathrm{SE}}(\sigma^2)$ is concave, and we can examine the concavity as discussed later.

To obtain the condition for the perfect reconstruction, we evaluate $\left.\dfrac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0}$ analytically. By the mathematical manipulation, we have

$$\left.\frac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0} := D(\boldsymbol{Q}) \tag{6.9}$$

$$= \frac{1}{\Delta}\sum_{\ell=1}^{L} p_\ell \left\{ Q_\ell p_{\mathrm{G}}(Q_\ell) - Q_{\ell+1} p_{\mathrm{G}}(Q_{\ell+1}) + \left(1+Q_\ell^2\right) P_{\mathrm{G}}(Q_\ell) \right. \\ \left. + \left(1+Q_{\ell+1}^2\right)\left(1 - P_{\mathrm{G}}(Q_{\ell+1})\right) \right\}, \tag{6.10}$$

where $\boldsymbol{Q} = [Q_1 \ \cdots \ Q_{L+1}]^{\mathsf{T}}$ (See Appendix 6.B). Since we can choose any $q_1, \ldots, q_L \geq 0$ in (1.36), we minimize (6.10) with respect to $Q_1, \ldots, Q_{L+1}$ as

$$D_{\min} = \min_{\boldsymbol{Q}} D(\boldsymbol{Q}) \text{ subject to } Q_1 \leq \cdots \leq Q_{L+1}. \tag{6.11}$$

Note that, in (6.11), we eliminate the constraint $Q_1 = -Q_{L+1}$. As we will see later, the optimal values of Q_1 and Q_{L+1} are $Q_1^{\mathrm{opt}} = -\infty$ and $Q_{L+1}^{\mathrm{opt}} = \infty$, respectively, and hence this relaxation does not change the optimal value $D_{\min}$. The optimization problem (6.11)

Algorithm 6.2 Parameter optimization

Input: $p_1, \ldots, p_L$

Output: $\boldsymbol{Q}^{\mathrm{opt}} = \left[Q_1^{\mathrm{opt}} \ \cdots \ Q_{L+1}^{\mathrm{opt}}\right]^{\mathsf{T}}$

for $\ell = L, L-1, \ldots, 2$ **do**
  $F_\ell(Q) = p_{\ell-1}\left\{-p_{\mathrm{G}}(Q) + Q\left(1 - P_{\mathrm{G}}(Q)\right)\right\} + p_\ell\left\{p_{\mathrm{G}}(Q) + QP_{\mathrm{G}}(Q)\right\}$
end for
$Q_{L+1}^{\mathrm{opt}} = \infty$
$G_L(Q) = F_L(Q)$
for $\ell = L, L-1, \ldots, 3$ **do**
  if $\hat{Q}(G_\ell(Q)) > \max_{j=2,\ldots,\ell-1} \hat{Q}\left(\sum_{k=j}^{\ell-1} F_k(Q)\right)$ **then**
    $Q_\ell^{\mathrm{opt}} = \hat{Q}(G_\ell(Q))$
    $G_{\ell-1}(Q) = F_{\ell-1}(Q)$
  else
    $Q_\ell^{\mathrm{opt}} = Q_{\ell-1}^{\mathrm{opt}}$
    $G_{\ell-1}(Q) = F_{\ell-1}(Q) + G_\ell(Q)$
  end if
end for
$Q_2^{\mathrm{opt}} = \hat{Q}(G_2(Q))$
$Q_1^{\mathrm{opt}} = -\infty$

can be solved via interior point methods [109] because $D(\boldsymbol{Q})$ is a convex function of $\boldsymbol{Q}$. We can also solve (6.11) with the following theorem, which enables us to theoretically analyze the performance of DAMP in some cases as described in Section 6.3.3. In what follows, for an equation $h(Q) = 0$ with a unique solution, we denote the solution by $\hat{Q}(h(Q))$, i.e., $h(\hat{Q}(h(Q))) = 0$.

Theorem 6.3.1. The unique minimizer $\boldsymbol{Q}^{\mathrm{opt}} = \left[Q_1^{\mathrm{opt}} \ \cdots \ Q_{L+1}^{\mathrm{opt}}\right]^{\mathsf{T}}$ of the optimization problem (6.11) can be obtained by Algorithm 6.2.

Proof. See Appendix 6.C. □

By using Algorithm 6.2, we can obtain the unique minimizer $\boldsymbol{Q}^{\mathrm{opt}}$ and the corresponding minimum value $D_{\min}$ of $\left.\dfrac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0}$. From (6.5), the soft thresholding function

with the optimal parameters Q_ℓ^{opt} is written as

$$\left[\eta^{\mathrm{S}}\left(\boldsymbol{u};\gamma\right)\right]_n = \begin{cases} r_1 & \left(u_n < r_1 + \gamma Q_2^{\mathrm{opt}}\right) \\ \vdots & \vdots \\ u_n - \gamma Q_k^{\mathrm{opt}} & \left(r_{k-1} + \gamma Q_k^{\mathrm{opt}} \le u_n < r_k + \gamma Q_k^{\mathrm{opt}}\right) \\ r_k & \left(r_k + \gamma Q_k^{\mathrm{opt}} \le u_n < r_k + \gamma Q_{k+1}^{\mathrm{opt}}\right) \\ \vdots & \vdots \\ r_L & \left(r_L + \gamma Q_L^{\mathrm{opt}} \le u_n\right) \end{cases}, \tag{6.12}$$

where $\left[\eta^{\mathrm{S}}\left(\boldsymbol{u};\gamma\right)\right]_n$ denotes the nth element of $\eta^{\mathrm{S}}(\boldsymbol{u};\gamma)$.

The DAMP algorithm with $\eta^{\mathrm{S}}(\cdot;\cdot)$, which we call *soft thresholding DAMP* henceforth, provides the perfect reconstruction in the large system limit if $D_{\min} < 1$ and the function $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2) = \mathrm{E}\left[\left\{\eta^{\mathrm{S}}\left(X + \frac{\sigma}{\sqrt{\Delta}}Z; \frac{\sigma}{\sqrt{\Delta}}\right) - X\right\}^2\right]$ is concave. Since we have

$$\frac{d^2\Psi_{\mathrm{SE}}}{d(\sigma^2)^2} = \frac{\sqrt{\Delta}}{2\sigma^5}\sum_{\ell=1}^{L} p_\ell \sum_{k=1}^{L} \left(-r_\ell + r_k\right)^3 \left\{-p_{\mathrm{G}}\left(T_{\ell,k,k}\right) + p_{\mathrm{G}}\left(T_{\ell,k,k+1}\right)\right\}, \tag{6.13}$$

where $T_{\ell,k,k'} = \frac{\sqrt{\Delta}}{\sigma}\left(-r_\ell + r_k\right) + Q_{k'}$ (See Appendix 6.D), the concavity of $\Psi_{\mathrm{SE}}(\sigma^2)$ depends on Q_ℓ. By evaluating (6.13) with $\boldsymbol{Q} = \boldsymbol{Q}^{\mathrm{opt}}$, we can investigate whether $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2)$ is concave or not. If $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2)$ is concave and $D_{\min} < 1$, soft thresholding DAMP can perfectly reconstruct the discrete-valued vector in the large system limit.

6.3.3 Examples of Asymptotic Analysis

We show three examples of the analysis for DAMP via state evolution. Although it is difficult to prove the concavity of $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2)$ by the direct calculation in general, we can confirm that $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2)$ is concave in the following examples.

Example 6.3.1 (Binary vector)**.** As the simplest example, we firstly consider the reconstruction of a binary vector $\boldsymbol{x} \in \{r_1, r_2\}^N$ with $\Pr(x_n = r_1) = p_1$ and $\Pr(x_n = r_2) = p_2$ $(= 1 - p_1)$. The binary vector reconstruction appears in CDMA multiuser detection and signal detection for MIMO systems with BPSK or QPSK. By using Algorithm 6.2, we can obtain the optimal parameters of the soft thresholding function.

In the noise-free case, soft thresholding DAMP provides the perfect reconstruction in the large system limit if $D_{\min} < 1$. Figure 6.1 shows the phase transition line of soft thresholding DAMP, where $D_{\min} = 1$. Note that the line is the boundary between the

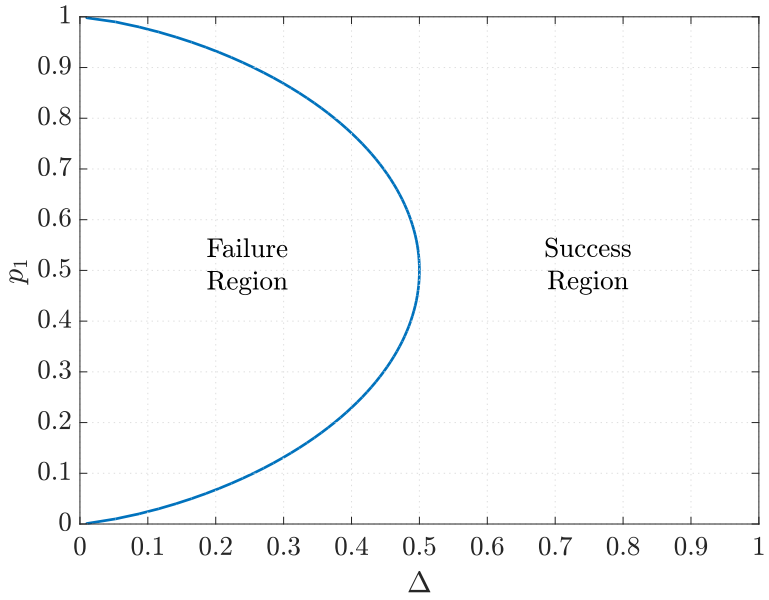


Figure 6.1: Phase transition of soft thresholding DAMP for binary vector

success and failure regions of DAMP in the large system limit. In the left region of the curve, the MSE of the estimate obtained by DAMP does not converge to zero. In the right region, DAMP can provide the perfect reconstruction of $\boldsymbol{x}$. For example, the figure shows that DAMP requires at least $N/2$ observations to accurately reconstruct an N-dimensional uniformly distributed binary vector with $p_1 = 0.5$. This result coincides with the theoretical analysis for the regularization-based method and the transform-based method [57] as well as the box relaxation [95]. Moreover, our analysis also provides the required number of measurements for the asymmetric distribution with $p_1 \neq 0.5$, which has not been obtained in [57] and [95]. It should be noted that $D_{\min}$ in (6.11) is independent of r_1 and r_2, and hence the phase transition line in Fig. 6.1 is identical for any r_1 and r_2 in the noise-free case.

In the noisy case, the asymptotic MSE at the fixed point of soft thresholding DAMP, i.e., the value of σ^2 satisfying $\sigma^2 = \Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2 + \Delta\sigma_{\mathrm{v}}^2)$, can be obtained numerically by iterating $\sigma_{t+1}^2 = \Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma_t^2 + \Delta\sigma_{\mathrm{v}}^2)$. Figure 6.2 shows the result for the binary vector $\boldsymbol{x} \in \{-1, 1\}^N$ with $\Pr(x_n = -1) = p_1$, $\Pr(x_n = 1) = 1 - p_1$, and $\sigma_{\mathrm{v}}^2 = 0.01$. We can see that the asymptotic MSE becomes smaller when the measurement ratio Δ increases.

Example 6.3.2 (Possibly sparse discrete-valued vector)**.** The reconstruction of a possibly sparse discrete-valued vector, such as $\boldsymbol{x} \in \{-1, 0, 1\}^N$ and $\boldsymbol{x} \in \{-3, -1, 0, 1, 3\}^N$, also arises in some problems, e.g., multiuser detection for machine-to-machine communications [26] and error recovery for MIMO signal detection [65]. Although some methods have been proposed for the reconstruction of the discrete-valued sparse vector [110–114],

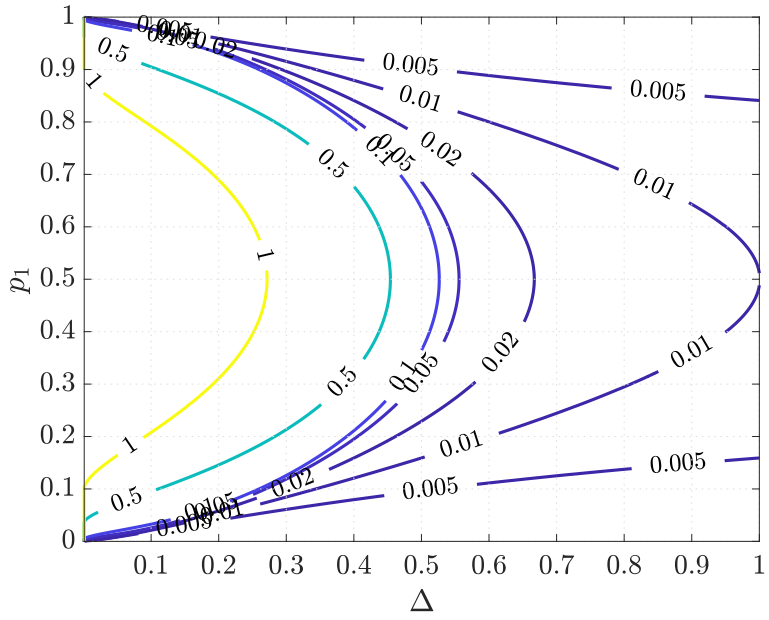


Figure 6.2: MSE at the fixed point in the noisy case ($\boldsymbol{x} \in \{-1, 1\}^N$, $\Pr(x_n = -1) = p_1$, $\Pr(x_n = 1) = 1 - p_1$, $\sigma_{\mathrm{v}}^2 = 0.01$)

their theoretical analyses have not been obtained.

In Fig. 6.3, we show the phase transition line for $\boldsymbol{x} \in \{-r, 0, r\}^N$ ('binary') and $\boldsymbol{x} \in \{-3r, -r, 0, r, 3r\}^N$ ($r > 0$) ('quad') in the noise-free case. For $\boldsymbol{x} \in \{-r, 0, r\}^N$, we assume $\Pr(x_n = 0) = p$ and $\Pr(x_n = -r) = \Pr(x_n = r) = (1 - p)/2$. For $\boldsymbol{x} \in \{-3r, -r, 0, r, 3r\}^N$, we assume $\Pr(x_n = 0) = p$ and $\Pr(x_n = -3r) = \Pr(x_n = -r) = \Pr(x_n = r) = \Pr(x_n = 3r) = (1 - p)/4$. The dashed line '$\ell_1$' shows the phase transition line of the original AMP algorithm for compressed sensing, which utilizes only the sparsity of the unknown vector. If the unknown vector is discrete-valued, the DAMP algorithm requires a less number of measurements compared to the AMP algorithm. However, as the possible candidates for non-zero value increases, more number of measurements is required for the perfect reconstruction.

