Accelerated and Deep Expectation Maximization for One-Bit MIMO-OFDM Detection

Mingjie Shao , *Member, IEEE*, Wing-Kin Ma , *Fellow, IEEE*, Junbin Liu , *Graduate Student Member, IEEE*, and Zihao Huang

Abstract—In this paper we study the expectation maximization (EM) technique for one-bit MIMO-OFDM detection (OMOD). Arising from the recent interest in massive MIMO with one-bit analog-to-digital converters, OMOD is a massive-scale problem. EM is an iterative method that can exploit the OFDM structure to process the problem in a per-iteration efficient fashion. In this study we analyze the convergence rate of EM for a class of approximate maximum-likelihood OMOD formulations, or, in a broader sense, a class of problems involving regression from quantized data. We show how the SNR and channel conditions can have an impact on the convergence rate. We do so by making a connection between the EM and the proximal gradient methods in the context of OMOD. This connection also gives us insight to build new accelerated and/or inexact EM schemes. The accelerated scheme has faster convergence in theory, and the inexact scheme provides us with the flexibility to implement EM more efficiently, with convergence guarantee. Furthermore we develop a deep EM algorithm, wherein we take the structure of our inexact EM algorithm and apply deep unfolding to train an efficient structured deep net. Simulation results show that our accelerated exact/inexact EM algorithms run much faster than their standard EM counterparts, and that the deep EM algorithm gives promising detection and runtime performances.

Index Terms—One-bit MIMO-OFDM detection, expectation maximization, convergence analysis, deep unfolding.

I. INTRODUCTION

IN massive multiple-input multiple-output (MIMO), we have recently seen enormous interest in techniques that are related to low-resolution or coarsely quantized (CQ) signals from a fundamental signal processing viewpoint. This trend is driven by

our interest in replacing high-resolution analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) with lower resolution ones, so that we can reduce the hardware cost and energy consumption of the radio-frequency (RF) front ends. In the conventional scenario of high-resolution ADCs/DACs, quantization errors are insignificant and are typically treated as part of the background noise. In the low-resolution scenario, recent research has suggested that we should exploit the low-resolution or quantization structure to better harness the problem. For example, for MIMO downlink precoding with one-bit DACs, a popular direction is to directly design the one-bit MIMO transmit signals by discrete optimization; see, e.g., [1], [2], [3], [4] and the references therein. In addition to exploiting the low-resolution structure, there have been endeavors that apply Sigma-Delta modulation—a widely-used technique in temporal DACs/ADCs—to spatially shape the quantization errors such that users in a designated angular sector will experience much less quantization error effects [5], [6].

Another meaningful problem in massive MIMO with low-resolution ADCs/DACs is MIMO detection. The problem typically appears in the uplink, and the goal is to detect signals transmitted by different users over the same RF band. While numerous MIMO detection methods were already developed in the conventional high-resolution ADC scenario, they do not assume the presence of quantization and may not be directly applied to the low-resolution ADC scenario. We have indeed seen research on applying linear receivers to the low-resolution scenario [7], [8], [9], but recent research appears to favor exploiting the quantization structure. The key idea that underlies the majority of low-resolution or CQ MIMO detection methods is that we consider maximum likelihood (ML) or related inference approaches, and the likelihood function is derived from quantized signals (quantization is treated as a function, and we do not model quantization errors as noise). This rationale gave rise to a variety of CQ MIMO detection methods; e.g., expectation maximization (EM) [10], convex and non-convex proximal gradient (PG) methods [11], [12], [13], deep learning [13], [14], [15], and posterior inference via approximate message passing [16] and variational Bayes [17].

A. The Problem of Interest: CQ MIMO-OFDM Detection

In MIMO detection it is common to assume frequency-flat channels. Yet, in the current mobile systems, we consider frequency-selective channels. This is not a problem in the

high-resolution scenario, because the current systems usually use the orthogonal frequency division multiplexing (OFDM) scheme, and OFDM breaks the frequency-selective channels into many parallel frequency-flat channels. This means that the MIMO detection methods for frequency-flat channels can be directly applied. Unfortunately, the same does not hold for the CQ case. The presence of quantization destroys the structure required to turn the problem into parallel frequency-flat channels. Subsequently, when we consider MIMO-OFDM detection with low-resolution ADCs, we have to engage with the full form of the problem. The arising difficulty is that the problem is a massive-scale problem; the problem size scales with the OFDM size, which is of the order of hundreds. While in principle we can directly apply the CQ MIMO detection methods in the frequency-flat channel case to the MIMO-OFDM case, it is computationally infeasible to do so due to the massive problem size. It is necessary to exploit the OFDM structure to efficiently realize MIMO detection methods.

There are far fewer studies for CQ MIMO-OFDM detection [10], [17], [18], [19]. In [18], the PG method for the box ML relaxation formulation was developed. The PG method is able to exploit the OFDM structure to have per-iteration efficient implementation. In [10], [19], the EM method for the Gaussian maximum a posteriori formulation was built. The EM method is iterative, and at each iteration we solve a problem—the so-called M-step problem—that takes the same form as MIMO-OFDM *without* quantization. Hence we can exploit the OFDM structure to have per-iteration efficient implementation, and this makes the EM method appealing in OMOD. The same benefit was also seen in the application of the variational Bayes method (which is related to EM) to CQ MIMO-OFDM detection [17].

B. This Work

In this paper we study the EM technique for CQ MIMO-OFDM detection. In particular, we consider MIMO-OFDM detection with one-bit ADCs, or simply one-bit MIMO-OFDM detection (OMOD). We are interested in EM convergence rate analysis for a class of approximate ML OMOD formulations. This aspect has not been seriously studied in CQ MIMO detection. Our analysis draws insights from the analysis techniques in PG methods [20], [21], [22], [23]. We will reveal a hidden connection between the PG and EM methods, and from that connection we establish EM convergence results by taking key ideas from the PG analysis techniques. We will show that (a) EM converges at a rate of at least $1/k$, where k denotes the iteration number; that (b) EM should converge faster than PG, in theory. Also, we show how the SNR and channel conditions can have an impact on the convergence rate. In particular, our analysis indicates that the convergence rate is slower as the SNR is higher; this applies to both PG and EM. This high-SNR slow convergence phenomenon was noticed by numerical results in previous studies [13], [18], and, to the best of our knowledge, this is the first time we see a theoretical result for such phenomenon.

Our EM convergence study also leads to new OMOD algorithms. In first-order optimization, it is well-known that the PG method can be accelerated to a convergence rate of at least $1/k^2$ if we employ extrapolation [20], [21]; we assume convex problems. Using the connection between the PG and EM methods, we apply the acceleration idea to EM. The result is an accelerated EM scheme for OMOD, which has the $1/k^2$ convergence rate guarantee and runs faster than the standard EM as will be shown in our numerical results. In addition we consider an inexact EM scheme, where we want to solve the M-step problems not too precisely to save complexity. By taking insight from the study of inexact PG [22], we provide theoretical conditions under which the inexact EM scheme will have convergence guarantees. The inexact EM scheme can be used together with acceleration, and our numerical results will demonstrate that the accelerated inexact EM algorithm for OMOD has significant merits in runtime savings.

Lately we have seen much interest in applying deep learning to MIMO detection [13], [15], [24], [25], [26], [27]. The most successful approach at present is deep unfolding [28]. The principle is to see a model-based algorithm as a deep net, take the structure, and use data-driven learning to try to learn a better deep net. We apply deep unfolding to our inexact EM algorithm for OMOD; that is, simply speaking, we take the structure of how efficient EM can handle MIMO-OFDM, and we use that to build an efficient highly-structured deep net. Our numerical results will show that the deep inexact EM algorithm is competitive in both detection and runtime performances, compared to the previous OMOD algorithms.

It is worth noting that the convergence analysis and the accelerated and inexact EM schemes in this study are, in a broad sense, also applicable to a variety of formulations arising from regression from CQ data, such as CQ MIMO channel estimation [29], [30] and CQ compressive sensing [31].

C. Related Works

We should discuss related works. Our EM convergence analysis plays a large part in this study, and one may question if the existing convergence analyses [32], [33], [34], [35] provide the same or better results. The existing studies mostly consider a general framework, covering a variety of optimization schemes and a rich class of problems. Our study, on the other hand, exploits the specific structure of our OMOD problem; e.g., the previously mentioned result of high-SNR slow convergence is unique to our problem. In terms of the analysis techniques used, our analysis is considered closest to Mairal's study [34], which considers a general majorization minimization (MM) framework and can cover EM. Mairal's study takes insight from the PG analysis techniques through some assumptions. We also draw insight from the PG analysis techniques, but we do so by best using our problem's specific structure. As we will discuss in details, our result arguably provides better characterization of the convergence behaviors.

One key result in this study is accelerated EM for OMOD. Indeed, Mairal's study [34] has described the same kind of acceleration for MM. To the best of our knowledge, no application was provided in Mairal's study and in subsequent studies. Our application is probably the first, showing the efficiency

of such accelerated EM or MM. We should also discuss the application of deep unfolding. While deep unfolding has been widely considered in MIMO detection, it has not been used in CQ MIMO-OFDM detection.

The preliminary version of this study was presented in a conference [36]. It studied inexact EM as a working idea. It does not consider EM convergence analysis, accelerated EM and deep EM, which are the main contributions in this paper.

D. Organization and Notations

This paper is organized as follows. Section II, III and IV review the EM basics, the MIMO-OFDM detection problem and the OMOD problem, respectively. In Section V, we perform EM convergence analysis and establish the accelerated and inexact EM methods. Section VI develops an accelerated inexact EM algorithms for OMOD. Section VII builds a deep inexact EM algorithm via deep unfolding. Section VIII provides the numerical results, and we conclude in Section IX.

Some basic notations are described as follows. The sets of real, non-negative, non-positive and complex numbers are denoted by $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_-$ and $\mathbb{C}$, respectively; we denote $\mathrm{j} = \sqrt{-1}$; boldface lowercase letters, e.g., $\boldsymbol{x}$, represent column vectors; the ith element of a vector $\boldsymbol{x}$ is denoted by x_i; boldface capital letters, e.g., $\boldsymbol{X}$, represent matrices; the superscripts $\top$, H, -1, and $\dagger$ denote the transpose, Hermitian transpose, inverse, and pseudoinverse, respectively; $\Re(\boldsymbol{x})$ and $\Im(\boldsymbol{x})$ denote the real and imaginary parts of $\boldsymbol{x}$, respectively; $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ means that $\boldsymbol{x} = [\,\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_m^\top\,]^\top$; $\|\cdot\|$ denotes the Euclidean norm; $\boldsymbol{0}$ and $\boldsymbol{I}$ denote an all-zero vector and an identity matrix, respectively; $\mathrm{Diag}(\boldsymbol{x})$ denotes a diagonal matrix whose (i,i)th element is x_i; $\boldsymbol{x} \sim \mathcal{D}$ means that $\boldsymbol{x}$ is a random vector following a probability distribution $\mathcal{D}$; $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the distribution function corresponding to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

II. A REVIEW OF EM FOR QUANTIZED REGRESSION

Before we delve into the details of OMOD, we would like to first review the classic concepts of expectation maximization (EM) [37] in the context of quantized linear regression (QLR). In particular, the arising notion of iterative linear regression is worth appreciating, and we will need some of the details later.

We consider a basic QLR problem for which we have an observation $\boldsymbol{y} \in \{\pm 1\}^m$ following a model

$$\boldsymbol{y} = \mathrm{sgn}(\boldsymbol{x}), \quad \boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{v}, \tag{1}$$

where $\mathrm{sgn} : \mathbb{R}^m \to \{\pm 1\}^m$ is the sign function; $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is known; $\boldsymbol{\theta} \in \mathbb{R}^n$ is a model parameter; $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ is noise; σ^2 is known. Note that $\boldsymbol{x}$ is unobserved. The problem is to estimate $\boldsymbol{\theta}$ from $\boldsymbol{y}$. As we will see later, OMOD can be viewed as an instance of QLR, with $\boldsymbol{A}$ and $\boldsymbol{\theta}$ having certain structures. We consider the maximum-likelihood (ML) estimation

$$\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^n} L(\boldsymbol{\theta}) := \log p(\boldsymbol{y}|\boldsymbol{\theta}), \tag{2}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the distribution of $\boldsymbol{y}$ given $\boldsymbol{\theta}$; the same notation will be used to denote the distributions of other random variables. It can be shown that $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^m p(y_i|\boldsymbol{\theta})$,

$$p(y_i|\boldsymbol{\theta}) = \int_{\mathcal{X}(y_i)} p(x_i|\boldsymbol{\theta}) \mathrm{d}x_i, \quad p(x_i|\boldsymbol{\theta}) = \mathcal{N}(x_i; \boldsymbol{a}_i^\top \boldsymbol{\theta}, \sigma^2),$$

where $\boldsymbol{a}_i$ denotes the ith row of $\boldsymbol{A}$; $\mathcal{X}(y_i) = \mathbb{R}_+$ if $y = 1$, and $\mathcal{X}(y_i) = \mathbb{R}_-$ if $y = -1$. This leads to

$$L(\boldsymbol{\theta}) = \sum_{i=1}^m \log \Phi\left(\frac{y_i \boldsymbol{a}_i^\top \boldsymbol{\theta}}{\sigma}\right), \tag{3}$$

where $\Phi(z) = \int_{-\infty}^z e^{-t^2/2} \mathrm{d}t / \sqrt{2\pi}$ is the standard cumulative Gaussian distribution. It is known that L is concave [11], [38].

The ML estimation problem (2) has no closed-form solution in general, and EM is a popular way to solve the ML problem. Intuitively, the idea is to use the fact that estimating $\boldsymbol{\theta}$ is easy if $\boldsymbol{x}$ is observable. To describe the EM method, let

$$G(\boldsymbol{\theta}|\boldsymbol{\theta}') = \underbrace{\mathbb{E}_{\boldsymbol{x}}[\log(p(\boldsymbol{x}|\boldsymbol{\theta}))|\boldsymbol{y},\boldsymbol{\theta}']}_{:=Q(\boldsymbol{\theta}|\boldsymbol{\theta}')} - \underbrace{\mathbb{E}_{\boldsymbol{x}}[\log(p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}'))|\boldsymbol{y},\boldsymbol{\theta}']}_{:=H(\boldsymbol{\theta}')}.$$

Here, $\mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})|\boldsymbol{y},\boldsymbol{\theta}'] = \int_{\mathcal{X}} f(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}') \mathrm{d}\boldsymbol{x}$ denotes a conditional expectation of $\boldsymbol{x}$; $\mathcal{X}$ denotes the support of $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}')$. It is known that $G(\boldsymbol{\theta}|\boldsymbol{\theta}') \leq L(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$, and $G(\boldsymbol{\theta}'|\boldsymbol{\theta}') = L(\boldsymbol{\theta}')$ for all $\boldsymbol{\theta}'$; i.e., $G(\boldsymbol{\theta}|\boldsymbol{\theta}')$ is a lower bound of $L(\boldsymbol{\theta})$. EM considers a successive lower-bound approximation

$$\boldsymbol{\theta}^{k+1} \in \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^n} G(\boldsymbol{\theta}|\boldsymbol{\theta}^k), \quad k = 0, 1, 2, \ldots \tag{4}$$

An important question is whether $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ is tractable. By noting $p(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{A}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$, one can verify that

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = -\frac{1}{2\sigma^2} \|\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}'] - \boldsymbol{A}\boldsymbol{\theta}\|^2 + \mathrm{const.}, \tag{5}$$

where "const." does not depend on $\boldsymbol{\theta}$. It can be shown that

$$\mathbb{E}_{x_i}[x_i|\boldsymbol{y},\boldsymbol{\theta}'] = \boldsymbol{a}_i^\top \boldsymbol{\theta}' + y_i \sigma \frac{\phi(y_i \boldsymbol{a}_i^\top \boldsymbol{\theta}'/\sigma)}{\Phi(y_i \boldsymbol{a}_i^\top \boldsymbol{\theta}'/\sigma)}, \tag{6}$$

where $\phi(t) = e^{-t^2/2}/\sqrt{2\pi}$; see, e.g., [29], for details. To prove (6), one needs to show that $p(x_i|y_i,\boldsymbol{\theta})$ is a truncated Gaussian distribution, namely, $p(x_i|y_i,\boldsymbol{\theta}) = p(x_i|\boldsymbol{\theta})/p(y_i|\boldsymbol{\theta})$, with $\mathcal{X}(y_i)$ being the support; that $\mathbb{E}_{x_i}[x_i|\boldsymbol{y},\boldsymbol{\theta}']$ is the mean of $p(x_i|y_i,\boldsymbol{\theta}')$; that a truncated Gaussian mean has an explicit expression [29]. The EM method can hence be written as

$$\boldsymbol{\theta}^{k+1} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^n} \|\boldsymbol{x}^k - \boldsymbol{A}\boldsymbol{\theta}\|^2, \quad k = 0, 1, 2, \ldots, \tag{7}$$

where $\boldsymbol{x}^k = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^k]$ is computed by (6). The result is an *iterative linear regression*: we estimate $\boldsymbol{x}$ by the conditional mean $\boldsymbol{x}^k$, inferred from $\boldsymbol{y}$ and the previous estimate $\boldsymbol{\theta}^k$; and then we estimate $\boldsymbol{\theta}$ by the linear regression in (7), which is easy to solve. This iterative linear regression process looks appealing and makes sense.

Some history should be noted. The EM method for QLR dates back to as early as 1980 in statistics, with application to blood lead data [39]. It also arose in signal processing, with applications to MIMO channel estimation and detection from coarsely quantized signals [10], [19], [29], [30].

III. A REVIEW OF THE MIMO-OFDM BACKGROUND

In this section, we review some background of MIMO-OFDM detection. Section III-A describes the MIMO-OFDM signal model. Section III-B shows the subcarrier decoupling concept, which, as will be seen later, plays a pivotal role in the application of EM to OMOD.

A. MIMO-OFDM Model

Consider a wireless uplink scenario where a number of users simultaneously send signals to a multi-antenna base station (BS). The transmitted signals undergo frequency selective channel effects, upon reception at the BS; and the user side employs the OFDM transmission scheme. The received signals at the BS can be modeled as

$$\boldsymbol{r}_m = \sum_{n=1}^{N} \boldsymbol{H}_{m,n} \boldsymbol{F}^{\mathsf{H}} \boldsymbol{s}_n + \boldsymbol{\nu}_m, \quad m = 1, \dots, M, \quad (8)$$

where $\boldsymbol{r}_m \in \mathbb{C}^W$ is a block of discrete-time signals received by the BS's mth antenna; W is the block length, or the OFDM size; M is the number of antennas at the BS; $\boldsymbol{F} \in \mathbb{C}^{W \times W}$ is the unitary discrete Fourier transform (DFT) matrix; $\boldsymbol{s}_n \in \mathbb{C}^W$ is a block of symbols sent by user n; N is the number of users; $\boldsymbol{H}_{m,n} \in \mathbb{C}^{W \times W}$ is a circulant matrix, constructed from a discrete-time impulse response $\boldsymbol{h}_{m,n} = (h_{0,m,n}, \dots, h_{W'-1,m,n}, 0, \dots, 0) \in \mathbb{C}^W$ of the channel from user n to the BS's mth antenna, with $W' \ll W$; $\boldsymbol{\nu}_m$ is circular complex Gaussian noise with mean $\boldsymbol{0}$ and covariance $\sigma_{\mathsf{C}}^2 \boldsymbol{I}$, and with $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_M$ being independent. The symbols $s_{w,n}$'s are drawn from a QAM constellation set

$$\mathcal{S} = \{s \in \mathbb{C} \mid \Re(s), \Im(s) \in \{\pm 1, \pm 3, \dots, \pm(2D-1)\}\},$$

where D is a positive integer. The above model is standard, and we refer the reader to the literature (e.g., [40]) for details. For conciseness, we let $\boldsymbol{r} = (\boldsymbol{r}_1, \dots, \boldsymbol{r}_M)$ and express (8) as

$$\boldsymbol{r} = \mathcal{H}\boldsymbol{s} + \boldsymbol{\nu}, \quad (9)$$

where $\boldsymbol{s} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_N)$, $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_M)$,

$$\mathcal{H} = \begin{bmatrix} \boldsymbol{H}_{1,1}\boldsymbol{F}^{\mathsf{H}} & \dots & \boldsymbol{H}_{1,N}\boldsymbol{F}^{\mathsf{H}} \\ \vdots & & \vdots \\ \boldsymbol{H}_{M,1}\boldsymbol{F}^{\mathsf{H}} & \dots & \boldsymbol{H}_{M,N}\boldsymbol{F}^{\mathsf{H}} \end{bmatrix}. \quad (10)$$

B. Subcarrier Decoupling in Unquantized MIMO-OFDM

Consider the problem of ML detection of the symbol vector $\boldsymbol{s}$ from the received signal $\boldsymbol{r}$, given information of the channel $\mathcal{H}$. It is well known that the ML detection problem can be formulated as

$$\hat{\boldsymbol{s}} \in \arg\min_{\boldsymbol{s} \in \mathcal{S}^{NW}} \|\boldsymbol{r} - \mathcal{H}\boldsymbol{s}\|^2. \quad (11)$$

At first look, this is a massive-scale problem; the OFDM size W is often of the order of hundreds. But it is widely known that (11) can be decoupled into many smaller problems. Since $\boldsymbol{H}_{m,n}$ is circulant, we have the eigendecomposition

$$\boldsymbol{H}_{m,n} = \boldsymbol{F}^{\mathsf{H}} \boldsymbol{D}_{m,n} \boldsymbol{F}, \quad \boldsymbol{D}_{m,n} = \mathrm{Diag}(\tilde{\boldsymbol{h}}_{m,n}), \quad (12)$$

where $\tilde{\boldsymbol{h}}_{m,n} = \boldsymbol{F}\boldsymbol{h}_{m,n}$. Let $\tilde{\boldsymbol{r}}_m = \boldsymbol{F}\boldsymbol{r}_m$. One can show that

$$\|\boldsymbol{r} - \mathcal{H}\boldsymbol{s}\|^2 = \sum_{m=1}^{M} \left\| \tilde{\boldsymbol{r}}_m - \sum_{n=1}^{N} \boldsymbol{D}_{m,n} \boldsymbol{s}_n \right\|^2$$

$$= \sum_{w=1}^{W} \|\check{\boldsymbol{r}}_w - \check{\boldsymbol{H}}_w \check{\boldsymbol{s}}_w\|^2,$$

where $\check{\boldsymbol{r}}_w = (\tilde{r}_{1,w}, \dots, \tilde{r}_{M,w}) \in \mathbb{C}^M$, $\check{\boldsymbol{H}}_w = [\tilde{h}_{w,m,n}]_{m,n} \in \mathbb{C}^{M \times N}$ and $\check{\boldsymbol{s}}_w = (s_{w,1}, \dots, s_{w,N}) \in \mathbb{C}^N$ are the received signal, MIMO channel response and symbol vector at subcarrier w, respectively. We can hence rewrite (11) as

$$\min_{\check{\boldsymbol{s}}_w \in \mathcal{S}^N} \|\check{\boldsymbol{r}}_w - \check{\boldsymbol{H}}_w \check{\boldsymbol{s}}_w\|^2, \text{ for } w = 1, \dots, W. \quad (13)$$

We see that each problem in (13) has a much smaller size, and can be dealt with independently.

The decomposition of (11) into (13) will be called *subcarrier decoupling,* in what follows. In MIMO-OFDM detection, subcarrier decoupling is commonly used to break the problem into many parallel smaller-size MIMO detection problems.

IV. OMOD PROBLEM STATEMENT AND PRIOR STUDY

In this section, we describe the background of OMOD. Sections IV-A and IV-B provide the problem formulation and statement. Sections IV-C and IV-D review the existing methods by proximal gradient and EM, respectively.

A. One-Bit MIMO-OFDM Detection (OMOD)

OMOD considers the same MIMO-OFDM scenario in Section III-A, but with one-bit ADCs. In the MIMO-OFDM detection in Section III, we assume that signal acquisitions at the BS have negligible quantization errors; to put it another way, the BS is equipped with high-resolution ADCs. For the one-bit ADC case, the acquired signals are

$$\boldsymbol{q}_m = \mathrm{sgn}(\Re(\boldsymbol{r}_m)) + \mathrm{j} \cdot \mathrm{sgn}(\Im(\boldsymbol{r}_m)), \quad m = 1, \dots, M, \quad (14)$$

where the $\boldsymbol{r}_m$'s are the unquantized MIMO-OFDM signals in (8). As mentioned in the Introduction, the one-bit ADC case is motivated by the massive MIMO scenario where we want to use cheap ADCs to reduce the hardware cost and energy consumption of the RF front ends.

The OMOD problem is formulated as follows. Applying the signal model (9) to (14) gives

$$\boldsymbol{x} := \begin{bmatrix} \Re(\boldsymbol{r}) \\ \Im(\boldsymbol{r}) \end{bmatrix} = \underbrace{\begin{bmatrix} \Re(\mathcal{H}) & -\Im(\mathcal{H}) \\ \Im(\mathcal{H}) & \Re(\mathcal{H}) \end{bmatrix}}_{:=\boldsymbol{A}} \underbrace{\begin{bmatrix} \Re(\boldsymbol{s}) \\ \Im(\boldsymbol{s}) \end{bmatrix}}_{:=\boldsymbol{\theta}} + \underbrace{\begin{bmatrix} \Re(\boldsymbol{\nu}) \\ \Im(\boldsymbol{\nu}) \end{bmatrix}}_{:=\boldsymbol{v}}, \quad (15)$$

$$\boldsymbol{y} := \begin{bmatrix} \Re(\boldsymbol{q}) \\ \Im(\boldsymbol{q}) \end{bmatrix} = \mathrm{sgn}(\boldsymbol{x}), \quad (16)$$

where $\boldsymbol{q} = (\boldsymbol{q}_1, \dots, \boldsymbol{q}_M)$, $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, $\sigma^2 = \sigma_{\mathsf{C}}^2/2$. The OMOD signal model in (15)–(16) is seen to take the form of the basic QLR model in (1) in Section II, and hence we can apply

the results in Section II to OMOD. Specifically, the ML detection of the symbol vector $\boldsymbol{\theta}$ from the one-bit MIMO-OFDM signal $\boldsymbol{y}$, given information of $\mathcal{H}$ and σ^2, can be formulated as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{2NW}} f(\boldsymbol{\theta}) := -\sum_{i=1}^{2MW} \log \Phi(\boldsymbol{b}_i^\top \boldsymbol{\theta})$$

$$\text{s.t. } \theta_i \in \mathcal{D} := \{\pm 1, \pm 3, \dots, \pm(2D-1)\}, \forall i, \quad (17)$$

where $\boldsymbol{b}_i = y_i \boldsymbol{a}_i / \sigma$; $\boldsymbol{a}_i$ is the ith row of $\boldsymbol{A}$; or,

$$\boldsymbol{B} := [\, \boldsymbol{b}_1, \dots, \boldsymbol{b}_{2MW} \,]^\top = \tfrac{1}{\sigma} \text{Diag}(\boldsymbol{y}) \boldsymbol{A}. \quad (18)$$

There is an important aspect to note. The subcarrier decoupling in Section III-B, which allows us to decouple the massive-scale unquantized MIMO-OFDM problem into many smaller problems, does not work for OMOD. This is because f in (17) does not possess the structure required to apply subcarrier decoupling. This means that we will have to confront the massive-scale nature of MIMO-OFDM in the one-bit case.