Example 6.3.3 (Uniformly distributed discrete-valued vector)**.** Finally, we analyze the reconstruction of $\boldsymbol{x} \in \{r_1, \ldots, r_L\}^N$ with the uniform distribution $p_1 = \cdots = p_L = 1/L$. The signal detection for MIMO systems with quadrature amplitude modulation (QAM) can be reduced to such reconstruction problem.

By Algorithm 6.2, we have $Q_L^{\mathrm{opt}} = \cdots = Q_2^{\mathrm{opt}} = 0$. The resultant soft thresholding function is equivalent to that of the AMP algorithm with the box relaxation [115]. The condition for the perfect reconstruction in the noise-free case is $D_{\min} = (L - 1)/(\Delta L) < 1 \Leftrightarrow \Delta > (L - 1)/L$, which means that soft thresholding DAMP requires more than $(L - 1)N/L$ measurements to reconstruct a N dimensional vector with the uniform

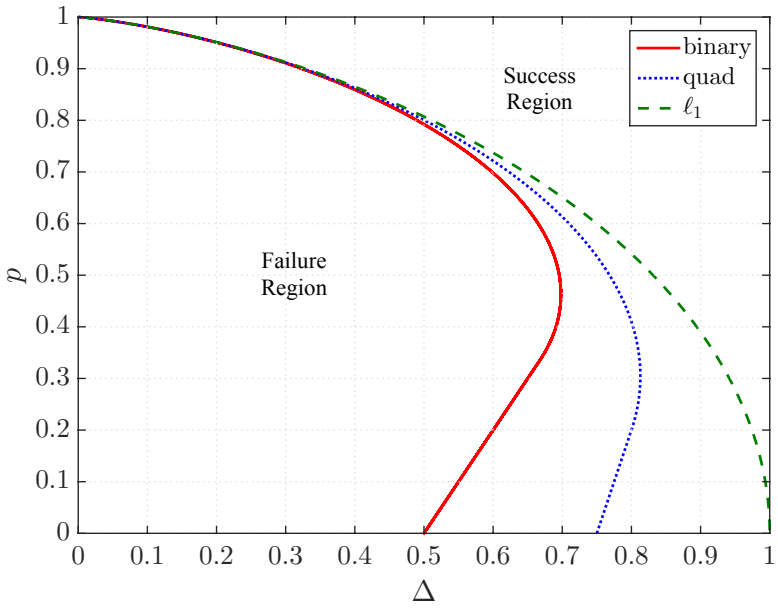


Figure 6.3: Phase transition of soft thresholding DAMP for possibly sparse discrete-valued vector

distribution of L values. This threshold is the same as that for the transform-based method [57]. It should be noted that the analysis of DAMP can also be applied even for non-uniform distribution, while uniform distributions are assumed for the analyses in [57] and [115].

6.4 Application to SOAV Optimization

In this section, we propose a method to determine the parameters q_ℓ of the SOAV optimization (1.36) on the basis of the asymptotic analysis of soft thresholding DAMP.

The DAMP algorithm proposed in the previous section has low computational complexity and its asymptotic performance can be predicted by state evolution, which provides the optimal parameters of the soft thresholding function. In the derivation, however, we take the large system limit and assume that the measurement matrix $\boldsymbol{A}$ is composed of i.i.d. elements. Hence, the DAMP algorithm suffers from the performance degradation when the problem size is not large or the measurement matrix is composed of correlated elements.

The SOAV optimization can be solved with the proximal splitting methods [61] as a convex optimization problem [26]. The convex optimization algorithms do not require any assumptions on the measurement matrix and can obtain the minimizer even when

the problem size is small. Thus, when the measurement matrix is composed of correlated elements or the problem size is not large enough, the convex optimization-based approach using parameters $q_1, \ldots, q_L$ obtained from the optimal values $Q_1^{\mathrm{opt}}, \ldots, Q_L^{\mathrm{opt}}$ in the previous section might outperform the DAMP algorithm.

We thus derive the parameters q_ℓ^{opt} corresponding to the optimal parameters Q_ℓ^{opt} in a casual manner. From the definitions of Q_k in (6.6), we can obtain q_ℓ from Q_ℓ and $Q_{\ell+1}$ as $q_\ell = (-Q_\ell + Q_{\ell+1})/2$. Since $Q_2^{\mathrm{opt}}, \ldots, Q_L^{\mathrm{opt}}$ are finite, the corresponding coefficients $q_2^{\mathrm{opt}}, \ldots, q_{L-1}^{\mathrm{opt}}$ given by

$$q_\ell^{\mathrm{opt}} = \frac{1}{2}\left(-Q_\ell^{\mathrm{opt}} + Q_{\ell+1}^{\mathrm{opt}}\right) \tag{6.14}$$

are also finite. On the other hand, $q_1^{\mathrm{opt}} = q_L^{\mathrm{opt}} = \infty$ follows from $Q_1^{\mathrm{opt}} = -\infty$ and $Q_{L+1}^{\mathrm{opt}} = \infty$. The objective function of the SOAV optimization includes q_1^{opt} and q_L^{opt} in the form $q_1^{\mathrm{opt}}|s - r_1| + q_L^{\mathrm{opt}}|s - r_L|$, and the term becomes infinity when $s \le r_1$ or $s \ge r_L$. When $r_1 < s < r_L$, however, the term is computed as

$$\begin{aligned} q_1^{\mathrm{opt}}|s - r_1| + q_L^{\mathrm{opt}}|s - r_L| &= (q_1^{\mathrm{opt}} - q_L^{\mathrm{opt}})s + \text{const.} && (6.15)\\ &= \frac{1}{2}(Q_2^{\mathrm{opt}} + Q_L^{\mathrm{opt}})s + \text{const.} && (6.16) \end{aligned}$$

because we have

$$Q_2 + Q_L = 2(q_1 - q_L) \tag{6.17}$$

from (6.6), where "const." is a constant independent of s. Since Q_2^{opt} and Q_L^{opt} are finite, $Q_2^{\mathrm{opt}} + Q_L^{\mathrm{opt}}$ is also finite and hence we have

$$q_1^{\mathrm{opt}}|s - r_1| + q_L^{\mathrm{opt}}|s - r_L| = \begin{cases} \dfrac{1}{2}(Q_2^{\mathrm{opt}} + Q_L^{\mathrm{opt}})s + \text{const.} & (r_1 < s < r_L) \\ \infty & (\text{otherwise}) \end{cases}, \tag{6.18}$$

where the infinity for $s \notin (r_1, r_L)$ corresponds to the box constraint $r_1 < s < r_L$. Therefore, we have the SOAV optimization problem corresponding to the optimal parameters for soft thresholding DAMP as

$$\begin{aligned} \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{s}\in\mathbb{R}^N} &\left\{ \sum_{\ell=1}^{L} q_\ell^{\mathrm{opt}} \|\boldsymbol{s} - r_\ell \mathbf{1}\|_1 + \frac{\lambda}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 \right\} \\ &\text{subject to } \; r_1\mathbf{1} \le \boldsymbol{s} \le r_L\mathbf{1}, \end{aligned} \tag{6.19}$$

where $q_2^{\mathrm{opt}}, \ldots, q_{L-1}^{\mathrm{opt}}$ are given by (6.14) and q_1^{opt}, q_L^{opt} (≥ 0) must be chosen to satisfy $q_1^{\mathrm{opt}} - q_L^{\mathrm{opt}} = (Q_2^{\mathrm{opt}} + Q_L^{\mathrm{opt}})/2$. Note that we relax the constraint as $r_1\mathbf{1} \le \boldsymbol{s} \le r_L\mathbf{1}$ because

Algorithm 6.3 Beck-Teboulle proximal gradient algorithm for SOAV optimization (6.19)

Input: $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, $\lambda \in \mathbb{R}$

Output: $\boldsymbol{x}^{T_{\mathrm{itr}}} \in \mathbb{R}^N$

1: $\boldsymbol{x}^1 = \boldsymbol{0}$, $\boldsymbol{z}^1 = \boldsymbol{0}$, $\tau_1 = 1$, $\gamma^{-1} \geq \lambda \|\boldsymbol{A}\|_2^2$

2: **for** $t = 1$ to $T_{\mathrm{itr}} - 1$ **do**

3: $\quad \boldsymbol{x}^{t+1} = \eta^{\mathrm{S}}\left(\boldsymbol{z}^t + \gamma\lambda \boldsymbol{A}^{\mathsf{T}}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}^t\right); \gamma\right)$

4: $\quad \tau_{t+1} = \dfrac{1 + \sqrt{4\tau_t^2 + 1}}{2}$

5: $\quad \omega_t = 1 + \dfrac{\tau_t - 1}{\tau_{t+1}}$

6: $\quad \boldsymbol{z}^{t+1} = \omega_t \boldsymbol{x}^{t+1} + (1 - \omega_t)\, \boldsymbol{x}^t$

7: **end for**

the unknown vector $\boldsymbol{x}$ may have r_1 and r_L. The problem (6.19) can be rewritten as

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{s} \in \mathbb{R}^N} \left\{ h_1(\boldsymbol{s}) + h_2(\boldsymbol{s}) \right\}, \tag{6.20}$$

where $h_1(\boldsymbol{s}) = \sum_{\ell=1}^{L} q_\ell^{\mathrm{opt}} \|\boldsymbol{s} - r_\ell \boldsymbol{1}\|_1 + \iota(\boldsymbol{s})$, $h_2(\boldsymbol{s}) = \lambda \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 / 2$, and

$$\iota(\boldsymbol{s}) = \begin{cases} 0 & (r_1 \boldsymbol{1} \leq \boldsymbol{s} \leq r_L \boldsymbol{1}) \\ \infty & (\text{otherwise}) \end{cases}. \tag{6.21}$$

An important fact here is that the proximity operator of $h_1(\boldsymbol{x})$ is given by $\mathrm{prox}_{\gamma h_1}(\boldsymbol{u}) = \eta^{\mathrm{S}}(\boldsymbol{u}; \gamma)$ because $q_1^{\mathrm{opt}}, \ldots, q_L^{\mathrm{opt}}, Q_1^{\mathrm{opt}}, \ldots, Q_L^{\mathrm{opt}}$ satisfy (6.6) and $\iota(\cdot)$ restricts the value of $\mathrm{prox}_{\gamma h_1}(\boldsymbol{u})$ as $r_1 \leq [\mathrm{prox}_{\gamma h_1}(\boldsymbol{u})]_n \leq r_L$. Hence, the convex optimization problem (6.19) can be efficiently solved by proximal splitting methods [61] using $\eta^{\mathrm{S}}(\boldsymbol{u}; \gamma)$. As an example, we show Beck-Teboulle proximal gradient algorithm [61, 62] for the optimization problem (6.20) in Algorithm 6.3.

As we can see from the following example, the proposed parameters q_ℓ^{opt} are different from those of the original SOAV optimization in general.

Example 6.4.1. We consider the reconstruction of $\boldsymbol{x} \in \{-1, 0, 1\}^N$. The distribution of $\boldsymbol{x}$ is assumed to be $\Pr(x_n = 0) = 0.2$ and $\Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$. In this case, we have $Q_1^{\mathrm{opt}} = -\infty$, $Q_2^{\mathrm{opt}} = Q_3^{\mathrm{opt}} = 0$, and $Q_4^{\mathrm{opt}} = \infty$ by Algorithm 6.2. Hence, the proposed parameters satisfy $q_1^{\mathrm{opt}} - q_3^{\mathrm{opt}} = 0$ and $q_2^{\mathrm{opt}} = 0$. Since we have $q_1^{\mathrm{opt}} \|\boldsymbol{s} + \boldsymbol{1}\|_1 + q_2^{\mathrm{opt}} \|\boldsymbol{s}\|_1 + q_3^{\mathrm{opt}} \|\boldsymbol{s} - \boldsymbol{1}\|_1 = 2q_1^{\mathrm{opt}} N$ (= const.) for $-\boldsymbol{1} \leq \boldsymbol{s} \leq \boldsymbol{1}$ in this case, the proposed optimization problem is given by

$$\begin{aligned} \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{s} \in \mathbb{R}^N} & \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 \\ \text{subject to} & \ -\boldsymbol{1} \leq \boldsymbol{s} \leq \boldsymbol{1}. \end{aligned} \tag{6.22}$$

The optimization problem (6.22) is quite different from the original SOAV optimization [60] and the regularization-based method [57], where $(q_1, q_2, q_3) = (0.4, 0.2, 0.4)$ and $(q_1, q_2, q_3) = (1, 1, 1)$, respectively. Note that the box relaxation optimization (6.22) has been considered for the reconstruction of the binary vector $\boldsymbol{x} \in \{-1, 1\}^N$ (e.g., [54, 95]). The proposed approach results in the box relaxation optimization (6.22) even when $\Pr(x_j = 0) = 0.2$.

6.5 Bayes optimal DAMP

In this section, we provide Bayes optimal DAMP on the basis of the state evolution. In the DAMP algorithm in Algorithm 6.1, we can use different functions as $\eta(\cdot;\cdot)$ instead of the soft thresholding function (6.5), (6.12). Moreover, the state evolution formula (6.7) is still valid for different $\eta(\cdot;\cdot)$ as far as it is Lipschitz continuous. In the literature of compressed sensing, the AMP algorithm is called *Bayes optimal* if the function $\left[\eta^{\mathrm{B}}(\boldsymbol{u};\gamma)\right]_n = \mathrm{E}\left[X \mid X + \gamma Z = u_n\right]$ is used instead of the soft thresholding function [105, 116]. x Note that $\eta^{\mathrm{B}}\left(\cdot; \dfrac{\sigma}{\sqrt{\Delta}}\right)$ is the minimizer of $\tilde{\Psi}_{\mathrm{SE}}(\sigma^2) = \mathrm{E}\left[\left\{\tilde{\eta}\left(X + \dfrac{\sigma}{\sqrt{\Delta}}Z\right) - X\right\}^2\right]$ when we consider $\tilde{\Psi}_{\mathrm{SE}}(\sigma^2)$ as the functional of a function $\tilde{\eta}(\cdot)$.