B. Problem Statement

Our study will revolve around the following formulation

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{2NW}} f(\boldsymbol{\theta}) + h(\boldsymbol{\theta}), \quad (19)$$

where $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a penalty function that takes a component-wise separable form

$$h(\boldsymbol{\theta}) = \sum_{i=1}^{2NW} h(\theta_i). \quad (20)$$

The above formulation is an approximation of the ML OMOD in (17). Consider the following two examples.

1) Gaussian Maximum a Posteriori (GMAP) [10]: Consider

$$h(\boldsymbol{\theta}) = \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2, \quad (21)$$

where we choose $\lambda^{-1} = \mathbb{E}[|\theta|^2] = ((2D)^2 - 1)/3$, assuming θ being uniformly distributed on $\mathcal{D}$. The corresponding formulation (19) is maximum a posteriori estimation under a Gaussian prior; i.e., $\max_{\boldsymbol{\theta}} \log(p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))$ with prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{0}, \lambda^{-1}\boldsymbol{I})$. This means that we approximate the symbols as Gaussian random variables.

2) Box [18]: Consider

$$h(\boldsymbol{\theta}) = \mathbb{I}_{[-U,U]^{2NW}}(\boldsymbol{\theta}), \quad U = 2D - 1, \quad (22)$$

where $\mathbb{I}_{\mathcal{X}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathcal{X}$, and $\mathbb{I}_{\mathcal{X}}(\boldsymbol{x}) = \infty$ if $\boldsymbol{x} \notin \mathcal{X}$. This corresponds to a relaxation of the ML OMOD detector (17) by replacing the constellation set $\mathcal{D}$ with an interval $[-U, U]$.

We are interested in studying how problem (19), which has a massive scale, can be handled in an efficient fashion. The next two subsections describe two existing methods for (19).

C. Existing Method: Proximal Gradient

One way to handle problem (19) is to apply the proximal gradient (PG) method [23]. The PG method for (19) is given by

$$\boldsymbol{\theta}^{k+1} = \text{prox}_{\eta_k h}(\boldsymbol{\theta}^k - \eta_k \nabla f(\boldsymbol{\theta}^k)), \quad k = 0, 1, 2, \dots, \quad (23)$$

where $\eta_k > 0$ is a step size; $\nabla f(\boldsymbol{\theta})$ is the gradient of f at $\boldsymbol{\theta}$, and it can be shown that

$$\nabla f(\boldsymbol{\theta}) = -\sum_{i=1}^{2MW} \frac{\phi(\boldsymbol{b}_i^\top \boldsymbol{\theta})}{\Phi(\boldsymbol{b}_i^\top \boldsymbol{\theta})} \boldsymbol{b}_i = -\boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}), \quad (24)$$

for which $\psi(\boldsymbol{z}) = (\psi(z_1), \dots, \psi(z_m))$, $\psi(z) = \phi(z)/\Phi(z)$;

$$\text{prox}_h(\boldsymbol{x}) \in \arg \min_{\boldsymbol{z} \in \mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 + h(\boldsymbol{z})$$

is the proximal operator associated with h. We typically consider cases for which $\text{prox}_h(\boldsymbol{x})$ is easy to compute. For example, for the box case $h = \mathbb{I}_{[-U,U]^n}$, we have

$$[\text{prox}_h(\boldsymbol{x})]_i = \begin{cases} U, & x_i \geq U \\ x_i, & x_i \in [-U, U] \\ -U, & x_i \leq -U \end{cases} \quad (25)$$

We want to examine the complexity of the PG method. The most expensive part lies in computing the gradients $\nabla f(\boldsymbol{\theta}^k)$. At first sight, computing $\nabla f(\boldsymbol{\theta})$ in (24) seems expensive since $\boldsymbol{B}$ is very large. But one can use the OFDM structure to reduce the complexity. It was shown in [18] that $\nabla f(\boldsymbol{\theta})$ can be computed with a complexity of

$$\mathcal{O}(2MW(\log(W) + N + C)), \quad (26)$$

where C denotes the complexity to numerically compute Φ, given a precision[1]. For the reader's convenience, in the supplemental material we describe the details with the computation of $\nabla f(\boldsymbol{\theta})$. Simply speaking, to compute $\nabla f(\boldsymbol{\theta})$ efficiently, we need to run fast Fourier transforms (FFTs) or inverse FFTs (IFFTs) for a total of $2M$ times. To sum up, the per-iteration complexity of the PG method is (26).

The PG method for OMOD was studied in [18], wherein the efficient FFT/IFFT-based method for computing the gradients was proposed, and the box relaxation was considered.

D. Existing Method: Expectation Maximization

Another way to handle problem (19) is to apply the EM method. By following the EM method reviewed in Section II, we can show the following. In accordance with the E-step (5) of the EM method in Section II, the EM surrogate function of the objective function f of the OMOD problem (19) is given by

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}') = \frac{1}{2\sigma^2}\|\boldsymbol{x}' - \boldsymbol{A}\boldsymbol{\theta}\|^2 + \text{const.}$$
$$= \frac{1}{2\sigma^2}\|\boldsymbol{r}' - \mathcal{H}\boldsymbol{s}\|^2 + \text{const.} := g(\boldsymbol{s}|\boldsymbol{s}'), \quad (27)$$

where $\boldsymbol{x}' = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}']$ is a conditional mean of the unquantized signal $\boldsymbol{x}$ in (15)–(16), and, similarly, $\boldsymbol{r}' = \mathbb{E}_{\boldsymbol{r}}[\boldsymbol{r}|\boldsymbol{q}, \boldsymbol{s}']$ is

---

[1]The function Φ has no closed form, and it is typically evaluated by an off-the-shelf algorithm (e.g., `erf` on MATLAB).

a conditional mean of the unquantized MIMO-OFDM signal $\boldsymbol{r}$ in (9). The EM method is given by

$$\boldsymbol{s}^{k+1} \in \arg \min_{\boldsymbol{s} \in \mathbb{C}^{NW}} g(\boldsymbol{s}|\boldsymbol{s}^k) + h(\boldsymbol{s}), \qquad (28)$$

where, for convenience, we denote $h(\boldsymbol{s}) = h(\Re(\boldsymbol{s})) + h(\Im(\boldsymbol{s}))$. The merit of the EM method is that it allows us to apply the subcarrier decoupling in Section III-B. To describe this, let

$$\boldsymbol{r}_m^k = \mathbb{E}_{\boldsymbol{r}_m}[\boldsymbol{r}_m|\boldsymbol{q}, \boldsymbol{s}^k], \ \tilde{\boldsymbol{r}}_m^k = \boldsymbol{F}\boldsymbol{r}_m^k, \ \check{\boldsymbol{r}}_w^k = (\tilde{r}_{1,w}^k, \ldots, \tilde{r}_{M,w}^k). \qquad (29)$$

By the subcarrier decoupling concept in Section III-B, we can decompose (28) into

$$\check{\boldsymbol{s}}_w^{k+1} \in \arg \min_{\check{\boldsymbol{s}}_w \in \mathbb{C}^N} \frac{1}{2\sigma^2}\|\check{\boldsymbol{r}}_w^k - \check{\boldsymbol{H}}_w\check{\boldsymbol{s}}_w\|^2 + h(\check{\boldsymbol{s}}_w), \qquad (30)$$

for $w = 1, \ldots, W$. The basic description of the EM method is complete.

Intuitively, the EM method seems more promising than the PG method in the previous subsection. The EM method has no parameter, such as step size, to select; it inherits the merits of subcarrier decoupling, which seems to suggest better efficiency; it performs more optimization in one iteration, specifically, (30), which seems to suggest faster convergence.

The story is actually more complex, as one tries to unravel. We should not forget the fact that the conditional means $\boldsymbol{r}_m^k$'s need to be computed. Examining the equations closely (see (6), (8), (9), and (15)), one will find that $\boldsymbol{r}_m^k$ is given by

$$\boldsymbol{r}_m^k = \boldsymbol{z}_m^k + \boldsymbol{\zeta}_m^k, \qquad (31a)$$

$$\boldsymbol{z}_m^k = \sum_{n=1}^N \boldsymbol{H}_{m,n}\boldsymbol{F}^\mathsf{H}\boldsymbol{s}_n^k = \boldsymbol{F}^\mathsf{H}(\sum_{n=1}^N \boldsymbol{D}_{m,n}\boldsymbol{s}_n^k), \qquad (31b)$$

$$\Re(\boldsymbol{\zeta}_m^k) = \sigma\left(\Re(\boldsymbol{q}_m) \odot \psi\left(\tfrac{1}{\sigma}\Re(\boldsymbol{q}_m) \odot \Re(\boldsymbol{z}_m^k)\right)\right), \qquad (31c)$$

$$\Im(\boldsymbol{\zeta}_m^k) = \sigma\left(\Im(\boldsymbol{q}_m) \odot \psi\left(\tfrac{1}{\sigma}\Im(\boldsymbol{q}_m) \odot \Im(\boldsymbol{z}_m^k)\right)\right), \qquad (31d)$$

for all m, n, where $\odot$ denotes the element-wise product. It can be verified that the complexity of (29) and (31) is

$$\mathcal{O}(2MW(\log(W) + N/2 + C)); \qquad (32)$$

particularly, we need FFTs and IFFTs for a total of $2M$ times. The complexity in (32) is the per-iteration complexity of EM, without counting the complexity required to solve (30). We see that (32) is *not* much better than the per-iteration complexity (26) for the PG method. When the complexity of solving (30) factors in, the per-iteration complexity of the EM method is likely to be higher than that of PG. The next question then goes to whether EM leads to faster convergence. We will try to answer this question in the subsequent sections.

The EM method for OMOD was studied in [10], wherein the efficient subcarrier-decoupling implementation was proposed, and GMAP was considered. For GMAP, the problems in (30) have closed-form solutions, and this advantage was exploited.

V. CONVERGENCE ANALYSIS AND NEW INSIGHTS

In the last section, we asked the question of whether the EM method would converge faster than the PG method in the application of OMOD. In this section we perform convergence analysis with EM. We will show that, in theory, the EM method should converge faster than the PG method. However, this is not the most striking result. The insight gained from our analysis will lead us to develop an accelerated variant of the EM method, which converges even faster. It will also enable us to consider inexact variants of the EM methods, which provides flexibility with implementations. We will also expand our analysis to include the non-convex case; i.e., instances for which the penalty function h is non-convex.

This section is organized as follows. Section V-A gives the problem setup. Sections V-B and V-C describe the convergence results for the PG and EM methods, respectively. Section V-D reveals the insight behind our EM convergence analysis. Section V-E describes a result that indicates that the PG and EM methods can converge slowly for high SNRs. Taking insight from the EM convergence analysis, Section V-F and V-G develop the accelerated and inexact EM methods, respectively. Section V-H considers the convergence analysis for the non-convex case. It is worth noting that, while we develop these results for OMOD, they can be applied to other QLR problems.

A. Problem Setup and Preliminary Concepts

Let us write down our main problem, problem (19), again:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^n} F(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + h(\boldsymbol{\theta}), \qquad (33)$$

where $f(\boldsymbol{\theta}) = -\sum_{i=1}^m \log \Phi(\boldsymbol{b}_i^\top \boldsymbol{\theta})$ is convex differentiable; $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper, lower semi-continuous and convex. We assume that problem (33) has an optimal solution, and we will use the notation $\boldsymbol{\theta}^\star$ to denote an optimal solution to problem (33). In our study, we will see problem (33) as a problem with a general $\boldsymbol{B}$ and h. Hence, we basically consider a class of QLR problems, not just OMOD.

We also recall some basic formulas. The gradient $\nabla f(\boldsymbol{\theta})$ can be written as

$$\nabla f(\boldsymbol{\theta}) = -\boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}); \qquad (34)$$

where $\boldsymbol{B} = [\, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_m \,]^\top$; $\psi(\boldsymbol{z}) = (\psi(z_1), \ldots, \psi(z_m))$;

$$\psi(z) = \frac{e^{-z^2/2}}{\int_{-\infty}^z e^{-t^2/2}\mathrm{d}t}. \qquad (35)$$

The EM surrogate function of f can be written as

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}') = \tfrac{1}{2}\|\boldsymbol{z}' - \boldsymbol{B}\boldsymbol{\theta}\|^2 + \text{const.} \qquad (36)$$
$$\boldsymbol{z}' = \boldsymbol{B}\boldsymbol{\theta}' + \psi(\boldsymbol{B}\boldsymbol{\theta}'); \qquad (37)$$

see (5)–(6). We have $g(\boldsymbol{\theta}|\boldsymbol{\theta}') \geq f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$, and $g(\boldsymbol{\theta}'|\boldsymbol{\theta}') = f(\boldsymbol{\theta}')$ for all $\boldsymbol{\theta}'$.

Some concepts and notations should be introduced. Let $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and let $\mathcal{X} \subseteq \mathbb{R}^n$. The domain of φ is defined as $\text{dom}\,\varphi = \{\boldsymbol{\theta} \in \mathbb{R}^n | \varphi(\boldsymbol{\theta}) < +\infty\}$. A differentiable φ is said to have L_φ-Lipschitz continuous gradient on $\mathcal{X}$ if

$$\|\nabla\varphi(\boldsymbol{\theta}) - \nabla\varphi(\boldsymbol{\theta}')\| \leq L_\varphi\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \text{ for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{X}.$$

Consider a minimization problem $\min_{\boldsymbol{\theta} \in \mathbb{R}^n} \varphi(\boldsymbol{\theta})$. A point $\hat{\boldsymbol{\theta}} \in \mathbb{R}^n$ is said to be a critical point of the problem if $\boldsymbol{0} \in \partial\varphi(\hat{\boldsymbol{\theta}})$,

where $\partial\varphi(\boldsymbol{\theta})$ is the limiting subdifferential of φ at $\boldsymbol{\theta}$; see [41] and the references therein. A point $\hat{\boldsymbol{\theta}}$ is said to be an ε-critical point of the problem if $\mathrm{dist}(\mathbf{0}, \partial\varphi(\hat{\boldsymbol{\theta}})) \le \varepsilon$, where $\mathrm{dist}(\boldsymbol{y}, \mathcal{X}) = \min_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{y}-\boldsymbol{x}\|$ denotes the distance between $\boldsymbol{y}$ and $\mathcal{X}$. If φ is convex, then any critical point of the problem is an optimal solution to the problem. The notations $\sigma_{\max}(\boldsymbol{X})$ and $\sigma_{\min}(\boldsymbol{X})$ denote the largest and smallest singular values of $\boldsymbol{X}$, respectively; $\langle\cdot,\cdot\rangle$ denotes the inner product.

B. Proximal Gradient

Consider the PG method for problem (33):

$$\boldsymbol{\theta}^{k+1} = \mathrm{prox}_{\eta h}(\boldsymbol{\theta}^k - \eta\nabla f(\boldsymbol{\theta}^k)), \quad k = 0, 1, 2, \dots, \quad (38)$$

where we assume a constant step size $\eta > 0$. The following result is well known.

Fact 1 (see, e.g., [23]): *Consider problem* (33) *and the accompanied problem setup, and consider the PG method* (38). *Suppose that f has L_f-Lipschitz continuous gradient on* $\mathrm{dom}\,F$. *If the step size is chosen as $\eta = 1/L_f$, then we have*

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \le \frac{L_f}{k}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star\|^2, \quad k \ge 1.$$

Our question is whether f possesses the desired Lipschitz continuous gradient property. This appears to have not been answered before, and we show that the answer is yes.

Proposition 1: *The function f in* (33) *has L_f-Lipschitz continuous gradient on $\mathbb{R}^n$, where $L_f = \sigma_{\max}(\boldsymbol{B})^2$. Consequently, the PG method* (38) *with step size $\eta = 1/\sigma_{\max}(\boldsymbol{B})^2$ has the following convergence result for problem* (33):

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \le \frac{\sigma_{\max}(\boldsymbol{B})^2}{k}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star\|^2, \quad k \ge 1. \quad (39)$$

The proof of Proposition 1 is given in Appendix A. Proposition 1 gives several implications. First, it confirms that the PG method guarantees convergence to the optimal solution with a convergence rate of at least $1/k$. Second, it indicates how the convergence scales with the problem instance $\boldsymbol{B}$; specifically, it scales with $\sigma_{\max}(\boldsymbol{B})^2$. This will be useful in our comparison with the EM method later. Third, it suggests how we can select the step size, specifically, by $\eta = 1/\sigma_{\max}(\boldsymbol{B})^2$. In the OMOD application, we can show from (18) that

$$\sigma_{\max}(\boldsymbol{B})^2 = \frac{\sigma_{\max}(\boldsymbol{\mathcal{H}})^2}{\sigma^2} = \max_{w=1,\dots,W}\frac{\sigma_{\max}(\check{\boldsymbol{H}}_w)^2}{\sigma^2}, \quad (40)$$

which gives us a simple way to select the step size. Note that the prior OMOD work [18] selects the step size manually.

C. Expectation Maximization

Next, consider the EM method for problem (33):

$$\boldsymbol{\theta}^{k+1} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} g(\boldsymbol{\theta}|\boldsymbol{\theta}^k) + h(\boldsymbol{\theta}), \quad k = 0, 1, 2, \dots \quad (41)$$

We have the following convergence result.

Proposition 2: *Consider problem* (33) *and the accompanied problem setup, and consider the EM method* (41). *We have*

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \le \frac{1}{k}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2, \quad k \ge 1. \quad (42)$$

The proof of Proposition 2 is given in Appendix B1, and we will describe the insight behind Proposition 2 in the next subsection. Eq. (42) shows that the EM method has a convergence rate of at least $1/k$—same as PG. It is interesting to compare the convergence results of PG and EM. Since

$$\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2 \le \sigma_{\max}(\boldsymbol{B})^2\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star\|^2, \quad (43)$$

the EM convergence result (42) is seen to be at least no worse than the PG result in (39). In fact, (43) suggests that the EM method should converge faster than the PG method, particularly when the gap in (43) is large.

D. Insight Behind the Convergence Results, and Prior Work

It is interesting to see what is the insight behind the EM convergence result in Proposition 2. To see it, we need to understand how PG convergence results were established in the literature. The idea is to see PG as a majorization-minimization (MM) method. Specifically, consider the MM method

$$\boldsymbol{\theta}^{k+1} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} u(\boldsymbol{\theta}|\boldsymbol{\theta}^k) + h(\boldsymbol{\theta}), \quad k = 0, 1, 2, \dots \quad (44)$$

where $u(\cdot|\boldsymbol{\theta}')$ is a majorant of f at $\boldsymbol{\theta}'$; i.e., $u(\boldsymbol{\theta}|\boldsymbol{\theta}') \ge f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$, $u(\boldsymbol{\theta}'|\boldsymbol{\theta}') = f(\boldsymbol{\theta}')$ for all $\boldsymbol{\theta}'$. The PG method (38) with step size $\eta = 1/L_f$ is seen as an MM, with

$$u(\boldsymbol{\theta}|\boldsymbol{\theta}') = f(\boldsymbol{\theta}') + \langle\nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle + \frac{L_f}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \quad (45)$$

Many PG convergence proofs make strong use of the identity (45). As an important discovery, we found that the EM surrogate function g in (36) possesses the following structure

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}') = f(\boldsymbol{\theta}') + \langle\nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle + \frac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|^2. \quad (46)$$

We see that (46) resembles (45). This has tremendous insights. It suggests that we may use the PG convergence proof concepts in the literature to show the convergence for EM. In fact, the proof of Proposition 2 is conceptually the same as that of Fact 1.

In view of the importance of (46), we provide the derivations of (46) here. From (36)–(37), we have

$$\begin{aligned}\|\boldsymbol{z}' - \boldsymbol{B}\boldsymbol{\theta}\|^2 &= \|\psi(\boldsymbol{B}\boldsymbol{\theta}') - \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|^2 \\ &= \|\psi(\boldsymbol{B}\boldsymbol{\theta}')\|^2 - 2\langle\psi(\boldsymbol{B}\boldsymbol{\theta}'), \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}')\rangle + \|\boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|^2 \\ &= \|\psi(\boldsymbol{B}\boldsymbol{\theta}')\|^2 + 2\langle\nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle + \|\boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|^2,\end{aligned}$$

where the last equation is due to $\nabla f(\boldsymbol{\theta}') = -\boldsymbol{B}^\top\psi(\boldsymbol{B}\boldsymbol{\theta}')$ in (34). By setting $\boldsymbol{\theta} = \boldsymbol{\theta}'$ in the above equation, we get

$$\|\boldsymbol{z}' - \boldsymbol{B}\boldsymbol{\theta}'\|^2 = \|\psi(\boldsymbol{B}\boldsymbol{\theta}')\|^2.$$

Applying the above two equations to (36), and using $g(\boldsymbol{\theta}'|\boldsymbol{\theta}') = f(\boldsymbol{\theta}')$, we obtain (46). The proof is complete.

The prior works on EM convergence analyses should be mentioned and compared. The question is whether the available convergence analysis results [32], [33], [34], [35] can lead to the same or better claim than our result in (42). This aspect is complex, and we provide a detailed account of it

in the supplemental material. In summary, the best result we can find is

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \leq \frac{\sigma_{\max}(\boldsymbol{B})^2}{k+1}(2R^2), \tag{47}$$

for some constant $R > 0$; it comes from Mairal's analysis [34]. Our result in (42) appears to be better than (47); e.g., it is unconvincing to use (47) to argue that EM should converge faster than PG.

E. A Slow Convergence Phenomenon

The PG and EM theoretical bounds in (39) and (42) are seen to be loose if $\boldsymbol{B}$ has a large magnitude. From (18), the magnitude of $\boldsymbol{B}$ is seen to increase with the SNR; see also (40). This gives an impression that, under high SNRs, the PG and EM methods could converge slowly. Further analysis shows that this is true for our interested choices of h.

Proposition 3: *Consider problem* (33) *and the accompanied problem setup. Consider either the PG method* (38) *with* $\eta = 1/\sigma_{\max}(\boldsymbol{B})^2$, *or the EM method* (41). *Suppose that* $\boldsymbol{\theta}^k$ *is* ε-*optimal in the sense that*

$$\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star\| \leq \varepsilon$$

for some optimal solution $\boldsymbol{\theta}^\star$ *to problem* (33) *and for* $\varepsilon > 0$.
(a) Suppose that $h(\boldsymbol{\theta}) = \mathbb{I}_{\mathcal{X}}(\boldsymbol{\theta})$ *for some non-empty convex closed set* $\mathcal{X}$. *Then* k *must satisfy*

$$k \geq \sigma_{\max}(\boldsymbol{B})\frac{\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \varepsilon}{\sqrt{m}}$$

for the PG method; and

$$k \geq \sigma_{\min}^+(\boldsymbol{B})\frac{\|\boldsymbol{P}(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - \varepsilon}{\sqrt{m}}$$

for the EM method, where $\sigma_{\min}^+(\boldsymbol{B})$ *denotes the smallest positive singular value of* $\boldsymbol{B}$, *and* $\boldsymbol{P} = \boldsymbol{B}^\top(\boldsymbol{B}^\top\boldsymbol{B})^\dagger\boldsymbol{B}$ *is an orthogonal projection matrix, which performs projection onto the subspace spanned by* $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$.
(b) Suppose that $h(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|^2/2$, $\lambda > 0$. *Then* k *must satisfy*

$$k \geq \sigma_{\max}(\boldsymbol{B})^2\frac{\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \varepsilon}{\lambda\|\boldsymbol{\theta}^0\| + \sigma_{\max}(\boldsymbol{B})\sqrt{m}} \tag{48}$$

for the PG method; and

$$k \geq (\sigma_{\min}^+(\boldsymbol{B}))^2\frac{\|\boldsymbol{P}(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - \varepsilon}{\lambda\|\boldsymbol{P}\boldsymbol{\theta}^0\| + \sigma_{\min}^+(\boldsymbol{B})\sqrt{m}} \tag{49}$$

for the EM method.

The proof of Proposition 3 is given in Appendix C. Proposition 3 shows that the number of iterations required by the PG or EM method must increase with the magnitude of $\boldsymbol{B}$, or the SNR, assuming that the other variables are held fixed. Let us show the insight behind Proposition 3. As an example, consider $h(\boldsymbol{\theta}) = 0$ and the PG method (38) with step size $\eta = 1/\sigma_{\max}(\boldsymbol{B})^2$. From (34) one can show that

$$\|\nabla f(\boldsymbol{\theta})\| \leq \sigma_{\max}(\boldsymbol{B})\|\psi(\boldsymbol{B}\boldsymbol{\theta})\| \leq \sigma_{\max}(\boldsymbol{B})\sqrt{m}, \tag{50}$$

where the second inequality is due to $0 < \psi(z) < 1$ (see (35)). Applying (50) to the PG method (38) gives

$$\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| = \eta\|\nabla f(\boldsymbol{\theta}^k)\| \leq \sigma_{\max}(\boldsymbol{B})^{-1}\sqrt{m}. \tag{51}$$

We see that, if the magnitude of $\boldsymbol{B}$ is large, then every $\boldsymbol{\theta}^k$ changes little, and we have slow convergence. This is the idea that underlies the results in Proposition 3.

F. Accelerated EM

The PG-EM relationship revealed in Section V-D opens new possibilities. In first-order convex optimization, it is well known that the convergence of the PG method can be accelerated by applying extrapolation [20], [21]. We can accelerate the EM method in the same manner. Consider the following modified EM method

$$\boldsymbol{\theta}^{k+1} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} g(\boldsymbol{\theta}|\boldsymbol{\vartheta}^k) + h(\boldsymbol{\theta}), \tag{52}$$

$$\boldsymbol{\vartheta}^{k+1} = \boldsymbol{\theta}^{k+1} + \alpha_{k+1}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k), \tag{53}$$

for $k \geq 0$, where $\boldsymbol{\vartheta}^0 = \boldsymbol{\theta}^0$; $\{\alpha_k\}_{k\geq 1}$ is an extrapolation coefficient sequence. We employ the FISTA sequence

$$\alpha_{k+1} = \frac{t_k - 1}{t_{k+1}}, \tag{54}$$

where $t_0 = 1$,

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad k \geq 0. \tag{55}$$

The modified EM method in (52)–(55) will be called the accelerated EM method, in what follows. The accelerated EM method is the direct application of the extrapolation in accelerated PG to the standard EM method (41). Note that, if we replace $g(\boldsymbol{\theta}|\boldsymbol{\vartheta}^k)$ in (52) with $u(\boldsymbol{\theta}|\boldsymbol{\vartheta}^k)$ in (45), we return to the accelerated PG method. We have the following result.

Proposition 4: *Consider problem* (33) *and the accompanied problem setup, and consider the accelerated EM method* (52)–(55). *We have*

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \leq \frac{2}{(k+1)^2}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2, \quad k \geq 1. \tag{56}$$

The proof of Proposition 4 is given in Appendix B3. Proposition 4 indicates that accelerated EM has an accelerated convergence rate of at least $1/k^2$, which is the same as that of the accelerated PG method.