Although it is difficult in general to analytically calculate the optimal function $\eta^{\mathrm{B}}(\cdot;\cdot)$, we can obtain $\eta^{\mathrm{B}}(\cdot;\cdot)$ for the Bayes optimal DAMP because the distribution of X is discrete in our problem. The conditional probability of X can be written as

$$\Pr\left(X = r_\ell \mid X + \gamma Z = u_n\right) = \frac{1}{\zeta} p_\ell p_{\mathrm{G}}\left(\frac{u_n - r_\ell}{\gamma}\right), \tag{6.23}$$

where the normalizing constant ζ is given by $\zeta = \displaystyle\sum_{\ell=1}^{L} p_\ell p_{\mathrm{G}}\left(\frac{u_n - r_\ell}{\gamma}\right)$. From (6.23), we have

$$\left[\eta^{\mathrm{B}}(\boldsymbol{u};\gamma)\right]_n = \frac{\displaystyle\sum_{\ell=1}^{L} p_\ell r_\ell p_{\mathrm{G}}\left(\frac{u_n - r_\ell}{\gamma}\right)}{\displaystyle\sum_{\ell'=1}^{L} p_{\ell'} p_{\mathrm{G}}\left(\frac{u_n - r_{\ell'}}{\gamma}\right)}. \tag{6.24}$$

As a special case, when $r_1 = -1, r_2 = 1$ and $p_1 = p_2 = 0.5$, (6.24) can be reduced to $\left[\eta^{\mathrm{B}}(\boldsymbol{u};\gamma)\right]_n = \tanh\left(u_n/\gamma^2\right)$, which has been proposed for CDMA multiuser detection [37, 117].

The state evolution formula for Bayes optimal DAMP is given by $\sigma_{t+1}^2 = \Psi_{\mathrm{SE}}^{\mathrm{B}}(\sigma_t^2 + \Delta\sigma_{\mathrm{v}}^2)$, where $\Psi_{\mathrm{SE}}^{\mathrm{B}}(\sigma^2) = \mathrm{E}\left[\left\{\eta^{\mathrm{B}}\left(X + \dfrac{\sigma}{\sqrt{\Delta}}Z; \dfrac{\sigma}{\sqrt{\Delta}}\right) - X\right\}^2\right]$. Since $\eta^{\mathrm{B}}\left(\cdot; \dfrac{\sigma}{\sqrt{\Delta}}\right)$ is the

minimizer of $\tilde{\Psi}_{\mathrm{SE}}(\sigma^2)$, Bayes optimal DAMP provides the minimum MSE at each iteration in the large system limit. In the noise-free case, the sequence of the MSE $\{\sigma_t^2\}_{t=0,1,\ldots}$ obtained by $\sigma_{t+1}^2 = \Psi_{\mathrm{SE}}^{\mathrm{B}}(\sigma_t^2)$ converges to zero if $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2)$ is concave and $D_{\min} < 1$, because $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2) < \sigma^2$ in that case and hence $\sigma_{t+1}^2 = \Psi_{\mathrm{SE}}^{\mathrm{B}}(\sigma_t^2) \leq \Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma_t^2) < \sigma_t^2$. Thus, the required measurement ratio Δ for soft thresholding DAMP is an upper bound of that for Bayes optimal DAMP. However, since $\Psi_{\mathrm{SE}}^{\mathrm{B}}(\sigma^2)$ is not necessarily concave unlike $\Psi_{\mathrm{SE}}^{\mathrm{S}}(\sigma^2)$, it is difficult to obtain the necessary condition analytically for the perfect reconstruction by Bayes optimal DAMP.

A similar algorithm to Bayes optimal DAMP can be derived by using the discrete prior distribution in GAMP [43, 118] with scalar variances. The AMP-based algorithm similar to Bayes optimal DAMP has also been proposed for MIMO signal detection [42], where the reconstruction of complex discrete-valued vectors with uniform distributions is considered. However, these algorithms use update equations to obtain the effective variance, while the proposed DAMP algorithm uses the simple estimation $\hat{\theta}_t^2 = \|\boldsymbol{z}^t\|_2^2/N$. These conventional algorithms use the knowledge of the noise variance σ_{v}^2 unlike Bayes optimal DAMP.

6.6 Simulation Results

In this section, we evaluate the performance of the proposed algorithms via computer simulations.

Figure 6.4 shows the prediction of MSE via state evolution and the empirical MSE $\sigma_t^2 = \|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2/N$ with DAMP obtained by simulations. We set $\boldsymbol{x} \in \{-1, 1\}^N$, $\Pr(x_n = -1) = 0.2$, $\Pr(x_n = 1) = 0.8$, $\Delta = 0.5$, and $\sigma_{\mathrm{v}}^2 = 0$. We evaluate the performance for the different problem sizes of $N = 100, 500, 1000$, and 5000. The measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ is composed of i.i.d. Gaussian variables with zero mean and variance $1/M$. In the figure, "soft thresholding" denotes the performance of DAMP with the soft thresholding function $\eta^{\mathrm{S}}(\cdot;\cdot)$ and "Bayes optimal" denotes that of Bayes optimal DAMP with $\eta^{\mathrm{B}}(\cdot;\cdot)$. We can see that Bayes optimal DAMP has much smaller MSE with less number of iterations than soft thresholding DAMP. The figure also shows that the prediction with state evolution is close to the empirical performance in the large-scale systems.

In Figs. 6.5 and 6.6, we evaluate the average of SER defined as $\|Q(\boldsymbol{x}^t) - \boldsymbol{x}\|_0/N$, where $Q(\hat{\boldsymbol{x}}) = \arg\min_{\boldsymbol{s}\in\{r_1,\ldots,r_L\}^N} \|\boldsymbol{s} - \hat{\boldsymbol{x}}\|_1$. The distribution of the unknown vector is $\Pr(x_n = 0) = 0.2, \Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$, which has been considered in Example 6.4.1. The problem size is $(N, M) = (1000, 800)$ in Fig. 6.5 and $(N, M) = (100, 80)$ in Fig. 6.6. The measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ is composed of i.i.d. Gaussian variables with zero mean and variance $1/M$, and the SNR is defined as $N(1-p)/\left(M\sigma_{\mathrm{v}}^2\right)$. The number of iterations in the algorithms is fixed to 200. In the figures, "STDAMP" and

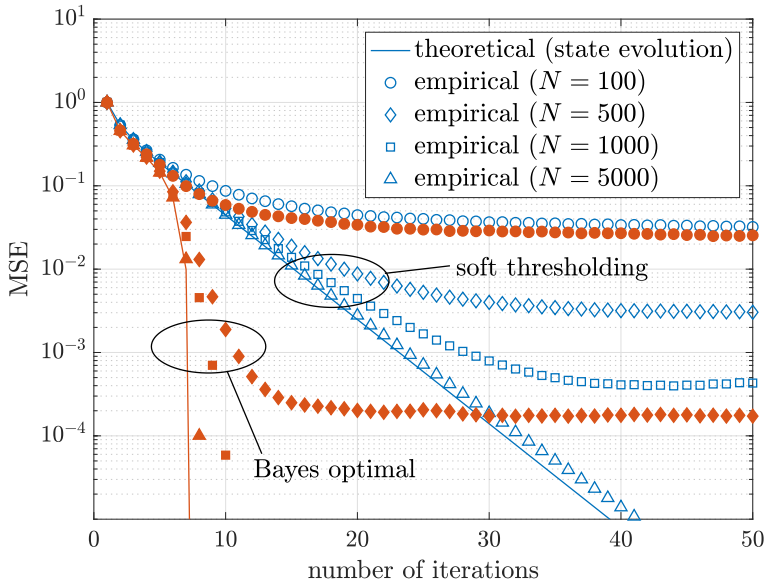


Figure 6.4: State evolution and empirical performance in the noise-free case ($\boldsymbol{x} \in \{-1, 1\}^N$, $\Pr(x_n = -1) = 0.2$, $\Pr(x_n = 1) = 0.8$, $\Delta = 0.5$, and $\sigma_{\mathrm{v}}^2 = 0$)

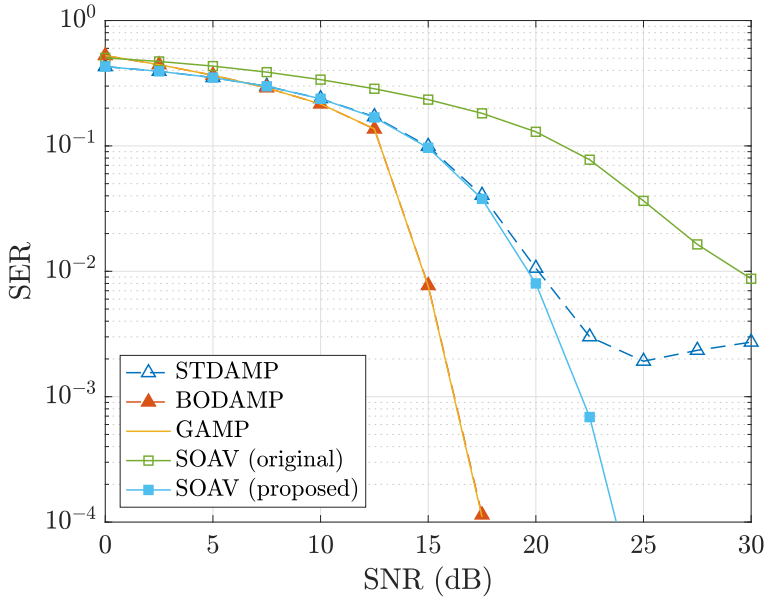


Figure 6.5: SER for i.i.d. Gaussian matrix ($\boldsymbol{x} \in \{-1, 0, 1\}^N$, $\Pr(x_n = 0) = 0.2$, $\Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$, and $(N, M) = (1000, 800)$)

"BODAMP" denote soft thresholding DAMP and Bayes optimal DAMP, respectively. For comparison, we also plot the performance of sum-product GAMP [43] with the

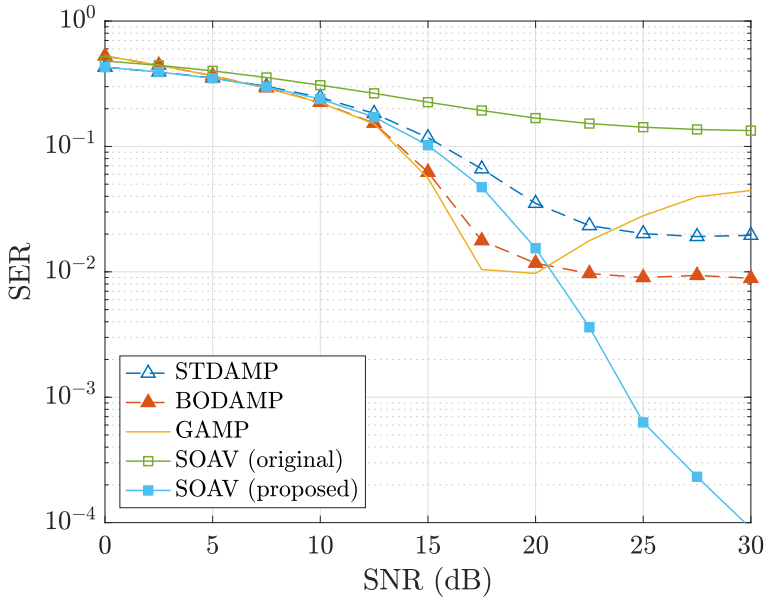


Figure 6.6: SER for i.i.d. Gaussian matrix ($\boldsymbol{x} \in \{-1, 0, 1\}^N$, $\Pr(x_n = 0) = 0.2$, $\Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$, and $(N, M) = (100, 80)$)

discrete-prior distribution as "GAMP". "SOAV (original)" and "SOAV (proposed)" represent the SOAV optimization with the original coefficients $q_\ell = p_\ell$ [60] and that with the proposed parameters q_ℓ^{opt}, respectively. The parameter for the original SOAV optimization λ is fixed as $\lambda = 10$. In the simulation, we have used Beck-Teboulle proximal gradient algorithm [61, 62] to solve these optimization problems. In both figures, we can see that the performance of the SOAV optimization with the proposed parameters is much better than the original ones. In Fig. 6.5, where $N = 1000$, Bayes optimal DAMP and GAMP have better SER performance than the other methods. As mentioned in Section 6.5, the GAMP algorithm uses the knowledge of the noise variance, which is not necessary in the proposed Bayes optimal DAMP. For the smaller-scale problem in Fig. 6.6, however, the performance of these methods severely degrades and the SOAV optimization with the proposed parameters can achieve the best SER performance for high SNR region. The difference of the error floor between Bayes optimal DAMP and GAMP may be caused by the estimation of the effective variance described in Section 6.5.

In Fig. 6.7, we show the SER performance for the correlated measurement matrix $\boldsymbol{A} = \boldsymbol{\Phi}_{\mathrm{R}}^{\frac{1}{2}} \boldsymbol{A}_{\mathrm{iid}} \boldsymbol{\Phi}_{\mathrm{T}}^{\frac{1}{2}}$. Here, $\boldsymbol{A}_{\mathrm{iid}} \in \mathbb{R}^{M \times N}$ is composed of i.i.d. Gaussian variables with zero mean and variance $1/M$. The (i, j) elements of the positive definite matrices $\boldsymbol{\Phi}_{\mathrm{R}}$ and $\boldsymbol{\Phi}_{\mathrm{T}}$ are given by $[\boldsymbol{\Phi}_{\mathrm{R}}]_{i,j} = J_0(|i - j| \cdot 2\pi d_{\mathrm{R}}/\nu)$ and $[\boldsymbol{\Phi}_{\mathrm{T}}]_{i,j} = J_0(|i - j| \cdot 2\pi d_{\mathrm{T}}/\nu)$, respectively. $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind and we set $d_{\mathrm{R}} = d_{\mathrm{T}} = \nu/2$ in the simulation. This model has been used for spatially correlated MIMO channels with equally spaced antennas [11]. The problem size and the distribution

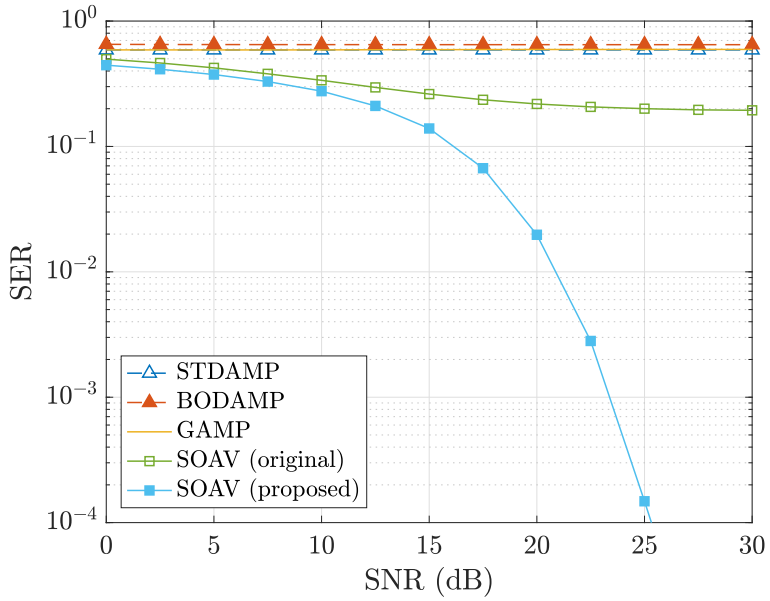


Figure 6.7: SER for correlated matrix ($\boldsymbol{x} \in \{-1, 0, 1\}^N$, $\Pr(x_n = 0) = 0.2$, $\Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$, and $(N, M) = (1000, 800)$)

of the unknown vector are the same as those in Fig. 6.5. From Fig. 6.7, we can see that the performance of AMP-based algorithms severely degrades because of the correlation. On the other hand, the approach based on the SOAV optimization with the proposed parameters works well even for the correlated measurement matrix.