G. Inexact EM

The EM method (41) requires us to solve the problems in (41) exactly. When the exact solutions are computationally non-negligible to compute, we may want to accept economical, but inexact, solutions. This motivates the study of inexact EM

$$\boldsymbol{\theta}^{k+1} \approx \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} G(\boldsymbol{\theta}|\boldsymbol{\theta}^k) := g(\boldsymbol{\theta}|\boldsymbol{\theta}^k) + h(\boldsymbol{\theta}), \tag{57}$$

for $k \geq 0$, where $\boldsymbol{\theta}^{k+1}$ is an approximate solution, produced by a solver that may not give high solution precision due to its limited complexity. The question is how robust the EM method is under solution errors.

Intriguingly, the same kind of questions has been addressed in the context of inexact PG [22]. We take insight from the aforementioned context to answer the question at hand. We characterize the inexactness of $\boldsymbol{\theta}^{k+1}$ by

$$\mathrm{dist}(\mathbf{0}, \partial G(\boldsymbol{\theta}^{k+1}|\boldsymbol{\theta}^k)) \leq \varepsilon_{k+1}, \qquad (58)$$

for some $\varepsilon_{k+1} \geq 0$; $\varepsilon_{k+1} = 0$ means that $\boldsymbol{\theta}^{k+1}$ is an optimal solution to the problem in (57). Also, we assume the following descent condition

$$G(\boldsymbol{\theta}^{k+1}|\boldsymbol{\theta}^k) \leq G(\boldsymbol{\theta}^k|\boldsymbol{\theta}^k). \qquad (59)$$

Eq. (59) makes sense in the following way: If we do a warm start with the inexact solver for problem (57) by using $\boldsymbol{\theta}^k$ as the initialization, the resulting approximate solution $\boldsymbol{\theta}^{k+1}$ should yield an objective value $G(\boldsymbol{\theta}^{k+1}|\boldsymbol{\theta}^k)$ better than the initial objective value $G(\boldsymbol{\theta}^k|\boldsymbol{\theta}^k)$. We have the result below.

Proposition 5: *Consider problem* (33) *and the accompanied problem setup. Consider the inexact EM method* (57), *with* (58)–(59) *being satisfied. Suppose that $\boldsymbol{B}$ has full column rank. For $k \geq 1$, we have*

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \leq \frac{1}{2k}\left(\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\| + 2\frac{\sum_{i=1}^{k} \varepsilon_i}{\sigma_{\min}(\boldsymbol{B})}\right)^2.$$

The proof of Proposition 5 is given in Appendix B2. Proposition 5 reveals two aspects. First, the inexact EM method can lead to convergence, with a rate of at least $1/k$, if the solution accuracy ε_k improves with k in such a way that $\sum_{i=1}^{\infty} \varepsilon_i < \infty$; i.e., ε_k decreases at a rate of k^{-p} for $p > 1$, or faster. Second, the inexact EM method may be more sensitive to solution errors if the smallest singular value $\sigma_{\min}(\boldsymbol{B})$ of $\boldsymbol{B}$ is smaller, i.e., $\boldsymbol{B}$ is more ill-conditioned. In massive MIMO, we should expect that $\boldsymbol{B}$ is reasonably well-conditioned.

We can also consider the accelerated version of inexact EM.

Proposition 6: *Consider problem* (33) *and the accompanied problem setup. Consider the accelerated version of the inexact EM method* (57), *wherein $\boldsymbol{\theta}^k$ is replaced by $\boldsymbol{\vartheta}^k$ in* (53)–(55). *Suppose that* (58) *holds, and that $\boldsymbol{B}$ has full column rank. For $k \geq 1$, we have*

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \leq \frac{2}{(k+1)^2}\left(\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\| + 2\frac{\sum_{i=1}^{k} i\,\varepsilon_i}{\sigma_{\min}(\boldsymbol{B})}\right)^2.$$

The proof of Proposition 6 is given in Appendix B4. Note that the above result does not require the descent condition (59). Proposition 6 looks similar to its non-accelerated counterpart in Proposition 5. The notable difference is that, to achieve a convergence rate of at least $1/k^2$ in Proposition 6, we need the solution accuracies ε_k's to satisfy $\sum_{i=1}^{\infty} i\,\varepsilon_i < \infty$; i.e., ε_k decreases at a rate of k^{-p}, $p > 2$, or faster. This means that the accelerated inexact EM method may require higher solution accuracies than the non-accelerated counterpart.

H. The Case of Non-Convex h

Our preceding analyses assume convex h. We also want to handle non-convex h. Consider the same inexact EM scenario in (57). The new difficulty is that the problems in (57) are non-convex. We assume that the inexact solver for the problems

in (57) can only guarantee finding an ε-critical point, rather than an optimal or ε-optimal solution. The ε-criticality of $\boldsymbol{\theta}^{k+1}$ is characterized by (58). Also, we assume that the descent condition (59) holds; our justification is the same as before. Furthermore, we modify $G(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ in (57) as

$$G(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = g(\boldsymbol{\theta}|\boldsymbol{\theta}^k) + \frac{\tau}{2}\|\boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}^k)\|^2 + h(\boldsymbol{\theta}) \qquad (60)$$

for some $\tau > 0$. We have the following result.

Proposition 7: *Consider problem* (33) *and the accompanied problem setup, except that h is non-convex. Consider the inexact modified EM method* (57) *and* (60), *with* (58)–(59) *being satisfied. Suppose that $\boldsymbol{B}$ has full column rank. We have*

$$\min_{i=1,\ldots,k} \mathrm{dist}_{\boldsymbol{R}}(\mathbf{0}, \partial F(\boldsymbol{\theta}^i)) \leq \frac{1}{\sqrt{k}}C, \qquad (61)$$

where $\mathrm{dist}_{\boldsymbol{R}}(\boldsymbol{y}, \mathcal{X}) = \min_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{R}}$, $\|\boldsymbol{x}\|_{\boldsymbol{R}} = (\boldsymbol{x}^\top \boldsymbol{R} \boldsymbol{x})^{1/2}$, $\boldsymbol{R} = (\boldsymbol{B}^\top \boldsymbol{B})^{-1}$, *and*

$$C = 2\left(\frac{(2+\tau)^2}{\tau}(F(\boldsymbol{\theta}^0) - F(\boldsymbol{\theta}^\star)) + \frac{\sum_{i=1}^{k}\varepsilon_i^2}{2\sigma_{\min}(\boldsymbol{B})^2}\right)^{1/2}.$$

The proof of Proposition 7 is given in Appendix B5. Eq. (61) employs an ellipsoidal extension of the ε-critical point definition to pin down the convergence. As an easier way to understand, consider the following expression which can be shown to be an implication of (61):

$$\min_{i=1,\ldots,k} \mathrm{dist}(\mathbf{0}, \partial F(\boldsymbol{\theta}^i)) \leq \frac{1}{\sqrt{k}}C\,\sigma_{\max}(\boldsymbol{B}). \qquad (62)$$

Eq. (62) suggests that the EM method leads to a δ-critical point in $\mathcal{O}(1/\delta^2)$ iterations, assuming that $\sum_{i=1}^{\infty} \varepsilon_i^2 < \infty$.

VI. ACCELERATED EM ALGORITHMS FOR OMOD

In this section we use the new theoretical insights gained in the previous section to build efficient algorithms for OMOD. Our attention is paid to the implementation of the accelerated inexact EM (AIEM) method.

A. An AIEM Algorithm for a General Convex h

Consider the OMOD problem (19) with a general convex h. By modifying EM OMOD method in Section IV-D as the AIEM method in Section V-G, we have the following AIEM OMOD method: Let $\boldsymbol{s}^0$ be the initialization, and let $\boldsymbol{s}^0_{\mathsf{ex}} = \boldsymbol{s}^0$. We perform, for $k = 0, 1, 2, \ldots$,

1. E-step: For all m, compute the conditional mean estimate $\boldsymbol{r}^k_m = \mathbb{E}_{\boldsymbol{r}_m}[\boldsymbol{r}_m|\boldsymbol{q}, \boldsymbol{s}^k_{\mathsf{ex}}]$ in the same way as (31).
2. M-step preparation: For all m, compute $\tilde{\boldsymbol{r}}^k_m = \boldsymbol{F}\boldsymbol{r}^k_m$ via FFT. Let $\check{\boldsymbol{r}}^k_w = (\tilde{r}^k_{1,w}, \ldots, \tilde{r}^k_{M,w})$ for all w.
3. M-step: For all w, compute the inexact solutions

$$\check{\boldsymbol{s}}^{k+1}_w \approx \arg\min_{\boldsymbol{s} \in \mathbb{C}^N} \frac{1}{\sigma^2} \underbrace{\frac{1}{2}\|\check{\boldsymbol{r}}^k_w - \check{\boldsymbol{H}}_w \boldsymbol{s}\|^2}_{:=\varphi^k_w(\boldsymbol{s})} + h(\boldsymbol{s}). \qquad (63)$$

Here $\{\check{\boldsymbol{s}}^{k+1}_w\}_{w=1}^W$ satisfies a solution quality condition

$$\|\boldsymbol{e}^{k+1}\| \leq \varepsilon_{k+1}, \qquad (64)$$

Algorithm 1 AIEM for problem (19) with convex h

1: **input:** s^0, $\{\alpha_k\}_{k\geq 1}$ in (54)–(55), $\{\varepsilon_k\}_{k\geq 1}$
2: set $s_{\text{ex}}^0 = s^0$, $\eta_w = 1/\sigma_{\max}(\check{H}_w)^2$ for all w
3: **for** $k = 0, 1, 2, \ldots$ **do**
4:     compute $z_m^k = F^{\mathsf{H}}(\sum_{n=1}^N D_{m,n} s_{\text{ex},n}^k)$ for all m
5:     compute ζ_m^k for all m via

$$\Re(\zeta_m^k) = \sigma\left(\Re(q_m) \odot \psi\left(\tfrac{1}{\sigma}\Re(q_m)\odot\Re(z_m^k)\right)\right)$$
$$\Im(\zeta_m^k) = \sigma\left(\Im(q_m) \odot \psi\left(\tfrac{1}{\sigma}\Im(q_m)\odot\Im(z_m^k)\right)\right)$$

6:     set $r_m^k = z_m^k + \zeta_m^k$ for all m
7:     compute $\tilde{r}_m^k = F r_m^k$ for all m
8:     set $\check{r}_w^k = (\tilde{r}_{1,w}^k, \ldots, \tilde{r}_{M,w}^k)$ for all w
9:     set $s_w^0 = \check{s}_w^k$, $s_{\text{ex},w}^0 = s_w^0$ for all w, set $j = 0$
10:     **repeat**
11:         $g_w^j = \check{H}_w^{\mathsf{H}}\check{H}_w s_{\text{ex},w}^j - \check{H}_w^{\mathsf{H}}\check{r}_w^k$ for all w
12:         $s_w^{j+1} = \text{prox}_{\eta_w \sigma^2 h}(s_{\text{ex},w}^j - \eta_w g_w^j)$ for all w
13:         $s_{\text{ex},w}^{j+1} = s_w^{j+1} + \alpha_{j+1}(s_w^{j+1} - s_w^j)$ for all w
14:         $j = j + 1$
15:         $\tilde{g}_w^j = \sigma^{-2}(\check{H}_w^{\mathsf{H}}\check{H}_w s_w^j - \check{H}_w^{\mathsf{H}}\check{r}_w^k)$ for all w
16:         $e_w^j = \min_{v \in \partial h(s_w^j)} \|\tilde{g}_w^j + v\|$ for all w
17:     **until** $\sum_{w=1}^W (e_w^j)^2 \leq \varepsilon_{k+1}^2$
18:     set s^{k+1} such that $\check{s}_w^{k+1} = s_w^j$ for all w
19:     $s_{\text{ex}}^{k+1} = s^{k+1} + \alpha_{k+1}(s^{k+1} - s^k)$
20:     if a stopping rule is met, exit and **output** s^{k+1}
21: **end for**

for some given $\varepsilon_{k+1} > 0$, where $e^{k+1} \in \mathbb{R}^W$ is given by

$$e_w^{k+1} = \min_{v \in \partial h(\check{s}_w^{k+1})} \|\sigma^{-2}\nabla\varphi_w^k(\check{s}_w^{k+1}) + v\|, \text{ for all } w.$$
$$(65)$$

4. Extrapolation for acceleration:

$$s_{\text{ex}}^{k+1} = s^{k+1} + \alpha_{k+1}(s^{k+1} - s^k), \qquad (66)$$

where $\{\alpha_k\}_{k\geq 1}$ is given by (54)–(55). Note that if we solve the problems in (63) exactly, the above method becomes the accelerated (exact) EM; that if we set $\alpha_k = 0$ for all k, the above method becomes the regular non-accelerated EM; and that (64) is identical to the solution quality requirement in (58). Proposition 6 suggests that we can choose $\varepsilon_k = C k^{-p}$ for some $C > 0$ and $p > 2$.

We have not specified the inexact solver for the M-step problems in (63). The AIEM method provides us with the freedom to choose the solver, and in this study we employ the accelerated PG (APG) method. Fixing k and w, the APG method for the problem in (63) is given by

$$s_w^{j+1} = \text{prox}_{\eta_j \sigma^2 h}(s_{\text{ex},w}^j - \eta_w \nabla\varphi_w^k(s_{\text{ex},w}^j)),$$
$$s_{\text{ex},w}^{j+1} = s_w^{j+1} + \alpha_{j+1}(s_w^{j+1} - s_w^j),$$

for $j \geq 0$, where $\eta_w = 1/\sigma_{\max}(\check{H}_w)^2$ is the step size, and $\nabla\varphi_w^k(s) = \check{H}_w^{\mathsf{H}}\check{H}_w s - \check{H}_w^{\mathsf{H}}\check{r}_w^k$. We stop the APG loop when $\{s_w^{j+1}\}_{w=1}^W$ meets the solution quality requirement (64). Assembling the above components together, we obtain the AIEM algorithm in Algorithm 1.

Algorithm 2 DIEM, inspired by AIEM in Algorithm 1

1: set $s^0 = 0$, $s_{\text{ex}}^0 = s^0$
2: **for** $k = 0, 1, 2, \ldots, K - 1$ **do**
3:     compute $z_m^k = F^{\mathsf{H}}(\sum_{n=1}^N D_{m,n} s_n^k)$ for all m
4:     compute ζ_m^k for all m via

$$\Re(\zeta_m^k) = (\sigma + \beta_k)\left(\Re(q_m) \odot \psi\left(\tfrac{1}{\sigma}\Re(q_m)\odot\Re(z_m^k)\right)\right)$$
$$\Im(\zeta_m^k) = (\sigma + \beta_k)\left(\Im(q_m) \odot \psi\left(\tfrac{1}{\sigma}\Im(q_m)\odot\Im(z_m^k)\right)\right)$$

5:     set $r_m^k = z_m^k + \zeta_m^k$ for all m
6:     compute $\tilde{r}_m^k = F r_m^k$ for all m
7:     set $\check{r}_w^k = (\tilde{r}_{1,w}^k, \ldots, \tilde{r}_{M,w}^k)$ for all w
8:     $g_w^k = \check{H}_w^{\mathsf{H}}\check{H}_w s_{\text{ex},w}^k - \check{H}_w^{\mathsf{H}}\check{r}_w^k$ for all w
9:     $s_w^{k+1} = \Omega_{\gamma_k}(s_{\text{ex},w}^k - \eta_k g_w^j)$ for all w
10:     $s_{\text{ex},w}^{k+1} = s_w^{k+1} + \alpha_{k+1}(s_w^{k+1} - s_w^k)$ for all w
11:     set s^{k+1} such that $\check{s}_w^{k+1} = s_w^{k+1}$ for all w
12: **end for**

B. Algorithms for Specific h and Some Implementation Details

The AIEM algorithm in the last subsection is developed for a general convex h. Let us consider the box case $h(\theta) = \mathbb{I}_{[-U,U]^n}(\theta)$. The prox operation is given by the clipping function (25). Eq. (65), which measures the solution quality (see also Line 16 of Algorithm 1), can be shown to be

$$e_w^{k+1} = \|\chi(\tilde{g}_w^{k+1}, \check{s}_w^{k+1})\|, \qquad (67)$$

where $\chi(g, \theta) = (\chi(g_1, \theta_1), \ldots, \chi(g_n, \theta_n))$; $\chi(g, \theta) = |g|$ if $g\theta \geq 0$ or $|\theta| < U$, and $\chi(g, \theta) = 0$ otherwise; $\tilde{g}_w^{k+1} = (\Re(\sigma^{-2}\nabla\varphi_w^k(\check{s}_w^{k+1})), \Im(\sigma^{-2}\nabla\varphi_w^k(\check{s}_w^{k+1})))$; $\check{s}_w^{k+1} = (\Re(\check{s}_w^{k+1}), \Im(\check{s}_w^{k+1}))$. We relegate the proof of (67) to the supplemental material.

Let us also consider the GMAP case $h(\theta) = \lambda\|\theta\|^2/2$. The problems in (63) can actually be solved in closed form [10]:

$$\check{s}_w^{k+1} = (\check{H}_w^{\mathsf{H}}\check{H}_w + \lambda\sigma^2 I)^{-1}\check{H}_w^{\mathsf{H}}\check{r}_w^k. \qquad (68)$$

Hence we can replace the APG solver (Line 10–17 of Algorithm 1) with (68), and the result is an accelerated EM.

VII. A Deep Inexact EM Algorithm for OMOD

In this section we apply deep unfolding [24], [28] to the inexact EM algorithm to build a data-driven detector for OMOD. The idea of deep unfolding is to alter an existing iterative algorithm by data-driven learning, thereby hoping to learn a better algorithm. Specifically, we see each iteration of the predecessor algorithm as a network layer, untie some of the parameters or alter some of the structures, and learn the untied parameters or new structures from data. We tried various ways to deep-unfold Algorithm 1. The successful one, by our empirical tests, is shown in Algorithm 2; we will call the algorithm *DIEM*. The untied parameters are marked red, and some notable changes marked blue.

The most notable alternation of DIEM is Line 9 of Algorithm 2. We replace the proximal operator $\text{prox}_{\eta_w\sigma^2 h}$ (Line 12 of AIEM) with a nonlinear activation function Ω_γ.

We use a multilevel sigmoid function; its real-valued scalar version is

$$\Omega_\gamma(\theta) = \sum_{\mu \in \{0, \pm 2, \ldots, \pm 2(D-1)\}} \varrho(\gamma(\theta - \mu)), \qquad (69)$$

where $\gamma > 0$, $\varrho : \bar{\mathbb{R}} \to [-1, 1]$ is a 0-centered sigmoid function:

$$\varrho(\theta) = \frac{2}{1 + e^{-\theta}} - 1;$$

we have $\Omega_\gamma(\mathbf{s}) = [\Re(\Omega_\gamma(s_i) + \jmath\, \Im(\Omega_\gamma(s_i))]_i$ for the complex-valued vector case. The sigmoid functions were used in deep MIMO detection [25] as the activation function. The intuitive idea is that Ω_γ approaches the decision function of the constellation set as $\gamma \to \infty$, and we want to use it to encourage the symbol estimates to lie closer to the constellation points.

The other alterations of DIEM are as follows: we abandon acceleration in the EM loop; we consider one-step APG with the inexact solver; we untie the step size η_k and extrapolation coefficient α_k of the APG in a layer-varying fashion; we add a parameter β_k to partially alter the noise standard deviation.

The training of DIEM is standard. Let $\boldsymbol{p} = \{\boldsymbol{s}, \mathcal{H}, \sigma, \boldsymbol{q}\}$ be a problem instance, let $\boldsymbol{\pi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}\}$ be the set of network parameters, and let $\boldsymbol{o}_{\boldsymbol{\pi}}(\boldsymbol{p})$ be the network output. We generate a large number of problem instances $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T$ according to the signal model. Then we learn $\boldsymbol{\pi}$ from $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T$ using a deep learning software; the objective is

$$\min_{\boldsymbol{\pi}} \sum_{t=1}^{T} \|\boldsymbol{s}_t - \boldsymbol{o}_{\boldsymbol{\pi}}(\boldsymbol{p}_t)\|^2,$$

where $\boldsymbol{s}_t$ is the symbol vector of the problem instance $\boldsymbol{p}_t$.

We want to provide an interpretation for DIEM. Suppose we choose a non-convex penalty h with the OMOD (19) to better approximate the difficult constellation constraints. As a concurrent study, we recently showed that there exists a constellation-promoting penalty function h such that its proximal operator may be approximated by the sigmoid activation function (69); see [42] for details. As studied in Section V-H, the corresponding non-convex OMOD problem (19) can be handled by inexact EM. Following Section VI, we can show that the non-convex inexact EM can be implemented by nearly the same procedure as the non-accelerated version of Algorithm 1. From this view, DIEM can be regarded as the deep unfolding of the non-convex inexact EM.

VIII. NUMERICAL RESULTS

In this section numerical results are provided to illustrate the detection and runtime performance of the new algorithms. We also give additional numerical results in the supplemental material.

A. Simulation Settings

We use Monte Carlo simulations to test the performance of the algorithms. The signals are generated by the model in Section III-A. The channel is generated by a multipath model; see, e.g., [43] and the reference therein. To describe the model,

let $\bar{\boldsymbol{h}}_{l,n} = (h_{l,1,n}, \ldots, h_{l,M,n})$, $l = 0, \ldots, W' - 1$, be the lth coefficient of the time-domain channel impulse response from user k to all the BS antennas. Assuming that the BS antennas are arranged as a uniform linear array, we model each $\bar{\boldsymbol{h}}_{l,n}$ as

$$\bar{\boldsymbol{h}}_{l,n} = \sum_{j=1}^{J} \alpha_{j,l,n} \boldsymbol{a}(\theta_{j,l,n}),$$

where $\boldsymbol{a}(\theta) = (1, e^{-\jmath \frac{2\pi d}{\lambda} \sin(\theta)}, \ldots, e^{-\jmath(M-1) \frac{2\pi d}{\lambda} \sin(\theta)})$ is the response of the array to a signal coming from an angle $\theta \in (-\pi/2, \pi/2)$; d is the inter-antenna spacing; λ is the wavelength; J is the number of paths; $\alpha_{j,l,n}$ and $\theta_{j,l,n}$ are the complex channel gain and angle of a path, respectively. In our simulations, we set $W' = 16$, $J = 4$ and $d = \lambda/2$; each $\alpha_{j,l,n}$ is randomly generated, following a circular complex Gaussian distribution with mean zero and variance $1/J$; each $\theta_{j,l,n}$ is randomly generated, following a uniform distribution on $(-\pi/2, \pi/2)$. The number of trials of our simulations is 500. In addition, the SNR is defined as $\mathsf{SNR} = \mathbb{E}[\|\boldsymbol{s}\|^2]/\mathbb{E}[\|\boldsymbol{v}_n\|^2]$.

Our simulations consider the following algorithms.

1. GMAP-EM: the standard EM method for GMAP OMOD, proposed previously in [10] and reviewed in Section IV-D;
2. GMAP-AEM: the accelerated EM method for GMAP OMOD, built in Section VI;
3. BOX-PG: the PG method for box OMOD, proposed previously in [18] and reviewed in Section IV-C;
4. BOX-AIEM: the accelerated inexact EM method for box OMOD, built in Section VI;
5. BOX-EM: the standard EM method for box OMOD, implemented by BOX-AIEM with a high solution accuracy;
6. DIEM: the deep algorithm in Section VII;
7. ZF: direct application of the zero-forcing detector in unquantized MIMO-OFDM to OMOD.

We tested all the algorithms by MATLAB 8.5 on the same desktop with Intel i7-7700 processor and 16GB RAM memory.

The parameter settings with the PG and EM algorithms (the non-deep ones) are as follows. The stopping condition is either $\|\boldsymbol{s}^{k+1} - \boldsymbol{s}^k\|/\|\boldsymbol{s}^k\| \le 5 \times 10^{-4}$ or $k > 1000$. For BOX-AIEM, we set the M-step solution accuracies as $\varepsilon_k = 2NWk^{-2.1}$. For BOX-EM, we set $\varepsilon_k = 2NW \times 10^{-4}$. For BOX-PG, we choose the step size η as the reciprocal of (40); see Proposition 1. We set the noise power parameter σ^2 as

$$\sigma = \sigma_{\text{actual}} + \sigma_0, \qquad (70)$$

where σ^2_{actual} is the actual noise power; $\sigma_0 > 0$ is some given constant. Doing so is to avoid slow convergence. As shown in Proposition 3, the PG and EM methods can converge slowly for small σ^2. In our simulations, we set $\sigma_0 = 3$.