Next, we evaluate the DAMP performance when the measurement matrix is a partial DCT matrix. The measurement vector $\boldsymbol{y}$ is assumed to be written as

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{D}\boldsymbol{x} + \boldsymbol{v}, \tag{6.25}$$

where $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ is the DCT matrix and its (i, j) element is given by

$$d_{i,j} = \begin{cases} \sqrt{\dfrac{1}{N}} & (i = 1) \\ \sqrt{\dfrac{2}{N}} \cos\left(\dfrac{\pi}{2N}(i-1)(2j-1)\right) & (i \neq 1) \end{cases} \tag{6.26}$$

in the simulations. The selection matrix $\boldsymbol{S} \in \mathbb{R}^{M \times N}$ is composed by randomly selecting M rows of the $N \times N$ identity matrix. Figures 6.8 and 6.9 show the SER performance for the partial DCT matrix. The distribution of the unknown vector is $\Pr(x_n = 0) = 0.2, \Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$. The problem size is $(N, M) = (1024, 768)$ in Fig. 6.8 and $(N, M) = (128, 96)$ in Fig. 6.9. In the figures, "Turbo-CS" denotes the performance of the algorithm based on turbo compressed sensing [103, 104], which has

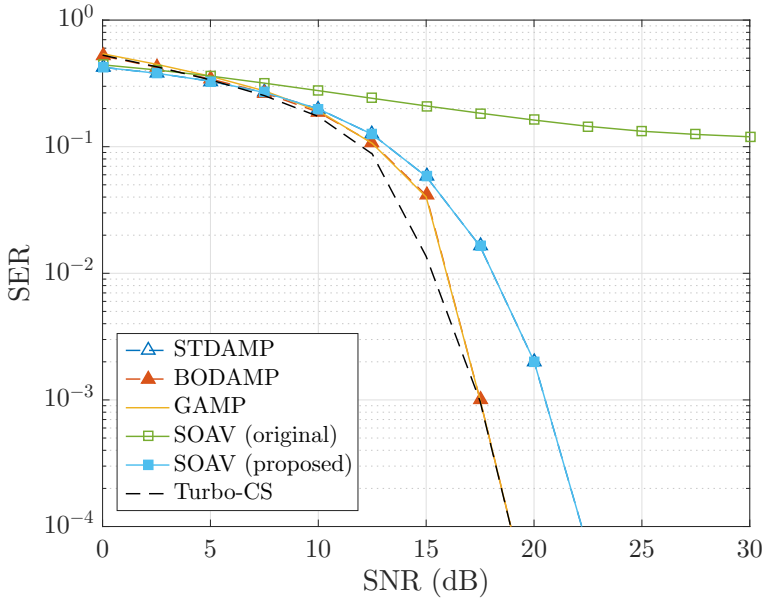


Figure 6.8: SER for partial DCT matrix ($\boldsymbol{x} \in \{-1, 0, 1\}^N$, $\Pr(x_n = 0) = 0.2$, $\Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$, and $(N, M) = (1024, 768)$)

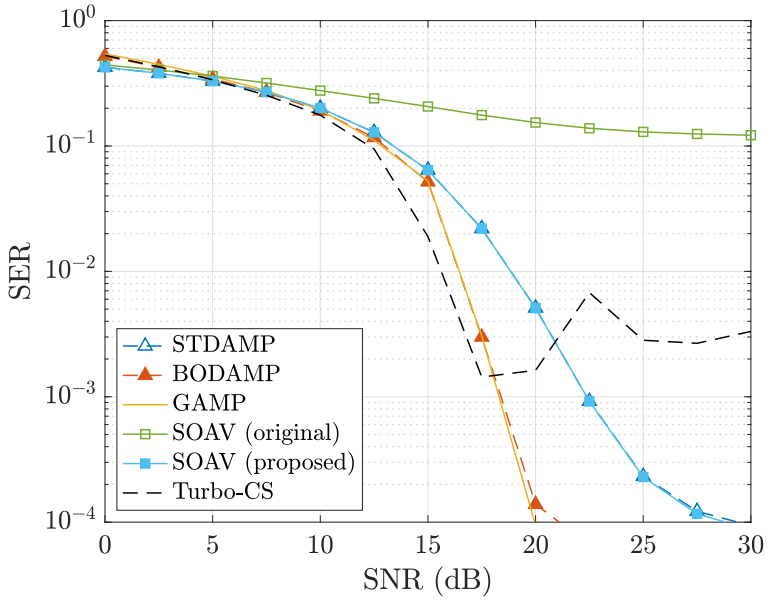


Figure 6.9: SER for partial DCT matrix ($\boldsymbol{x} \in \{-1, 0, 1\}^N$, $\Pr(x_n = 0) = 0.2$, $\Pr(x_n = -1) = \Pr(x_n = 1) = 0.4$, and $(N, M) = (128, 96)$)

been proposed for the measurement with a partial DFT matrix. Although Turbo-CS achieves the best SER in Fig. 6.8, it has the error floor in Fig. 6.9 possibly because

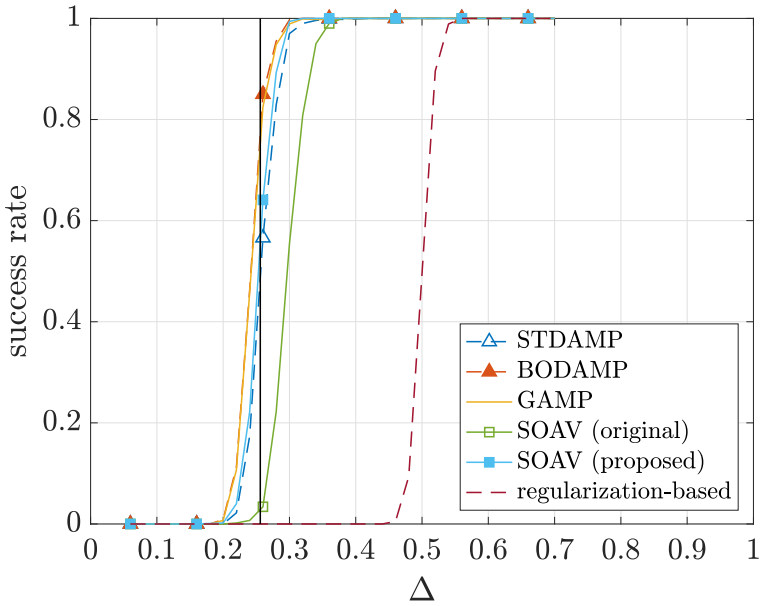


Figure 6.10: Success rate in the noise-free case ($\boldsymbol{x} \in \{0,1\}^N$, $\Pr(x_n = 0) = 0.1$, $\Pr(x_n = 1) = 0.9$, and $N = 1000$)

the problem size is rather small. In [13], a similar phenomenon can be observed for IO-LAMA [42], which is an AMP-based MIMO signal detection scheme. The SERs at the error floor will depend both on the algorithms and the structure of the measurement matrix. From the figures, we observe that Bayes optimal DAMP and GAMP achieve good performance even if the problem size is not very large when the measurement matrix is a partial DCT matrix. We note again that Bayes optimal DAMP does not require the knowledge of noise variance unlike GAMP .

In Figs. 6.10 and 6.11, we empirically evaluate the rate of the success recovery in the sense that $\boldsymbol{Q}(\boldsymbol{x}^t) = \boldsymbol{x}$ after $t = 300$ iterations. Figure 6.10 shows the success rate for the binary vector $\boldsymbol{x} \in \{0,1\}^N$ with $N = 1000$. The distribution of the unknown vector is given by $\Pr(x_n = 0) = 0.1$ and $\Pr(x_n = 1) = 0.9$. The measurement matrix is an i.i.d. Gaussian matrix. We consider the noise-free case and hence the SOAV optimization problem is given by

$$\begin{aligned} \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{s}\in\mathbb{R}^N} \; & q_1 \|\boldsymbol{s}\|_1 + q_2 \|\boldsymbol{s} - \boldsymbol{1}\|_1 \\ \text{subject to} \; & \boldsymbol{y} = \boldsymbol{A}\boldsymbol{s}, \end{aligned} \tag{6.27}$$

which is solved by Douglas-Rachford algorithm [61, 101] in the simulation. In the figure, "regularization-based" denotes the regularization-based method [57], which solves (6.27) with $q_1 = q_2 = 1$. The vertical line corresponds to the value of Δ for $D_{\min} = 1$ obtained from Fig. 6.1. In the large system limit, the left side of each vertical

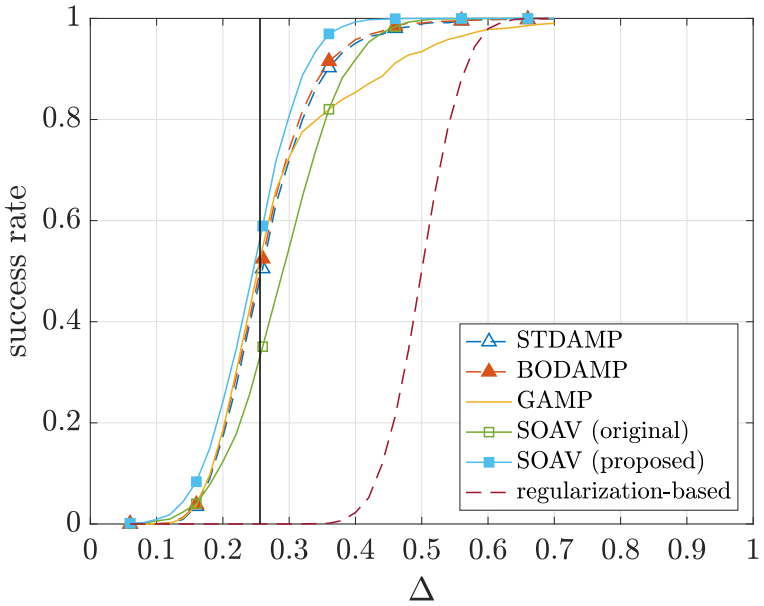


Figure 6.11: Success rate in the noise-free case ($\boldsymbol{x} \in \{0,1\}^N$, $\Pr(x_n = 0) = 0.1$, $\Pr(x_n = 1) = 0.9$, and $N = 100$)

line is the failure region and the right side is the success region of soft thresholding DAMP in the sense that $\sigma_t^2 \to 0$ ($t \to \infty$). The success rate of soft thresholding DAMP rapidly increases around the vertical line. Moreover, Bayes optimal DAMP, GAMP , and the SOAV optimization with the proposed parameters can achieve slightly better success rates than that of soft thresholding DAMP. Their success rates are also better than those of the SOAV optimization with original coefficients and the regularization-based method. One of possible reasons that the recovery rate is not equal to one in the success region near the boundary is that we restrict the maximum number of iterations as $t = 300$. Another reason will be that the problem size here is finite and not large enough. In Fig. 6.11, we evaluate the recovery rate for $\boldsymbol{x} \in \{0,1\}^N$ with $N = 100$. Since the problem size is smaller than that in Fig. 6.10, the performance of the AMP-based algorithms is worse than that of the SOAV optimization with the proposed parameters.

6.7 Conclusion

In this chapter, we have proposed the algorithm for the discrete-valued vector reconstruction, referred to as DAMP. We have analytically evaluated the asymptotic performance of soft thresholding DAMP and have derived the condition for the perfect reconstruction in the large system limit via state evolution. The optimization algorithm for the parameters of the soft thresholding function enables us to analyze the performance theoretically

in some cases. By using the analysis of soft thresholding DAMP, we have also proposed the method to determine the parameters of the SOAV optimization. Moreover, we have provided Bayes optimal DAMP, which gives much smaller MSE compared to soft thresholding DAMP. Via computer simulations, we have shown that DAMP can reconstruct the discrete-valued vector from its underdetermined linear measurements and the empirical performance agrees well with our theoretical results for large-scale problems. For smaller-scale problems, the SOAV optimization with the proposed parameters can achieve better performance than the AMP-based algorithms. We have also shown that Bayes optimal DAMP works well for partial DCT measurement matrices.