The training of DIEM is implemented on Pytorch 1.2 platform for Python 3.5.2 (the test of DIEM is on MATLAB). The training optimizer is stochastic gradient descent, with step size 10^{-3}. The SNR range in our training is the SNR range in our simulations, with an additional margin of 2 to 5dB. The number of layers is $K = 20$. The open source code is available at `https://github.com/shaomingjie/DeepEM_MIMO_det.git`.
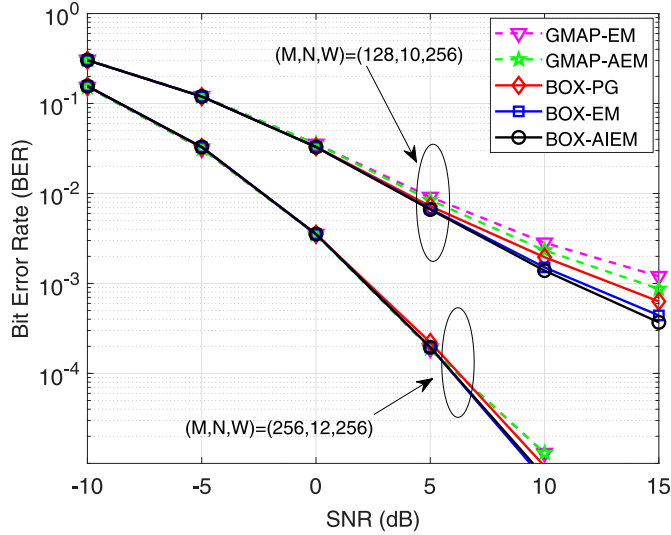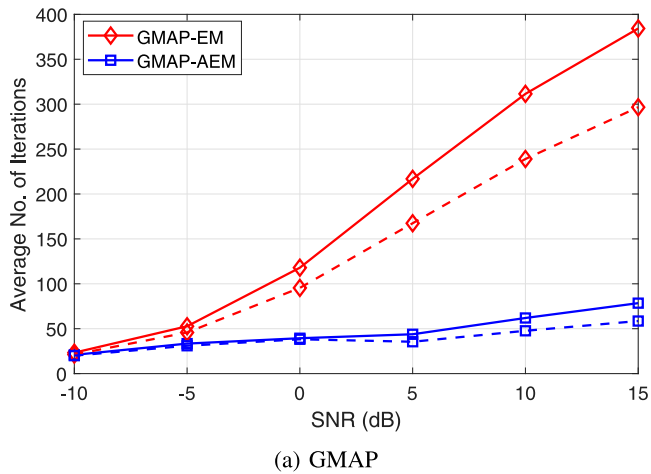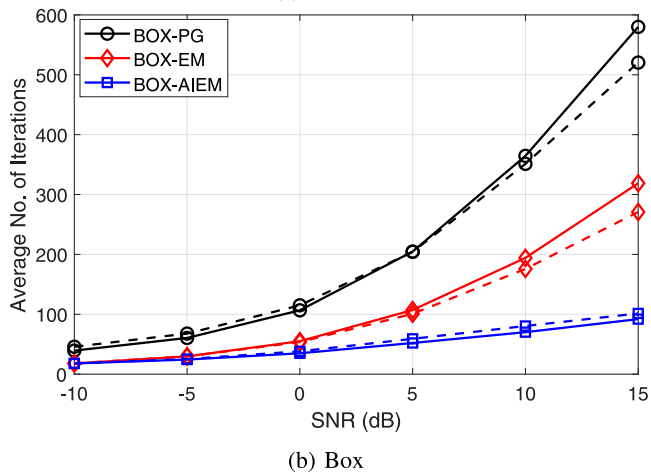
Fig. 1. BERs of the PG and EM algorithms. 16-QAM. Dashed: GMAP-based algorithms; solid: box-based algorithms.
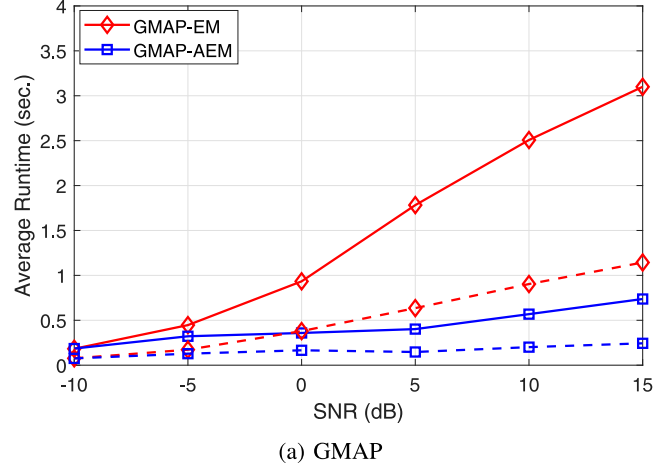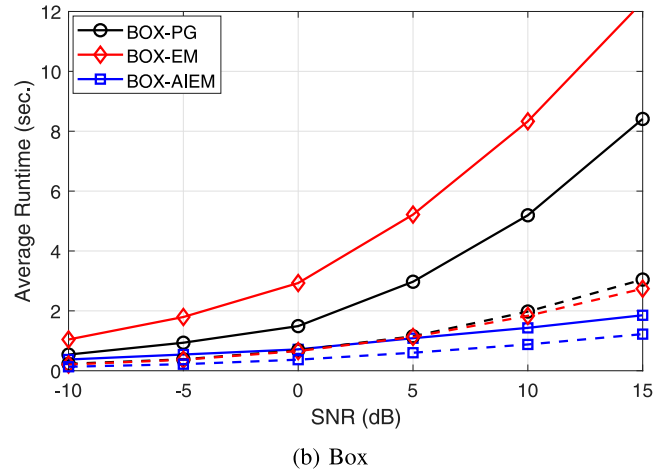

(a) GMAP

(b) Box

Fig. 2. Average number of iterations of the PG and EM algorithms. 16-QAM. Solid lines: $(M, N, W) = (256, 12, 256)$, dashed lines: $(M, N, W) = (128, 10, 256)$.


(a) GMAP

(b) Box

Fig. 3. Average runtimes of the PG and EM algorithms. 16-QAM. Solid lines: $(M, N, W) = (256, 12, 256)$, dashed lines: $(M, N, W) = (128, 10, 256)$.

## B. Comparison of EM and Accelerated Exact/Inexact EM

In this subsection we focus on comparison of the standard EM method and our modified EM methods. Fig. 1 shows the bit error rate (BER) performance of the GMAP- and box-based algorithms. GMAP-AEM is seen to yield nearly the same performance as GMAP-EM. This is expected since both solve the same problem, and the problem is convex. Similarly, we see the same behaviors with the box-based algorithms. Fig. 2 shows the average numbers of EM or PG iterations used by the algorithms. As can be seen, the accelerated EM methods use much less iterations than the PG and standard EM methods. This demonstrates the benefit of accelerated EM.

From Fig. 2, we also observe a phenomenon, namely, that the numbers of iterations of the PG and EM algorithms rise with the SNR. This agrees with the theoretical results in Proposition 3.

Fig. 3 shows the average runtimes. It is more interesting to examine the box case. While we see in Fig. 2(b) that BOX-EM uses less numbers of iterations than BOX-PG, we see in
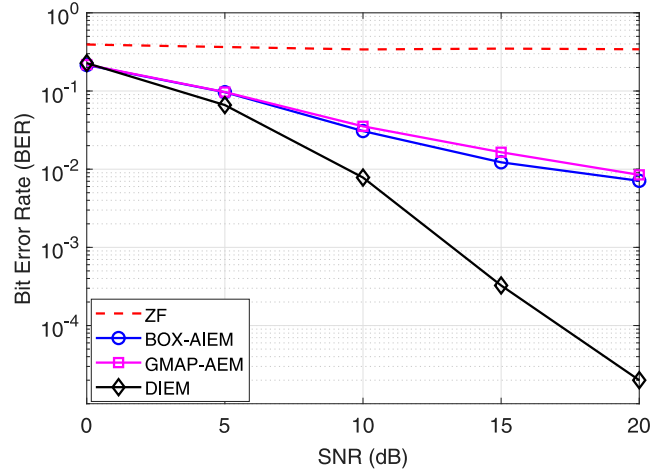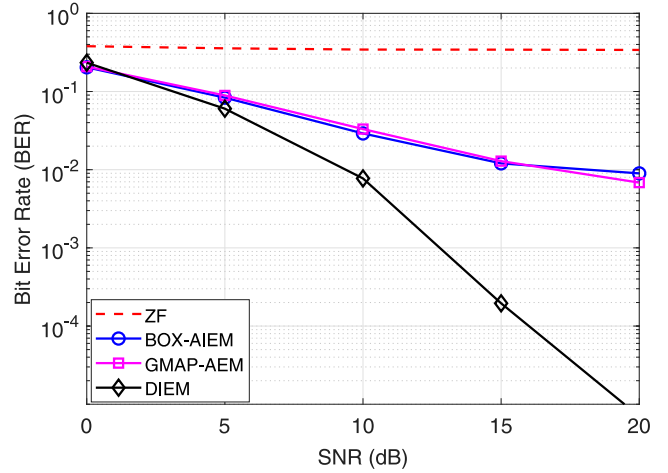
(a) $(M, N, W) = (128, 10, 256)$
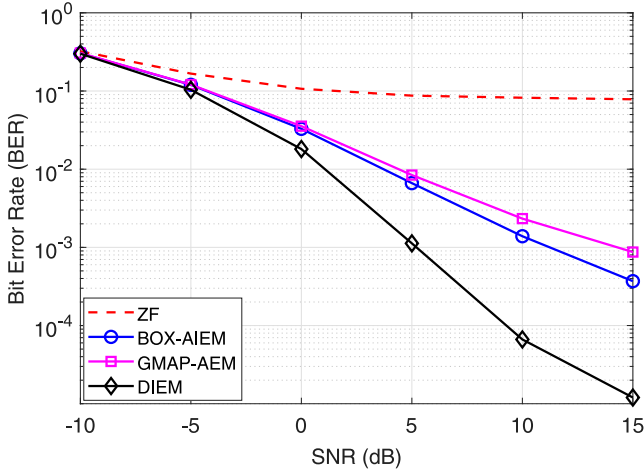
(a) $(M, N, W) = (128, 6, 256)$
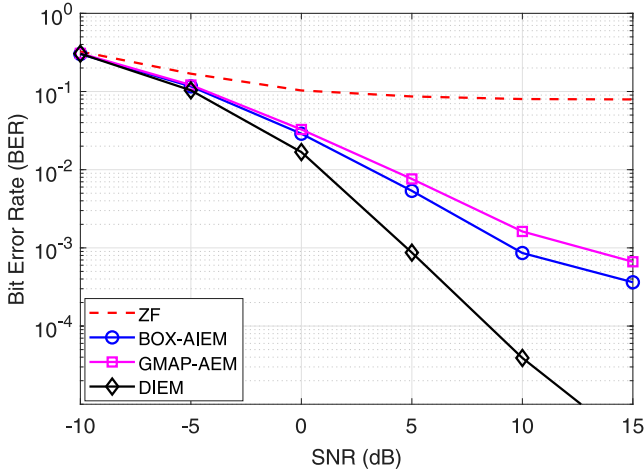
(b) $(M, N, W) = (256, 20, 512)$

(b) $(M, N, W) = (256, 12, 512)$

Fig. 4.   BER performance of DIEM; 16-QAM.

Fig. 5.   BER performance of DIEM; 64-QAM.

Fig. 3(b) that BOX-EM runs slower than BOX-PG. This is because the M-step in EM requires non-negligible computations. Moreover, BOX-AIEM is seen to be faster than BOX-EM (and also BOX-PG); the use of inexact M-step is a key factor for its runtime advantage.

C. Performance Comparison for DIEM

In this subsection we demonstrate the performance of DIEM. Figs. 4–5 show the BER performance of DIEM under various system settings. We benchmark DIEM against ZF, GMAP-AEM and BOX-AIEM. DIEM is seen to exhibit remarkably better BER performance than the other algorithms. Tables I–II show the runtimes of the various algorithms. As can be seen, DIEM is much faster than the other algorithms. The reason for the efficiency of DIEM is that it uses a fixed number of layers, specifically, 20; and each layer has simple operations, specifically, one-step APG-like operations.

TABLE I
AVERAGE RUNTIMES OF DIFFERENT ALGORITHMS
$(M, N, W) = (128, 10, 256)$, 16-QAM

SNR (dB) Alg.	-10	-5	0	5	10	15
GMAP-AEM	0.076	0.13	0.17	0.23	0.28	0.33
BOX-AIEM	0.13	0.20	0.37	0.60	0.87	1.22
DIEM	0.094					

TABLE II
AVERAGE RUNTIMES OF DIFFERENT ALGORITHMS
$(M, N, W) = (128, 6, 256)$, 64-QAM

SNR (dB) Alg.	0	5	10	15	20
GMAP-AEM	0.12	0.17	0.26	0.32	0.44
BOX-AIEM	0.59	0.92	0.95	1.60	2.06
DIEM	0.082				

IX. CONCLUSION

To conclude, the EM convergence rate was analyzed for a class of approximate ML formulations for OMOD. The analysis revealed the possibility of slow convergence under high SNRs. It also led to accelerated and inexact EM schemes. The implementation of these schemes for OMOD was studied, and the use of deep unfolding in EM for OMOD was also considered. The resulting EM algorithms were numerically shown to have competitive performance in the OMOD application. In deep learning-based methods, empirical study plays a key role. As a future direction, it would be useful to use numerical experiments to thoroughly examine the reliability and efficiency of the proposed deep unfolding EM method under different conditions.

APPENDIX

A. Proof of Proposition 1

We will need the following result.

Lemma 1: *Consider the function $\psi(z)$ in (35). The function $\psi(z)$ is 1-Lipschitz continuous on $\mathbb{R}$, i.e., $|\psi(z) - \psi(z')| \leq 1$ for all $z, z' \in \mathbb{R}$.*

Proof of Lemma 1: A differentiable function $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz continuous on $\mathbb{R}$ if $|\mathrm{d}f(x)/\mathrm{d}x| \leq L$ for all x. Moreover, $\psi(z)$ is closely related to Mill's ratio, and it was shown in the context therein that $\mathrm{d}\psi(z)/\mathrm{d}z \leq 1$ for all z [44]. The desired result follows. ∎

From the expression of $\nabla f(\boldsymbol{\theta})$ in (34), we have

$$\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\| = \|\boldsymbol{B}^\top(\psi(\boldsymbol{B\theta}) - \psi(\boldsymbol{B\theta}'))\|$$
$$\leq \sigma_{\max}(\boldsymbol{B})\|\psi(\boldsymbol{B\theta}) - \psi(\boldsymbol{B\theta}')\|$$
$$\leq \sigma_{\max}(\boldsymbol{B})\|\boldsymbol{B\theta} - \boldsymbol{B\theta}'\|$$
$$\leq \sigma_{\max}(\boldsymbol{B})^2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|,$$

where we apply Lemma 1 to obtain the second inequality. The proof is complete.

B. Proof of Propositions 2 and 4–7

The proofs of Propositions 2, 4, 5, 6 and 7 follow the same principles as their PG counterparts [22], [23], although a careful verification remains necessary. The following lemma, which considers the structure of our problem, is important.