Appendix 6.A Derivation of $\Psi_{\mathrm{SE}}(\sigma^2)$

We firstly rewrite (6.8) as

$$\Psi_{\mathrm{SE}}(\sigma^2) = \sum_{\ell=1}^{L} p_\ell \Psi_{\mathrm{SE},\ell}(\sigma^2), \tag{6.28}$$

where

$$\Psi_{\mathrm{SE},\ell}(\sigma^2) = \int_{-\infty}^{\infty} \left\{ \eta\left(r_\ell + \frac{\sigma}{\sqrt{\Delta}} z; \frac{\sigma}{\sqrt{\Delta}} \right) - r_\ell \right\}^2 p_{\mathrm{G}}(z) dz. \tag{6.29}$$

From (6.4) and (6.5), we have

$$\eta\left(r_\ell + \frac{\sigma}{\sqrt{\Delta}} z; \frac{\sigma}{\sqrt{\Delta}} \right) = \begin{cases} r_\ell + \frac{\sigma}{\sqrt{\Delta}}(z - Q_1) & (z < T_{\ell,1,1}) \\ r_1 & (T_{\ell,1,1} \le z < T_{\ell,1,2}) \\ \quad \vdots & \quad \vdots \\ r_\ell + \frac{\sigma}{\sqrt{\Delta}}(z - Q_k) & (T_{\ell,k-1,k} \le z < T_{\ell,k,k}) \\ r_k & (T_{\ell,k,k} \le z < T_{\ell,k,k+1}) \\ \quad \vdots & \quad \vdots \\ r_\ell + \frac{\sigma}{\sqrt{\Delta}}(z - Q_{L+1}) & (T_{\ell,L,L+1} \le z) \end{cases}, \tag{6.30}$$

where

$$T_{\ell,k,k'} = \frac{\sqrt{\Delta}}{\sigma}(-r_\ell + r_k) + Q_{k'}. \tag{6.31}$$

We thus rewrite $\Psi_{\mathrm{SE},\ell}(\sigma^2)$ as

$$\begin{aligned}\Psi_{\mathrm{SE},\ell}(\sigma^2) = &\frac{\sigma^2}{\Delta}\int_{-\infty}^{T_{\ell,1,1}}(z-Q_1)^2\,p_{\mathrm{G}}(z)dz\\ &+\sum_{k=1}^{L}\int_{T_{\ell,k,k}}^{T_{\ell,k,k+1}}(r_k-r_\ell)^2\,p_{\mathrm{G}}(z)dz\\ &+\frac{\sigma^2}{\Delta}\sum_{k=2}^{L}\int_{T_{\ell,k-1,k}}^{T_{\ell,k,k}}(z-Q_k)^2\,p_{\mathrm{G}}(z)dz\\ &+\frac{\sigma^2}{\Delta}\int_{T_{\ell,L,L+1}}^{\infty}(z-Q_{L+1})^2\,p_{\mathrm{G}}(z)dz. \end{aligned}\tag{6.32}$$

For $a, b, Q \in \mathbb{R}$, we have

$$\begin{aligned}&\int_a^b (z-Q)^2\,p_{\mathrm{G}}(z)dz\\ &=\{-bp_{\mathrm{G}}(b)+ap_{\mathrm{G}}(a)+P_{\mathrm{G}}(b)-P_{\mathrm{G}}(a)\}\\ &\quad-2Q\{-p_{\mathrm{G}}(b)+p_{\mathrm{G}}(a)\}+Q^2\{P_{\mathrm{G}}(b)-P_{\mathrm{G}}(a)\},\end{aligned}\tag{6.33}$$

thus

$$\begin{aligned}\Psi_{\mathrm{SE},\ell}(\sigma^2) = &\sum_{k=1}^{L}(r_k-r_\ell)^2\left\{P_{\mathrm{G}}\left(T_{\ell,k,k+1}\right)-P_{\mathrm{G}}\left(T_{\ell,k,k}\right)\right\}\\ &+\frac{\sigma^2}{\Delta}\sum_{k=1}^{L}\Big[\left\{-T_{\ell,k,k}p_{\mathrm{G}}\left(T_{\ell,k,k}\right)+P_{\mathrm{G}}\left(T_{\ell,k,k}\right)\right\}\\ &\qquad+2Q_kp_{\mathrm{G}}\left(T_{\ell,k,k}\right)+Q_k^2P_{\mathrm{G}}\left(T_{\ell,k,k}\right)\Big]\\ &+\frac{\sigma^2}{\Delta}\sum_{k=2}^{L+1}\Big[\left\{T_{\ell,k-1,k}p_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)-P_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)\right\}\\ &\qquad-2Q_kp_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)-Q_k^2P_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)\Big]\\ &+\frac{\sigma^2}{\Delta}\left(1+Q_{L+1}^2\right).\end{aligned}\tag{6.34}$$

Hence, $\Psi_{\mathrm{SE}}(\sigma^2)$ in (6.28) can be obtained from (6.34).

Appendix 6.B Derivation of $\left.\frac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0}$

From (6.28), we have

$$\left.\frac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0} = \sum_{\ell=1}^{L} p_\ell \left.\frac{d\Psi_{\mathrm{SE},\ell}}{d(\sigma^2)}\right|_{\sigma\downarrow 0}. \tag{6.35}$$

With the derivative of $T_{\ell,k,k'}$ with respect to σ^2

$$T'_{\ell,k,k'} = \frac{dT_{\ell,k,k'}}{d(\sigma^2)} \left(= -\frac{\sqrt{\Delta}}{2\sigma^3}\left(-r_\ell + r_k\right)\right), \tag{6.36}$$

the derivative of (6.34) is given by

$$\begin{aligned}
&\frac{d\Psi_{\mathrm{SE},\ell}}{d(\sigma^2)} \\
&= \sum_{k=1}^{L} (r_k - r_\ell)^2 \left\{T'_{\ell,k,k+1} p_{\mathrm{G}}\left(T_{\ell,k,k+1}\right) - T'_{\ell,k,k} p_{\mathrm{G}}\left(T_{\ell,k,k}\right)\right\} \\
&\quad + \frac{1}{\Delta}\sum_{k=1}^{L}\left[\left\{-T_{\ell,k,k} p_{\mathrm{G}}\left(T_{\ell,k,k}\right) + P_{\mathrm{G}}\left(T_{\ell,k,k}\right)\right\} + 2Q_k p_{\mathrm{G}}\left(T_{\ell,k,k}\right) + Q_k^2 P_{\mathrm{G}}\left(T_{\ell,k,k}\right)\right] \\
&\quad + \frac{\sigma^2}{\Delta}\sum_{k=1}^{L}\left(T_{\ell,k,k} - Q_k\right)^2 T'_{\ell,k,k} p_{\mathrm{G}}\left(T_{\ell,k,k}\right) \\
&\quad + \frac{1}{\Delta}\sum_{k=2}^{L+1}\left[\left\{T_{\ell,k-1,k} p_{\mathrm{G}}\left(T_{\ell,k-1,k}\right) - P_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)\right\}\right. \\
&\qquad\qquad \left. -2Q_k p_{\mathrm{G}}\left(T_{\ell,k-1,k}\right) - Q_k^2 P_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)\right] \\
&\quad - \frac{\sigma^2}{\Delta}\sum_{k=2}^{L+1}\left(T_{\ell,k-1,k} - Q_k\right)^2 T'_{\ell,k-1,k} p_{\mathrm{G}}\left(T_{\ell,k-1,k}\right) \\
&\quad + \frac{1}{\Delta}\left(1 + Q_{L+1}^2\right) \\
&= \frac{1}{\Delta}\sum_{k=1}^{L}\left[\left\{-T_{\ell,k,k} p_{\mathrm{G}}\left(T_{\ell,k,k}\right) + P_{\mathrm{G}}\left(T_{\ell,k,k}\right)\right\} + 2Q_k p_{\mathrm{G}}\left(T_{\ell,k,k}\right) + Q_k^2 P_{\mathrm{G}}\left(T_{\ell,k,k}\right)\right] \\
&\quad + \frac{1}{\Delta}\sum_{k=2}^{L+1}\left[\left\{T_{\ell,k-1,k} p_{\mathrm{G}}\left(T_{\ell,k-1,k}\right) - P_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)\right\}\right.
\end{aligned} \tag{6.37}$$

$$-2Q_k p_{\mathrm{G}}\left(T_{\ell,k-1,k}\right) - Q_k^2 P_{\mathrm{G}}\left(T_{\ell,k-1,k}\right)\Big] + \frac{1}{\Delta}\left(1+Q_{L+1}^2\right). \tag{6.38}$$

From

$$\lim_{\sigma\downarrow 0} T_{\ell,k,k'} = \begin{cases} \infty & (\ell < k) \\ Q_{k'} & (\ell = k) \\ -\infty & (\ell > k) \end{cases}, \tag{6.39}$$

we conclude that

$$\begin{aligned}
&\left.\frac{d\Psi_{\mathrm{SE},\ell}}{d(\sigma^2)}\right|_{\sigma\downarrow 0} \\
&= \frac{1}{\Delta}\left[\left\{-Q_\ell p_{\mathrm{G}}\left(Q_\ell\right) + P_{\mathrm{G}}\left(Q_\ell\right)\right\} + 2Q_\ell p_{\mathrm{G}}\left(Q_\ell\right) + Q_\ell^2 P_{\mathrm{G}}\left(Q_\ell\right)\right] \\
&\quad + \frac{1}{\Delta}\sum_{k=\ell+1}^{L}\left(1+Q_k^2\right) \\
&\quad + \frac{1}{\Delta}\left[\left\{Q_{\ell+1} p_{\mathrm{G}}\left(Q_{\ell+1}\right) - P_{\mathrm{G}}\left(Q_{\ell+1}\right)\right\} - 2Q_{\ell+1} p_{\mathrm{G}}\left(Q_{\ell+1}\right) - Q_{\ell+1}^2 P_{\mathrm{G}}\left(Q_{\ell+1}\right)\right] \\
&\quad + \frac{1}{\Delta}\sum_{k=\ell+2}^{L+1}\left(-1-Q_k^2\right) \\
&\quad + \frac{1}{\Delta}\left(1+Q_{L+1}^2\right) \qquad (6.40) \\
&= \frac{1}{\Delta}\Big\{Q_\ell p_{\mathrm{G}}\left(Q_\ell\right) - Q_{\ell+1} p_{\mathrm{G}}\left(Q_{\ell+1}\right) + \left(1+Q_\ell^2\right) P_{\mathrm{G}}\left(Q_\ell\right) \\
&\qquad + \left(1+Q_{\ell+1}^2\right)\left(1 - P_{\mathrm{G}}\left(Q_{\ell+1}\right)\right)\Big\}. \qquad (6.41)
\end{aligned}$$

Hence, $\left.\frac{d\Psi_{\mathrm{SE}}}{d(\sigma^2)}\right|_{\sigma\downarrow 0}$ is straightforwardly obtained from (6.35) and (6.41) as in (6.10).

Appendix 6.C Proof of Theorem 6.3.1

Since $D(\boldsymbol{Q})$ is a monotonically increasing function of Q_1 and a monotonically decreasing function of Q_{L+1}, their optimal values are $Q_1^{\mathrm{opt}} = -\infty$ and $Q_{L+1}^{\mathrm{opt}} = \infty$, respectively. Thus, the optimization problem (6.11) can be reduced to

$$\begin{aligned}
D_{\min} = &\min_{Q_2,\ldots,Q_L} \tilde{D}(Q_2,\ldots,Q_L) \\
&\text{subject to } Q_2 \le \cdots \le Q_L,
\end{aligned} \tag{6.42}$$

where

$$\begin{aligned}
&\tilde{D}(Q_2,\ldots,Q_L) \\
&:= \frac{\Delta}{2} D(\boldsymbol{Q})\Big|_{Q_1=-\infty,Q_{L+1}=\infty} \qquad (6.43)\\
&= \frac{1}{2}\sum_{\ell=2}^{L}\Big[p_{\ell-1}\left\{-Q_\ell p_{\mathrm{G}}(Q_\ell)+\left(1+Q_\ell^2\right)(1-P_{\mathrm{G}}(Q_\ell))\right\} \\
&\qquad +p_\ell\left\{Q_\ell p_{\mathrm{G}}(Q_\ell)+\left(1+Q_\ell^2\right)P_{\mathrm{G}}(Q_\ell)\right\}\Big]. \qquad (6.44)
\end{aligned}$$

It is sufficient to confirm that $\tilde{D}(Q_2,\ldots,Q_L)$ is strictly convex and $Q_2^{\mathrm{opt}},\ldots,Q_L^{\mathrm{opt}}$ obtained by Algorithm 6.2 satisfies Karush-Kuhn-Tucker (KKT) conditions of (6.42).

6.C.1 Strict Convexity of $\tilde{D}(Q_2,\ldots,Q_L)$

To prove the strict convexity of $\tilde{D}(Q_2,\ldots,Q_L)$, we show that the Hessian $\nabla^2\tilde{D}$ is positive definite. The partial derivative of $\tilde{D}(Q_2,\ldots,Q_L)$ with respect to Q_ℓ $(\ell=2,\ldots,L)$ is given by

$$\frac{\partial\tilde{D}}{\partial Q_\ell} = p_{\ell-1}\left\{-p_{\mathrm{G}}(Q_\ell)+Q_\ell(1-P_{\mathrm{G}}(Q_\ell))\right\}+p_\ell\left\{p_{\mathrm{G}}(Q_\ell)+Q_\ell P_{\mathrm{G}}(Q_\ell)\right\}. \quad (6.45)$$

The second-order partial derivative can be written as

$$\frac{\partial^2\tilde{D}}{\partial Q_\ell^2} = p_{\ell-1}(1-P_{\mathrm{G}}(Q_\ell))+p_\ell P_{\mathrm{G}}(Q_\ell) > 0, \quad (6.46)$$

and

$$\frac{\partial^2\tilde{D}}{\partial Q_\ell\partial Q_{\ell'}} = 0\ (\ell\neq\ell'). \quad (6.47)$$

From (6.46) and (6.47), the Hessian $\nabla^2\tilde{D} = \mathrm{diag}\left(\frac{\partial^2\tilde{D}}{\partial Q_2^2},\ldots,\frac{\partial^2\tilde{D}}{\partial Q_L^2}\right)$ is positive definite and hence $\tilde{D}(Q_2,\ldots,Q_L)$ is a strictly convex function of $Q_2,\ldots,Q_L$.

6.C.2 KKT Conditions

Next, we prove that $Q_2^{\mathrm{opt}},\ldots,Q_L^{\mathrm{opt}}$ satisfies the KKT conditions of (6.42). We define the Lagrange function as

$$\mathcal{L}(Q_2,\ldots,Q_L) = \tilde{D}(Q_2,\ldots,Q_L)+\sum_{\ell=2}^{L-1}\mu_\ell(Q_\ell-Q_{\ell+1}), \quad (6.48)$$

where $\mu_2, \ldots, \mu_{L-1}$ are the KKT multipliers. Since the partial derivatives of $\tilde{D}(Q_2, \ldots, Q_L)$ are obtained as in (6.45), the KKT conditions can be written with $F_\ell(Q)$ $(\ell = 2, \ldots, L)$ defined in Algorithm 6.2.

KKT conditions of (6.42)

1. $F_2(Q_2) + \mu_2 = 0$,
   $F_\ell(Q_\ell) - \mu_{\ell-1} + \mu_\ell = 0$ $(\ell = 3, \ldots, L-1)$,
   $F_L(Q_L) - \mu_{L-1} = 0$.
2. $Q_\ell - Q_{\ell+1} \leq 0$ $(\ell = 2, \ldots, L-1)$.
3. $\mu_\ell \geq 0$ $(\ell = 2, \ldots, L-1)$.
4. $\mu_\ell(Q_\ell - Q_{\ell+1}) = 0$ $(\ell = 2, \ldots, L-1)$.

Before the investigation of the KKT conditions, we confirm that the equations $G_\ell(Q) = 0$ and $\sum_{k=j}^{\ell-1} F_k(Q) = 0$ have a unique solution and $\hat{Q}(\cdot)$ in Algorithm 6.2 can be defined properly. For $\ell = 2, \ldots, L$, we have $\lim_{Q \to -\infty} F_\ell(Q) = -\infty$, $\lim_{Q \to \infty} F_\ell(Q) = \infty$, and

$$\frac{dF_\ell}{dQ} = p_{\ell-1}\left(1 - P_{\mathrm{G}}(Q)\right) + p_\ell P_{\mathrm{G}}(Q) > 0, \tag{6.49}$$

which show that each $F_\ell(Q)$ is a strictly increasing function with the range $\mathbb{R}$. Since $G_\ell(Q)$ and $\sum_{k=j}^{\ell-1} F_k(Q)$ are sum of some $F_\ell(Q)$ $(\ell = 2, \ldots, L)$, they are also strictly increasing functions with the range $\mathbb{R}$ and the solutions of $G_\ell(Q) = 0$ and $\sum_{k=j}^{\ell-1} F_k(Q) = 0$ are unique.

We then prove that $Q_2^{\mathrm{opt}}, \ldots, Q_L^{\mathrm{opt}}$ obtained by Algorithm 6.2 and $\mu_\ell^{\mathrm{opt}} := G_{\ell+1}\left(Q_{\ell+1}^{\mathrm{opt}}\right)$ $(\ell = 2, \ldots, L-1)$ satisfy the KKT conditions, i.e.,

$$F_2\left(Q_2^{\mathrm{opt}}\right) + \mu_2^{\mathrm{opt}} = 0, \tag{6.50}$$

$$F_\ell\left(Q_\ell^{\mathrm{opt}}\right) - \mu_{\ell-1}^{\mathrm{opt}} + \mu_\ell^{\mathrm{opt}} = 0 \ (\ell = 3, \ldots, L-1), \tag{6.51}$$

$$F_L\left(Q_L^{\mathrm{opt}}\right) - \mu_{L-1}^{\mathrm{opt}} = 0, \tag{6.52}$$

$$Q_\ell^{\mathrm{opt}} - Q_{\ell+1}^{\mathrm{opt}} \leq 0 \ (\ell = 2, \ldots, L-1), \tag{6.53}$$

$$\mu_\ell^{\mathrm{opt}} \geq 0 \ (\ell = 2, \ldots, L-1), \tag{6.54}$$

$$\mu_\ell^{\mathrm{opt}}\left(Q_\ell^{\mathrm{opt}} - Q_{\ell+1}^{\mathrm{opt}}\right) = 0 \ (\ell = 2, \ldots, L-1). \tag{6.55}$$

In the following proofs of (6.50)–(6.55), we denote the condition $\hat{Q}(G_\ell(Q)) > \max_{j=2,\ldots,\ell-1} \hat{Q}\left(\sum_{k=j}^{\ell-1} F_k(Q)\right)$ in the line 8 of Algorithm 6.2 by H_ℓ.

proof of (6.50)–(6.52)

From the definition of $G_\ell(Q)$, we have $G_\ell\left(Q_\ell^{\mathrm{opt}}\right) = F_\ell\left(Q_\ell^{\mathrm{opt}}\right) + G_{\ell+1}\left(Q_{\ell+1}^{\mathrm{opt}}\right)$ for $\ell = 3, \ldots, L-1$. We thus obtain

$$F_\ell\left(Q_\ell^{\mathrm{opt}}\right) - \mu_{\ell-1}^{\mathrm{opt}} + \mu_\ell^{\mathrm{opt}} = F_\ell\left(Q_\ell^{\mathrm{opt}}\right) - G_\ell\left(Q_\ell^{\mathrm{opt}}\right) + G_{\ell+1}\left(Q_{\ell+1}^{\mathrm{opt}}\right) \tag{6.56}$$

$$= 0. \tag{6.57}$$

Similarly, we have $F_2\left(Q_2^{\mathrm{opt}}\right)+\mu_2^{\mathrm{opt}} = F_2\left(Q_2^{\mathrm{opt}}\right)+G_3\left(Q_3^{\mathrm{opt}}\right) = G_2\left(Q_2^{\mathrm{opt}}\right) = G_2\left(\hat{Q}(G_2(Q))\right) = 0$ and $F_L\left(Q_L^{\mathrm{opt}}\right) - \mu_{L-1}^{\mathrm{opt}} = F_L\left(Q_L^{\mathrm{opt}}\right) - G_L\left(Q_L^{\mathrm{opt}}\right) = 0$ because $G_L(Q) = F_L(Q)$.

proof of (6.53)

We firstly consider the case where the condition $H_{\ell+1}$ is satisfied. In this case, $Q_{\ell+1}^{\mathrm{opt}}$ is determined as $Q_{\ell+1}^{\mathrm{opt}} = \hat{Q}\left(G_{\ell+1}(Q)\right)$. We define ℓ' $(< \ell)$ as the maximum index that $H_{\ell'}$ is true, i.e., the conditions $H_\ell, H_{\ell-1}, \ldots, H_{\ell'+1}$ are not satisfied and the condition $H_{\ell'}$ is satisfied. By using Algorithm 6.2, we can obtain $G_{\ell'}(Q) = \sum_{k=\ell'}^{\ell} F_k(Q)$ and $Q_\ell^{\mathrm{opt}} = Q_{\ell-1}^{\mathrm{opt}} = \cdots = Q_{\ell'}^{\mathrm{opt}} = \hat{Q}\left(G_{\ell'}(Q)\right)$. We thus have $Q_{\ell+1}^{\mathrm{opt}} = \hat{Q}\left(G_{\ell+1}(Q)\right) > \hat{Q}\left(\sum_{k=\ell'}^{\ell} F_k(Q)\right) = \hat{Q}\left(G_{\ell'}(Q)\right) = Q_\ell^{\mathrm{opt}}$ and hence $Q_\ell^{\mathrm{opt}} - Q_{\ell+1}^{\mathrm{opt}} \leq 0$.

If the condition $H_{\ell+1}$ is not satisfied and $Q_{\ell+1}^{\mathrm{opt}} = Q_\ell^{\mathrm{opt}}$, we also have $Q_\ell^{\mathrm{opt}} - Q_{\ell+1}^{\mathrm{opt}} \leq 0$.

proof of (6.54)

If the condition $H_{\ell+1}$ is satisfied, $\mu_\ell^{\mathrm{opt}} = G_{\ell+1}\left(Q_{\ell+1}^{\mathrm{opt}}\right) = G_{\ell+1}\left(\hat{Q}(G_{\ell+1}(Q))\right) = 0$ and hence $\mu_\ell^{\mathrm{opt}} \geq 0$ holds.

Next, we assume that the condition $H_{\ell+1}$ is not satisfied. In this case, we have

$$\hat{Q}\left(G_{\ell+1}(Q)\right) \leq \max_{j=2,\ldots,\ell} \hat{Q}\left(\sum_{k=j}^{\ell} F_k(Q)\right). \tag{6.58}$$

We define ℓ' $(< \ell + 1)$ as the maximum index that $H_{\ell'}$ is true, i.e., the conditions $H_\ell, H_{\ell-1}, \ldots, H_{\ell'+1}$ are not satisfied and the condition $H_{\ell'}$ is satisfied. We can obtain

$$\hat{Q}\left(G_{\ell'}(Q)\right) > \max_{j=2,\ldots,\ell'-1} \hat{Q}\left(\sum_{k=j}^{\ell'-1} F_k(Q)\right), \tag{6.59}$$

$$G_{\ell'}(Q) = G_{\ell+1}(Q) + \sum_{k=\ell'}^{\ell} F_k(Q), \tag{6.60}$$

and $Q_{\ell+1}^{\mathrm{opt}} = Q_{\ell}^{\mathrm{opt}} = \cdots = Q_{\ell'}^{\mathrm{opt}} = \hat{Q}\left(G_{\ell'}(Q)\right)$. In what follows, we often use the following lemma.

Lemma 6.C.1. For strictly increasing functions $f(Q)$ and $g(Q)$ with the range $\mathbb{R}$, we have

$$\hat{Q}\left(f(Q)\right) < \hat{Q}\left(g(Q)\right) \iff \hat{Q}\left(f(Q)\right) < \hat{Q}\left(f(Q) + g(Q)\right) < \hat{Q}\left(g(Q)\right). \tag{6.61}$$

The proposition obtained by replacing all of $<$ with $\leq$ also holds.

To prove $\mu_{\ell}^{\mathrm{opt}} \geq 0$, we will show that

$$\hat{Q}\left(G_{\ell+1}(Q)\right) \leq \hat{Q}\left(\sum_{k=\ell'}^{\ell} F_k(Q)\right), \tag{6.62}$$

which results in $\hat{Q}\left(G_{\ell+1}(Q)\right) \leq \hat{Q}\left(G_{\ell'}(Q)\right)$ from (6.60) and Lemma 6.C.1. If $\hat{Q}\left(G_{\ell+1}(Q)\right) \leq \hat{Q}\left(G_{\ell'}(Q)\right)$ holds, we can obtain $\mu_{\ell}^{\mathrm{opt}} \geq 0$ as $\mu_{\ell}^{\mathrm{opt}} = G_{\ell+1}\left(Q_{\ell+1}^{\mathrm{opt}}\right) = G_{\ell+1}\left(Q_{\ell'}^{\mathrm{opt}}\right) = G_{\ell+1}\left(\hat{Q}\left(G_{\ell'}(Q)\right)\right) \geq G_{\ell+1}\left(\hat{Q}\left(G_{\ell+1}(Q)\right)\right) = 0$.

To show (6.62), we provide the proof by contradiction with the assumption

$$\hat{Q}\left(G_{\ell+1}(Q)\right) > \hat{Q}\left(\sum_{k=\ell'}^{\ell} F_k(Q)\right). \tag{6.63}$$

From (6.60), (6.63) and Lemma 6.C.1, we have

$$\hat{Q}\left(\sum_{k=\ell'}^{\ell} F_k(Q)\right) < \hat{Q}\left(G_{\ell'}(Q)\right) < \hat{Q}\left(G_{\ell+1}(Q)\right). \tag{6.64}$$

It follows from (6.59) and (6.64) that

$$\max_{j=2,\ldots,\ell'} \hat{Q}\left(\sum_{k=j}^{\ell} F_k(Q)\right) < \hat{Q}\left(G_{\ell'}(Q)\right) < \hat{Q}\left(G_{\ell+1}(Q)\right). \tag{6.65}$$

From (6.58) and (6.65), we can obtain

$$\hat{Q}\left(G_{\ell+1}(Q)\right) \leq \max_{j=\ell'+1,\ldots,\ell} \hat{Q}\left(\sum_{k=j}^{\ell} F_k(Q)\right). \tag{6.66}$$

We define ℓ_1 as $\ell_1 = \underset{j=\ell'+1,\ldots,\ell}{\arg\max}\ \hat{Q}\left(\sum_{k=j}^{\ell} F_k(Q)\right)$, which results in

$$\hat{Q}\left(G_{\ell+1}(Q)\right) \leq \hat{Q}\left(\sum_{k=\ell_1}^{\ell} F_k(Q)\right). \tag{6.67}$$

Since $\ell' + 1 \leq \ell_1 \leq \ell$, the conditions $H_\ell, \ldots, H_{\ell_1}$ are not satisfied and hence we have

$$\hat{Q}(G_{\ell_1}(Q)) \leq \max_{j=2,\ldots,\ell_1-1} \hat{Q}\left(\sum_{k=j}^{\ell_1-1} F_k(Q)\right), \tag{6.68}$$

$$G_{\ell_1}(Q) = G_{\ell+1}(Q) + \sum_{k=\ell_1}^{\ell} F_k(Q). \tag{6.69}$$

Lemma 6.C.1, (6.67), and (6.69) give

$$\hat{Q}\left(G_{\ell+1}(Q)\right) \leq \hat{Q}\left(G_{\ell_1}(Q)\right) \leq \hat{Q}\left(\sum_{k=\ell_1}^{\ell} F_k(Q)\right). \tag{6.70}$$

From (6.65) and (6.70), we have

$$\hat{Q}\left(G_{\ell'}(Q)\right) < \hat{Q}\left(G_{\ell+1}(Q)\right) \leq \hat{Q}\left(G_{\ell_1}(Q)\right). \tag{6.71}$$

With a similar approach to (6.69), we can also obtain

$$G_{\ell'}(Q) = G_{\ell_1}(Q) + \sum_{k=\ell'}^{\ell_1-1} F_k(Q) \tag{6.72}$$

and

$$\hat{Q}\left(\sum_{k=\ell'}^{\ell_1-1} F_k(Q)\right) \leq \hat{Q}\left(G_{\ell'}(Q)\right) < \hat{Q}\left(G_{\ell_1}(Q)\right). \tag{6.73}$$

It follows from (6.59) and (6.73) that

$$\max_{j=2,\ldots,\ell'} \hat{Q}\left(\sum_{k=j}^{\ell_1-1} F_k(Q)\right) < \hat{Q}\left(G_{\ell_1}(Q)\right). \tag{6.74}$$

If $\ell_1 = \ell' + 1$, (6.74) contradicts (6.68) and hence we can conclude (6.62) and $\mu_\ell^{\mathrm{opt}} \geq 0$. Otherwise $\ell_1 > \ell' + 1$, and in this case combining (6.68) and (6.74) gives

$$\hat{Q}(G_{\ell_1}(Q)) \leq \max_{j=\ell'+1,\ldots,\ell_1-1} \hat{Q}\left(\sum_{k=j}^{\ell_1-1} F_k(Q)\right). \tag{6.75}$$

We then define ℓ_2 as $\ell_2 = \operatorname*{arg\,max}_{j=\ell'+1,\ldots,\ell_1-1} \hat{Q}\left(\sum_{k=j}^{\ell_1-1} F_k(Q)\right)$. Here, note that $\ell' + 1 \leq \ell_2 < \ell_1 < \ell + 1$. By repeating the same manner, we have a sequence $\ell_1, \ell_2, \ldots, \ell_i$ satisfying

$\ell' + i - 1 \leq \ell_i < \ell_{i-1} < \cdots < \ell_1 < \ell + 1$. Since $\{\ell' + i - 1\}_{i=1,\ldots}$ is monotonically increasing and $\{\ell_i\}_{i=1,\ldots}$ is monotonically decreasing, there exists $\tilde{i}$ satisfying $\ell' + \tilde{i} - 1 = \ell_{\tilde{i}} < \ell_{\tilde{i}-1} < \cdots < \ell_1 < \ell + 1$. Moreover, similar to (6.74), we have

$$\hat{Q}\left(G_{\ell_{\tilde{i}}}(Q)\right) > \max_{j=2,\ldots,\ell'+\tilde{i}-2} \hat{Q}\left(\sum_{k=j}^{\ell_{\tilde{i}}-1} F_k(Q)\right) \tag{6.76}$$

$$= \max_{j=2,\ldots,\ell_{\tilde{i}}-1} \hat{Q}\left(\sum_{k=j}^{\ell_{\tilde{i}}-1} F_k(Q)\right). \tag{6.77}$$

However, (6.77) contradicts the fact that ℓ' $(< \ell + 1)$ is the maximum index of $H_{\ell'}$ being true because $\ell_{\tilde{i}} > \ell'$. We thus conclude (6.62) and $\mu_\ell^{\rm opt} \geq 0$.

proof of (6.55)

If the condition $H_{\ell+1}$ is satisfied and $Q_{\ell+1}^{\rm opt}$ is determined as $Q_{\ell+1}^{\rm opt} = \hat{Q}\left(G_{\ell+1}(Q)\right)$, $\mu_\ell^{\rm opt} = G_{\ell+1}\left(Q_{\ell+1}^{\rm opt}\right) = 0$ and hence $\mu_\ell^{\rm opt}\left(Q_\ell^{\rm opt} - Q_{\ell+1}^{\rm opt}\right) = 0$ holds. Otherwise, $Q_{\ell+1}^{\rm opt} = Q_\ell^{\rm opt}$ and hence $\mu_\ell^{\rm opt}\left(Q_\ell^{\rm opt} - Q_{\ell+1}^{\rm opt}\right) = 0$ also holds.

Appendix 6.D Derivation of $\frac{d^2\Psi_{\rm SE}}{d(\sigma^2)^2}$

From (6.38), the second derivative of $\Psi_{{\rm SE},\ell}(\sigma^2)$ is given by

$$\begin{aligned}
&\frac{d^2\Psi_{{\rm SE},\ell}}{d(\sigma^2)^2} \\
&= \frac{1}{\Delta}\sum_{k=1}^{L}\left\{T_{\ell,k,k}^2 T'_{\ell,k,k} p_{\rm G}\left(T_{\ell,k,k}\right) - 2Q_k T_{\ell,k,k} T'_{\ell,k,k} p_{\rm G}\left(T_{\ell,k,k}\right) + Q_k^2 T'_{\ell,k,k} p_{\rm G}\left(T_{\ell,k,k}\right)\right\} \\
&\quad - \frac{1}{\Delta}\sum_{k=2}^{L+1}\left\{T_{\ell,k-1,k}^2 T'_{\ell,k-1,k} p_{\rm G}\left(T_{\ell,k-1,k}\right) - 2Q_k T_{\ell,k-1,k} T'_{\ell,k-1,k} p_{\rm G}\left(T_{\ell,k-1,k}\right)\right. \\
&\qquad\qquad \left. + Q_k^2 T'_{\ell,k-1,k} p_{\rm G}\left(T_{\ell,k-1,k}\right)\right\} \qquad (6.78) \\
&= \frac{1}{\Delta}\sum_{k=1}^{L}\left(T_{\ell,k,k} - Q_k\right)^2 T'_{\ell,k,k} p_{\rm G}\left(T_{\ell,k,k}\right) \\
&\quad - \frac{1}{\Delta}\sum_{k=2}^{L+1}\left(T_{\ell,k-1,k} - Q_k\right)^2 T'_{\ell,k-1,k} p_{\rm G}\left(T_{\ell,k-1,k}\right) \qquad (6.79)
\end{aligned}$$

$$= \frac{\sqrt{\Delta}}{2\sigma^5} \sum_{k=1}^{L} (-r_\ell + r_k)^3 \left\{ -p_{\mathrm{G}} \left(T_{\ell,k,k} \right) + p_{\mathrm{G}} \left(T_{\ell,k,k+1} \right) \right\}, \tag{6.80}$$

which results in (6.13).

Chapter 7

Conclusion

7.1 Summary

In this thesis, we have proposed several algorithms for the discrete-valued vector reconstruction and analyze the reconstruction performance of some algorithms.

In Chapter 1, we have explained the discrete-valued vector reconstruction from underdetermined linear measurements and provided several applications in communication systems. The problem of the discrete-valued vector reconstruction appears in many applications, such as MIMO signal detection, channel equalization, decoding of NO-STBC, multiuser detection, and FTN signaling. We have also described conventional approaches for the discrete-valued vector reconstruction. Various low-complexity algorithms using the discrete nature have been proposed on the basis of message passing or convex optimization. The message passing-based approach can achieve good performance under some assumptions on the measurement matrix, whereas the convex optimization-based approach does not require such assumptions in the algorithm.

In Chapter 2, we have proposed the reconstruction algorithm for the binary vector via iterative convex optimization. The proposed IW-SOAV iterate the W-SOAV optimization and the update of the parameters in the objective function. For the W-SOAV optimization, we have derived the algorithm based on Douglas-Rachford algorithm, which is one of proximal splitting methods. In each iteration of IW-SOAV, we can improve the estimate of the unknown vector by using the tentative estimate in the previous iteration as the prior information. Simulation results show that the proposed method outperforms several conventional methods in massive overloaded MIMO signal detection and the decoding of NO-STBCs.

In Chapter 3, we have extended the conventional SOAV optimization to the SCSR optimization for the reconstruction of complex discrete-valued vector. The proposed SCSR optimization uses the weighted sum of some sparse regularizers in the complex-valued domain as a regularizer for the discrete-valued vector. Moreover, we extend

the SCSR optimization to the W-SCSR optimization and propose the iterative approach named IW-SCSR, where we iteratively solve the W-SCSR optimization with the update of the parameters in the objective function. By solving the optimization in the complex-valued domain, we can directly utilize the discrete nature of the unknown vector in the complex-valued domain. Simulation results show that the proposed IW-SCSR can achieve better performance than conventional methods in the applications of overloaded MIMO signal detection and channel equalization.

In Chapter 4, we have proposed the algorithms on the basis of the SSR optimization. The proposed SSR optimization uses possibly nonconvex sparse regularizer, whereas only convex objective functions have been considered in the conventional methods. For the SSR optimization, we have derived two algorithms on the basis of ADMM and PDS. The proposed approach can also be used for the reconstruction of the complex discrete-valued vector. Simulation results show that the proposed method with nonconvex regularizers outperforms the conventional convex optimization-based methods.

In Chapter 5, we have analyzed the asymptotic performance of the Box-SOAV optimization. By using the analysis using CGMT, we have characterized the SER of the Box-SOAV in the large system limit. We have also proposed the method to obtain the asymptotically optimal quantizer minimizing the asymptotic SER. From the result, we can also optimize the parameters in the objective function of the Box-SOAV optimization. Simulation results show that the empirical reconstruction performance agrees well with the theoretical result in large-scale problems.

In Chapter 6, we have applied the idea of the AMP algorithm for compressed sensing to the SOAV optimization. The resultant DAMP algorithm has low computational complexity and its asymptotic performance can be predicted with the state evolution. We have examined the asymptotic performance of the DAMP algorithm in the noise-free case and derived the required measurement ratio for the perfect reconstruction. Simulation results show that the performance of the DAMP algorithm is very close to the theoretical result in large-scale problems.

7.2 Future Work

There are several remaining topics for the discrete-valued vector reconstruction and its analysis. In this section, we provide the future work on the study in this thesis.

7.2.1 Interpretation of Iterative Approach

In Chapters 2 and 3, we have proposed the iterative approaches for the discrete-valued vector reconstruction, where we iterate the optimization and the parameter update in the objective function. However, the parameter update methods are rather heuristic and there is no theoretical justification for the convergence of the overall algorithm.

For similar iterative approaches such as iterative reweighted least squares (IRLS), the relation to the corresponding nonconvex optimization has been discussed [119, 120]. It would be interesting to reveal the relation between the proposed iterative approaches and the nonconvex optimization behind them. It might also provide a reasonable method for the parameter update.

7.2.2 Extension of Performance Analysis via CGMT

The extension of CGMT is also an interesting research direction. As in the performance analysis via CGMT in Chapter 5, we have to assume that the measurement matrix $\boldsymbol{A}$ is composed of i.i.d. zero mean Gaussian variables. Hence, the result cannot be directly used when the measurement matrix does not satisfy the assumption. For example, even when the measurement matrix has nonzero mean Gaussian elements, we would need to modify the result in some manner. In the context of compressed sensing, however, the performance of ℓ_1 optimization with the nonzero mean measurement matrix has been analyzed via the replica method in developed in statistical mechanics [121]. Since the replica method is not rigorous in part, it would be an interesting topic to prove the same result as in [121] via the rigorous approach using CGMT. As for message passing-based approaches, it has been shown that the performance of EP-based algorithm can be predicted for unitary invariant measurement matrices [51, 53]. It would also be valuable to obtain the analytical result for optimization-based methods with such class of measurement matrices.

The theoretical results for the SCSR optimization and the SSR optimization in Chapter 3 and 4 have not been obtained. The asymptotic analysis for the SCSR optimization in the complex-valued domain might be obtained by using a similar approach to [100]. For the SSR optimization, CGMT cannot be directly applied because the objective function is not convex. However, the upper bound of the reconstruction error might be obtained as it has been provided for the MAP method in [122].

7.2.3 Application of CGMT to Optimization Algorithm

Since some update equations of the optimization algorithm can be written in the form of an optimization problem, which can be solved more easily than the original optimization problem. By analyzing the subproblem in the update equation, it might be possible to obtain the evolution of the reconstruction error in the optimization algorithm like the state evolution for the AMP algorithm. Moreover, this would enable us to determine the asymptotically optimal parameters in the optimization algorithms such as the step size.

7.2.4 Practical Applications

As described in Chapter 1, various problems in communications can be expressed as the discrete-valued vector reconstruction. In some problems, the measurement matrix is not i.i.d. Gaussian and have some structure depending on the system model. Hence, the performance evaluation of the proposed algorithms is required to show the validity in the application. Moreover, the unknown vector in some problems have an additional structure other than the discreteness. In signal detection for MU-MIMO OFDM/SC-CP explained in Section 1.2, for example, the unknown transmitted signal vector has not only the discreteness but also the group sparsity. The use of such additional structure would further improve the performance of the reconstruction.

Bibliography

[1] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[2] A. Chockalingam and B. S. Rajan, *Large MIMO Systems*. Cambridge, U.K.: Cambridge University Press, 2014.

[3] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 4, pp. 1941–1988, Fourthquarter 2015.

[4] S. Verdú, *Multiuser Detection*, 1st ed. New York, NY, USA: Cambridge University Press, 1998.

[5] K. K. Wong, A. Paulraj, and R. D. Murch, "Efficient high-performance decoding for overloaded MIMO antenna systems," *IEEE Trans. Wirel. Commun.*, vol. 6, no. 5, pp. 1833–1843, May 2007.

[6] L. Bai, C. Chen, and J. Choi, "Lattice reduction aided detection for underdetermined MIMO systems: A pre-voting cancellation approach," in *Proc. IEEE 71st Vehicular Technology Conference*, May 2010, pp. 1–5.

[7] T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Low-complexity near-optimal signal detection in underdetermined large-MIMO systems," in *Proc. National Conference on Communications (NCC)*, Feb. 2012, pp. 1–5.

[8] T. Takahashi, S. Ibi, and S. Sampei, "Criterion of adaptively scaled belief for PDA in overloaded MIMO channels," in *Proc. 51st Asilomar Conference on Signals, Systems, and Computers*, Oct. 2017, pp. 1094–1098.

[9] S. Takabe, M. Imanishi, T. Wadayama, R. Hayakawa, and K. Hayashi, "Trainable projected gradient detector for massive overloaded MIMO channels: Data-driven tuning approach," *IEEE Access*, vol. 7, pp. 93 326–93 338, 2019.

[10] J. B. Anderson, F. Rusek, and V. Öwall, "Faster-than-Nyquist signaling," *Proc. IEEE*, vol. 101, no. 8, pp. 1817–1830, Aug. 2013.

[11] H. Shin and J. H. Lee, "Capacity of multiple-antenna fading channels: Spatial fading correlation, double scattering, and keyhole," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2636–2647, Oct. 2003.

[12] J. W. Choi and B. Shim, "New approach for massive MIMO detection using sparse error recovery," in *Proc. IEEE Global Communications Conference*, Dec. 2014, pp. 3754–3759.

[13] ——, "Detection of large-scale wireless systems via sparse error recovery," *IEEE Trans. Signal Process.*, vol. 65, no. 22, pp. 6038–6052, Nov. 2017.

[14] K. Hayashi, A. Nakai, R. Hayakawa, and S. Ha, "Uplink overloaded MU-MIMO OFDM signal detection methods using convex optimization," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2018, pp. 1421–1427.

[15] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications," *IEEE Signal Process. Mag.*, vol. 17, no. 3, pp. 29–48, May 2000.

[16] K. Hayashi, A. Nakai-Kasai, and R. Hayakawa, "An overloaded SC-CP IoT signal detection method via sparse complex discrete-valued vector reconstruction," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2019.

[17] T. Abe and T. Matsumoto, "Space-time turbo equalization in frequency-selective MIMO channels," *IEEE Trans. Veh. Technol.*, vol. 52, no. 3, pp. 469–475, May 2003.

[18] N. Souto and R. Dinis, "MIMO detection and equalization for single-carrier systems using the alternating direction method of multipliers," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1751–1755, Dec. 2016.

[19] H. Jafarkhani, *Space-Time Coding: Theory and Practice*, 1st ed. New York, NY, USA: Cambridge University Press, 2010.

[20] B. A. Sethuraman, B. S. Rajan, and V. Shashidhar, "Full-diversity, high-rate space-time block codes from division algebras," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2596–2616, Oct. 2003.

[21] S. K. Mohammed, A. Zaki, A. Chockalingam, and B. S. Rajan, "High-rate space–time coded large-MIMO systems: Low-complexity detection and channel estimation," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 6, pp. 958–974, Dec. 2009.

[22] N. Srinidhi, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "Low-complexity near-ML decoding of large non-orthogonal STBCs using reactive tabu search," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2009, pp. 1993–1997.

[23] M. Suneel, P. Som, A. Chockalingam, and B. S. Rajan, "Belief propagation based decoding of large non-orthogonal STBCs," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2009, pp. 2003–2007.

[24] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "Low-complexity near-MAP decoding of large non-orthogonal STBCs using PDA," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2009, pp. 1998–2002.

[25] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454–465, Feb. 2011.

[26] H. Sasahara, K. Hayashi, and M. Nagahara, "Multiuser detection based on MAP estimation with sum-of-absolute-values relaxation," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5621–5634, Nov. 2017.

[27] J. E. Mazo, "Faster-than-Nyquist signaling," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1451–1462, Oct. 1975.

[28] F. Han, M. Jin, and H. Zou, "Binary symbol recovery via ℓ_∞ minimization in faster-than-Nyquist signaling systems," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5282–5293, Oct. 2014.

[29] H. Sasahara, K. Hayashi, and M. Nagahara, "Symbol detection for faster-than-Nyquist signaling by sum-of-absolute-values optimization," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1853–1857, Dec. 2016.

[30] M. Pohst, "On the computation of lattice vectors of minimal length, successive minima and reduced bases with applications," *SIGSAM Bull*, vol. 15, no. 1, pp. 37–44, Feb. 1981.

[31] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.

[32] K. V. Vardhan, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity detector for large MIMO systems and multicarrier CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 473–485, Apr. 2008.

[33] P. Li and R. D. Murch, "Multiple output selection-LAS algorithm in large MIMO systems," *IEEE Commun. Lett.*, vol. 14, no. 5, pp. 399–401, May 2010.

[34] T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Random-restart reactive tabu search algorithm for detection in large-MIMO systems," *IEEE Commun. Lett.*, vol. 14, no. 12, pp. 1107–1109, Dec. 2010.

[35] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Francisco, CA, USA: Morgan Kaufmann, 1988.

[36] F. R. Kschischang, B. J. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[37] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A: Math. Gen.*, vol. 36, no. 43, pp. 11 111–11 121, Oct. 2003.

[38] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *PNAS*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[39] ——, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. IEEE Information Theory Workshop*, Jan. 2010, pp. 1–5.

[40] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[41] K. Hayashi, M. Nagahara, and T. Tanaka, "A user's guide to compressed sensing for communications systems," *IEICE Trans. Commun.*, vol. E96-B, no. 3, pp. 685–712, Mar. 2013.

[42] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2015, pp. 1227–1231.

[43] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jul. 2011, pp. 2168–2172.

[44] M. Bayati and A. Montanari, "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[45] A. Kuriya and T. Tanaka, "Performance degradation of AMP for small-sized problems," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2015, pp. 2802–2806.

[46] ——, "Effects of the approximations from BP to AMP for small-sized problems," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jul. 2016, pp. 770–774.

[47] F. Caltagirone, L. Zdeborová, and F. Krzakala, "On convergence of approximate message passing," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2014, pp. 1812–1816.

[48] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 2021–2025.

[49] S. Rangan, P. Schniter, and A. Fletcher, "On the convergence of approximate message passing with arbitrary matrices," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2014, pp. 236–240.

[50] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional MIMO systems," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2840–2849, Aug. 2014.

[51] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 1588–1592.

[52] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[53] K. Takeuchi, "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 368–386, Jan. 2020.

[54] P. H. Tan, L. K. Rasmussen, and T. J. Lim, "Constrained maximum-likelihood detection in CDMA," *IEEE Trans. Commun.*, vol. 49, no. 1, pp. 142–153, Jan. 2001.

[55] A. Yener, R. D. Yates, and S. Ulukus, "CDMA multiuser detection: A nonlinear programming approach," *IEEE Trans. Commun.*, vol. 50, no. 6, pp. 1016–1024, Jun. 2002.

[56] C. Thrampoulidis, W. Xu, and B. Hassibi, "Symbol error rate performance of box-relaxation decoders in massive MIMO," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3377–3392, Jul. 2018.

[57] A. Aïssa-El-Bey, D. Pastor, S. M. A. Sbaï, and Y. Fadlallah, "Sparsity-based recovery of finite alphabet solutions to underdetermined linear systems," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2008–2018, Apr. 2015.

[58] Y. Fadlallah, A. Aïssa-El-Bey, K. Amis, D. Pastor, and R. Pyndiah, "New iterative detector of MIMO transmission using sparse decomposition," *IEEE Trans. Veh. Technol.*, vol. 64, no. 8, pp. 3458–3464, Aug. 2015.

[59] Z. Hajji, A. Aïssa-El-Bey, and K. Amis, "Simplicity-based recovery of finite-alphabet signals for large-scale MIMO systems," *Digital Signal Processing*, vol. 80, pp. 70–82, Sep. 2018.

[60] M. Nagahara, "Discrete signal reconstruction by sum of absolute values," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1575–1579, Oct. 2015.

[61] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ser. Springer Optimization and Its Applications. New York, NY: Springer New York, 2011, vol. 49, pp. 185–212.

[62] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[63] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, May 2008.

[64] T. Wo and P. A. Hoeher, "A simple iterative gaussian detector for severely delay-spread MIMO channels," in *Proc. IEEE International Conference on Communications*, Jun. 2007, pp. 4598–4603.

[65] R. Hayakawa and K. Hayashi, "Error recovery for massive MIMO signal detection via reconstruction of discrete-valued sparse vector," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E100-A, no. 12, pp. 2671–2679, Dec. 2017.

[66] R. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, Jan. 1962.

[67] D. MacKay and R. Neal, "Near Shannon limit performance of low density parity check codes," *Electron. Lett.*, vol. 33, no. 6, pp. 457–458, Mar. 1997.

[68] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, Jan. 1976.

[69] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, Apr. 1992.

[70] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found Trends Mach Learn*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[71] L. Li, X. Wang, and G. Wang, "Alternating direction method of multipliers for separable convex optimization of real functions in complex variables," *Math. Probl. Eng.*, 2015.

[72] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc. F - Commun. Radar Signal Process.*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[73] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, "$L_{1/2}$ regularization," *Sci. China Inf. Sci.*, vol. 53, no. 6, pp. 1159–1169, Jun. 2010.

[74] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[75] Y. Zhang and W. Ye, "$L_{2/3}$ regularization: Convergence of iterative thresholding algorithm," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 350–357, Nov. 2015.

[76] F. Chen, L. Shen, and B. W. Suter, "Computing the proximity operator of the ℓ_p norm with $0 < p < 1$," *IET Signal Process.*, vol. 10, no. 5, pp. 557–565, 2016.

[77] P. Yin, Y. Lou, Q. He, and J. Xin, "Minimization of ℓ_{1-2} for compressed sensing," *SIAM J. Sci. Comput.*, vol. 37, no. 1, pp. A536–A563, Jan. 2015.

[78] Y. Lou and M. Yan, "Fast L1–L2 minimization via a proximal operator," *J. Sci. Comput.*, vol. 74, no. 2, pp. 767–785, Feb. 2018.

[79] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, no. 1, pp. 120–145, May 2011.

[80] L. Condat, "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, Aug. 2013.

[81] T. Liu and T. K. Pong, "Further properties of the forward–backward envelope with applications to difference-of-convex programming," *Comput Optim Appl*, vol. 67, no. 3, pp. 489–520, Jul. 2017.

[82] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge, UK ; New York, NY: Cambridge University Press, 2018.

[83] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Math. Program.*, vol. 137, no. 1, pp. 91–129, Feb. 2013.

[84] G. Li and T. K. Pong, "Global convergence of splitting methods for nonconvex composite optimization," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2434–2460, Jan. 2015.

[85] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, Jan. 2016.

[86] D. Hajinezhad, M. Hong, T. Zhao, and Z. Wang, "NESTT: A nonconvex primal-dual splitting method for distributed and stochastic optimization," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 3215–3223.

[87] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *J Sci Comput*, vol. 78, no. 1, pp. 29–63, Jan. 2019.

[88] T. Ikeda, M. Nagahara, and S. Ono, "Discrete-valued control of linear time-invariant systems by sum-of-absolute-values optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 6, pp. 2750–2763, Jun. 2017.

[89] T. Ikeda and M. Nagahara, "Discrete-valued model predictive control using sum-of-absolute-values optimization," *Asian J. Control*, vol. 20, no. 1, pp. 196–206, 2018.

[90] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *Proc. Conference on Learning Theory*, Jun. 2015, pp. 1683–1709.

[91] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high dimensions," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5592–5628, Aug. 2018.

[92] C. Thrampoulidis, A. Panahi, D. Guo, and B. Hassibi, "Precise error analysis of the LASSO," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 3467–3471.

[93] C. Thrampoulidis, A. Panahi, and B. Hassibi, "Asymptotically exact error analysis for the generalized ℓ_2^2-LASSO," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2015, pp. 2021–2025.

[94] I. B. Atitallah, C. Thrampoulidis, A. Kammoun, T. Y. Al-Naffouri, M. Alouini, and B. Hassibi, "The BOX-LASSO with application to GSSK modulation in massive MIMO systems," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 1082–1086.

[95] C. Thrampoulidis, E. Abbasi, W. Xu, and B. Hassibi, "BER analysis of the box relaxation for BPSK signal recovery," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 3776–3780.

[96] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "LASSO with non-linear measurements is equivalent to one with linear measurements," in *Proc. the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 3420–3428.

[97] C. Thrampoulidis and W. Xu, "The performance of box-relaxation decoding in massive MIMO with low-resolution ADCs," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, Jun. 2018, pp. 821–825.

[98] A. M. Alrashdi, I. B. Atitallah, T. Y. Al-Naffouri, and M. Alouini, "Precise performance analysis of the LASSO under matrix uncertainties," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov. 2017, pp. 1290–1294.

[99] A. M. Alrashdi, I. B. Atitallah, T. Ballal, C. Thrampoulidis, A. Chaaban, and T. Y. Al-Naffouri, "Optimum training for MIMO BPSK transmission," in *Proc. IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Jun. 2018, pp. 1–5.

[100] E. Abbasi, F. Salehi, and B. Hassibi, "Performance analysis of convex data detection in MIMO," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 4554–4558.

[101] P. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numer. Anal.*, vol. 16, no. 6, pp. 964–979, Dec. 1979.

[102] D. Pérez and Y. Quintana, "A survey on the Weierstrass approximation theorem," *Divulg. Matemáticas*, vol. 16, no. 1, pp. 231–247, 2008.

[103] J. Ma, X. Yuan, and L. Ping, "Turbo compressed sensing with partial DFT sensing matrix," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 158–161, Feb. 2015.

[104] Z. Xue, J. Ma, and X. Yuan, "Denoising-based turbo compressed sensing," *IEEE Access*, vol. 5, pp. 7193–7204, 2017.

[105] D. L. Donoho, A. Maleki, and A. Montanari, "The noise-sensitivity phase transition in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6920–6941, Oct. 2011.

[106] ——, "Message passing algorithms for compressed sensing: II. Analysis and validation," in *Proc. IEEE Information Theory Workshop*, Jan. 2010, pp. 1–5.

[107] D. Donoho and A. Montanari, "High dimensional robust M-estimation: Asymptotic variance via approximate message passing," *Probab. Theory Relat. Fields*, vol. 166, no. 3, pp. 935–969, Dec. 2016.

[108] M. Advani and S. Ganguli, "An equivalence between high dimensional Bayes optimal inference and M-estimation," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 3378–3386.

[109] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[110] Z. Tian, G. Leus, and V. Lottici, "Detection of sparse signals under finite-alphabet constraints," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 2349–2352.

[111] A. Müller, D. Sejdinovic, and R. Piechocki, "Approximate message passing under finite alphabet constraints," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 3177–3180.

[112] B. Shim, S. Kwon, and B. Song, "Sparse detection with integer constraint using multipath matching pursuit," *IEEE Commun. Lett.*, vol. 18, no. 10, pp. 1851–1854, Oct. 2014.

[113] S. Sparrer and R. F. Fischer, "Enhanced iterative hard thresholding for the estimation of discrete-valued sparse signals," in *Proc. 24th European Signal Processing Conference (EUSIPCO)*, Aug. 2016, pp. 71–75.

[114] N. M. B. Souto and H. A. Lopes, "Efficient recovery algorithm for discrete valued sparse signals using an ADMM approach," *IEEE Access*, vol. 5, pp. 19 562–19 569, 2017.

[115] C. Jeon, A. Maleki, and C. Studer, "On the performance of mismatched data detection in large MIMO systems," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jul. 2016, pp. 180–184.

[116] K. Mimura, "On introducing damping to Bayes optimal approximate message passing for compressed sensing," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec. 2015, pp. 659–662.

[117] A. Montanari and D. Tse, "Analysis of belief propagation for non-linear problems: The example of CDMA (or: How to prove Tanaka's formula)," in *Proc. IEEE Information Theory Workshop*, Mar. 2006, pp. 160–164.

[118] S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7464–7474, Dec. 2016.

[119] R. Chartrand and Wotao Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Las Vegas, NV, USA: IEEE, Mar. 2008, pp. 3869–3872.

[120] N. Mourad and J. P. Reilly, "Minimizing nonconvex functions for sparse vector reconstruction," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3485–3496, Jul. 2010.

[121] T. Tanaka, "Performance analysis of L1-norm minimization for compressed sensing with non-zero-mean matrix elements," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2018, pp. 401–405.

[122] C. Thrampoulidis, I. Zadik, and Y. Polyanskiy, "A simple bound on the BER of the MAP decoder for massive MIMO systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 4544–4548.

Copyright Notice