Lemma 2: *Consider problem (33) and the accompanied problem setup. Let $\boldsymbol{\theta}^k$ be given, and consider the following problem*

$$\boldsymbol{\theta}^{k+1} \approx \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^n} G(\boldsymbol{\theta}|\boldsymbol{\theta}^k) := g(\boldsymbol{\theta}|\boldsymbol{\theta}^k) + h(\boldsymbol{\theta}), \quad (71)$$

where we consider the alternate form of $g(\boldsymbol{\theta}|\boldsymbol{\theta}')$ in (46); and $\boldsymbol{\theta}^{k+1}$ denotes an approximate solution that satisfies

$$\mathrm{dist}(\boldsymbol{0}, \partial G(\boldsymbol{\theta}^{k+1}|\boldsymbol{\theta}^k)) \leq \varepsilon_{k+1}. \quad (72)$$

Let $\tilde{\boldsymbol{\theta}}$ be any point on $\mathbb{R}^n$. It holds that

$$F(\boldsymbol{\theta}^{k+1}) - F(\tilde{\boldsymbol{\theta}}) \leq \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}})\|^2 - \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \tilde{\boldsymbol{\theta}})\|^2$$
$$+ \varepsilon_{k+1}\|\boldsymbol{\theta}^{k+1} - \tilde{\boldsymbol{\theta}}\|. \quad (73)$$

Proof of Lemma 2: First we list some basic results. Since f and h are convex, we have, respectively,

$$f(\tilde{\boldsymbol{\theta}}) \geq f(\boldsymbol{\theta}^k) + \langle \nabla f(\boldsymbol{\theta}^k), \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^k \rangle, \quad (74)$$
$$h(\tilde{\boldsymbol{\theta}}) \geq h(\boldsymbol{\theta}^{k+1}) + \langle \boldsymbol{v}^{k+1}, \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{k+1} \rangle, \quad (75)$$

for any $\boldsymbol{v}^{k+1} \in \partial h(\boldsymbol{\theta}^{k+1})$. From (72), we have

$$\boldsymbol{e}^{k+1} \in \nabla g(\boldsymbol{\theta}^{k+1}|\boldsymbol{\theta}^k) + \partial h(\boldsymbol{\theta}^{k+1}),$$

for some $\|\boldsymbol{e}^{k+1}\| \leq \varepsilon_{k+1}$. From (46), it can be shown that

$$\nabla g(\boldsymbol{\theta}|\boldsymbol{\theta}') = \nabla f(\boldsymbol{\theta}') + \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}'). \quad (76)$$

The above two equations imply that

$$\boldsymbol{e}^{k+1} = \nabla f(\boldsymbol{\theta}^k) + \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k) + \boldsymbol{v}^{k+1}, \quad (77)$$

for some $\boldsymbol{v}^{k+1} \in \partial h(\boldsymbol{\theta}^{k+1})$, $\|\boldsymbol{e}^{k+1}\| \leq \varepsilon_{k+1}$.

Second, we put together the above results to show (73). We have the derivations in (78) at the bottom of next page, where (78a) is due to $f(\boldsymbol{\theta}) \leq g(\boldsymbol{\theta}|\boldsymbol{\theta}')$; (78b) is obtained by putting (46) into (78a); (78c) is obtained by putting (74) into (78b); (78d) is obtained by putting (75) and (77) into (78c); (78e) is due to the Cauchy-Schwarz inequality and $\|\boldsymbol{e}^{k+1}\| \leq \varepsilon_{k+1}$; and (78f) is obtained by appropriate manipulation of the terms in (78e). The proof is complete. ∎

1) Proof of Proposition 2: We start with the basic EM method. Applying Lemma 2 to the EM method (41), with $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star$ and $\varepsilon_k = 0$, we have

$$F(\boldsymbol{\theta}^{k+1}) - F(\boldsymbol{\theta}^\star) \leq \tfrac{1}{2}(\|\boldsymbol{B}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star)\|^2 - \|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^\star)\|^2),$$

for $k \geq 0$. It follows that, for $k \geq 1$,

$$\sum_{i=0}^{k-1}(F(\boldsymbol{\theta}^{i+1}) - F(\boldsymbol{\theta}^\star))$$
$$\leq \tfrac{1}{2}(\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2 - \|\boldsymbol{B}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star)\|^2).$$

Also, by the basic EM property $F(\boldsymbol{\theta}^{k+1}) \leq F(\boldsymbol{\theta}^k)$, we have

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \leq \frac{1}{k}\sum_{i=0}^{k-1}(F(\boldsymbol{\theta}^{i+1}) - F(\boldsymbol{\theta}^\star))$$
$$\leq \tfrac{1}{2k}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2.$$

2) Proof of Proposition 5: Next, we consider the inexact EM scenario in (57)–(59). For convenience, let $w_k = \sum_{i=0}^{k-1}(F(\boldsymbol{\theta}^{i+1}) - F(\boldsymbol{\theta}^\star))$. Following the same proof in the preceding subsection, but with $\varepsilon_k \neq 0$ in general, one can show that, for $k \geq 1$,

$$w_k + \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star)\|^2 \leq \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2 + \sum_{i=1}^{k} \varepsilon_i \|\boldsymbol{\theta}^i - \boldsymbol{\theta}^\star\|, \quad (79)$$

$$F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \leq \tfrac{1}{k}w_k. \quad (80)$$

Note that (80) is due to the descent condition $F(\boldsymbol{\theta}^{k+1}) \leq F(\boldsymbol{\theta}^k)$, enabled by the condition $G(\boldsymbol{\theta}^{k+1}|\boldsymbol{\theta}^k) \leq G(\boldsymbol{\theta}^k|\boldsymbol{\theta}^k)$ in (59). The new challenge lies in the last term of (79). To deal with it, assume that $\boldsymbol{B}$ has full column rank. We have

$$\sum_{i=1}^{k} \varepsilon_i \|\boldsymbol{\theta}^i - \boldsymbol{\theta}^\star\| \leq \frac{1}{\sigma_{\min}(\boldsymbol{B})} \sum_{i=1}^{k} \varepsilon_i \|\boldsymbol{B}(\boldsymbol{\theta}^i - \boldsymbol{\theta}^\star)\|, \quad (81)$$

where we have used the result $\|\boldsymbol{Bx}\| \geq \sigma_{\min}(\boldsymbol{B})\|\boldsymbol{x}\|$. Since $w_k \geq 0$, Eqs. (79) and (81) imply that

$$\tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star)\|^2 \leq \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2 + \sum_{i=1}^{k} \frac{\varepsilon_i}{\sigma_{\min}} \|\boldsymbol{B}(\boldsymbol{\theta}^i - \boldsymbol{\theta}^\star)\|, \quad (82)$$

where we denote $\sigma_{\min} = \sigma_{\min}(\boldsymbol{B})$ for convenience. Consider the following lemma.

Lemma 3 [22, Lemma 1]: *Let $\{u_k\}_{k \geq 0}$ be a non-negative sequence. Suppose that, for $k \geq 1$,*

$$u_k^2 \leq u_0^2 + \sum_{i=1}^{k} \lambda_i u_i,$$

where $\lambda_i \geq 0$ for all i. Then, for $k \geq 1$, we have

$$u_k \leq \tfrac{1}{2} \sum_{i=1}^{k} \lambda_i + \left(u_0^2 + \left(\tfrac{1}{2} \sum_{i=1}^{k} \lambda_i \right)^2 \right)^{1/2}.$$

Applying Lemma 3 to (82) gives

$$\tfrac{1}{\sqrt{2}} \|\boldsymbol{B}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star)\| \leq A_k + (u_0^2 + A_k^2)^{1/2} \leq u_0 + 2A_k, \quad (83)$$

for $k \geq 1$, where

$$u_0 = \tfrac{1}{\sqrt{2}} \|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|, \quad A_k = \frac{1}{\sqrt{2}\sigma_{\min}} \sum_{i=1}^{k} \varepsilon_i. \quad (84)$$

Now we use (81) and (83) to bound the last term of (79). From (79), we have

$$w_k \leq \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2 + \sum_{i=1}^{k} \varepsilon_i \|\boldsymbol{\theta}^i - \boldsymbol{\theta}^\star\|$$

$$\leq u_0^2 + \sum_{i=1}^{k} \frac{\sqrt{2}\varepsilon_i}{\sigma_{\min}}(u_0 + 2A_i)$$

$$\leq u_0^2 + \sum_{i=1}^{k} \frac{\sqrt{2}\varepsilon_i}{\sigma_{\min}}(u_0 + 2A_k) = u_0^2 + 2A_k(u_0 + 2A_k)$$

$$= (u_0 + 2A_k)^2, \quad (85)$$

where the second inequality is obtained by applying (81) and (83); and the third inequality is due to $A_k \geq A_i$ for $i \leq k$. Applying (85) and (84) to (80) leads to the final result in Proposition 5.

3) Proof of Proposition 4: To describe the proof for the accelerated EM method, we first write down some basic results [21]. The sequence $\{t_k\}_{k \geq 0}$ in (55) satisfies (a) $t_k^2 - t_k = t_{k+1}^2$; (b) $t_k \geq (k+2)/2$; and (c) $t_k \leq k+1$. Also, the following result is important.

Lemma 4: *Let $\tilde{\boldsymbol{\theta}} = t_k^{-1}\boldsymbol{\theta}^\star + (1 - t_k^{-1})\boldsymbol{\theta}^k$. It holds that*

$$t_k^2(F(\boldsymbol{\theta}^{k+1}) - F(\tilde{\boldsymbol{\theta}})) \geq t_k^2 v_{k+1} - t_{k-1}^2 v_k, \quad (86)$$

where $v_k = F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star)$.

Note that Lemma 4 uses only the convexity of F and the property $t_k^2 - t_k = t_{k+1}^2$. Applying Lemma 2 to the accelerated EM method (52)–(55), with $\tilde{\boldsymbol{\theta}}$ given by the one in Lemma 4, $\boldsymbol{\theta}^k = \boldsymbol{\vartheta}^k$ and $\varepsilon_k = 0$, we have

$$F(\boldsymbol{\theta}^{k+1}) - F(\boldsymbol{\theta}^\star) \leq \tfrac{1}{2}(\|\boldsymbol{B}(\boldsymbol{\vartheta}^k - \tilde{\boldsymbol{\theta}})\|^2 - \|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \tilde{\boldsymbol{\theta}})\|^2), \quad (87)$$

for $k \geq 0$. Let $\boldsymbol{u}^{k+1} = t_k\boldsymbol{\theta}^{k+1} - ((t_k - 1)\boldsymbol{\theta}^k + \boldsymbol{\theta}^\star)$ for $k \geq 0$. It can be verified that, for $k \geq 1$,

$$\boldsymbol{\theta}^{k+1} - \tilde{\boldsymbol{\theta}} = t_k^{-1}\boldsymbol{u}^{k+1}, \quad \boldsymbol{\vartheta}^k - \tilde{\boldsymbol{\theta}} = t_k^{-1}\boldsymbol{u}^k. \quad (88)$$

Combining (86)–(88) gives

$$t_k^2 v_{k+1} + \tfrac{1}{2}\|\boldsymbol{Bu}^{k+1}\|^2 \leq t_{k-1}^2 v_k + \tfrac{1}{2}\|\boldsymbol{Bu}^k\|^2, \quad k \geq 1.$$

The above inequality further implies that

$$t_k^2 v_{k+1} + \tfrac{1}{2}\|\boldsymbol{Bu}^{k+1}\|^2 \leq t_0^2 v_1 + \tfrac{1}{2}\|\boldsymbol{Bu}^1\|^2. \quad (89)$$

Applying Lemma 2 to the accelerated EM method (52), with $k = 1$, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star$ and $\varepsilon_k = 0$, and noting $\boldsymbol{\vartheta}^0 = \boldsymbol{\theta}^0$, we have

$$v_1 \leq \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2 - \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^\star)\|^2.$$

By noting that $t_0 = 1$ and $\boldsymbol{u}^1 = \boldsymbol{\theta}^1 - \boldsymbol{\theta}^\star$, we further obtain

$$t_0^2 v_1 + \tfrac{1}{2}\|\boldsymbol{Bu}^1\|^2 \leq \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^\star)\|^2.$$

$$F(\boldsymbol{\theta}^{k+1}) \leq g(\boldsymbol{\theta}^{k+1}|\boldsymbol{\theta}^k) + h(\boldsymbol{\theta}^{k+1}) \tag{78a}$$

$$= f(\boldsymbol{\theta}^k) + \langle \nabla f(\boldsymbol{\theta}^k), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k \rangle + \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k)\|^2 + h(\boldsymbol{\theta}^{k+1}) \tag{78b}$$

$$\leq f(\tilde{\boldsymbol{\theta}}) + \langle \nabla f(\boldsymbol{\theta}^k), \boldsymbol{\theta}^{k+1} - \tilde{\boldsymbol{\theta}} \rangle + \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k)\|^2 + h(\boldsymbol{\theta}^{k+1}) \tag{78c}$$

$$\leq F(\tilde{\boldsymbol{\theta}}) + \langle \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k) - \boldsymbol{e}^{k+1}, \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{k+1} \rangle + \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k)\|^2 \tag{78d}$$

$$\leq F(\tilde{\boldsymbol{\theta}}) + \varepsilon_{k+1}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{k+1}\| + \langle \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k), \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{k+1} \rangle + \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k)\|^2 \tag{78e}$$

$$= F(\tilde{\boldsymbol{\theta}}) + \varepsilon_{k+1}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{k+1}\| + \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}})\|^2 - \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta}^{k+1} - \tilde{\boldsymbol{\theta}})\|^2, \tag{78f}$$

Applying the above inequality to (89) gives

$$t_k^2 v_{k+1} \le \tfrac{1}{2}\|B(\theta^0 - \theta^\star)\|^2.$$

Finally, by applying $t_k \ge (k+2)/2$ to the above inequality, we obtain the final result in Proposition 4.

4) Proof of Proposition 6: The proof for the accelerated inexact EM scenario is a combination of the proof methods in the last two subsections. Following the same proof in the previous subsection, with $\varepsilon_k \ne 0$ in general, one can show that, for $k \ge 1$,

$$t_k^2 v_{k+1} + \tfrac{1}{2}\|Bu^{k+1}\|^2 \le t_{k-1}^2 v_k + \tfrac{1}{2}\|Bu^k\|^2 \\ + \tfrac{t_k \varepsilon_{k+1}}{\sigma_{\min}}\|Bu^{k+1}\|,$$

and that

$$t_0^2 v_1 + \tfrac{1}{2}\|Bu^1\|^2 \le \tfrac{1}{2}\|B(\theta^0 - \theta^\star)\|^2 + \tfrac{t_0 \varepsilon_1}{\sigma_{\min}}\|Bu^1\|.$$

The above two inequalities imply that, for $k \ge 0$,

$$t_k^2 v_{k+1} + \tfrac{1}{2}\|Bu^{k+1}\|^2 \le u_0^2 + \sum_{i=1}^{k+1} \frac{t_{i-1}\varepsilon_i}{\sigma_{\min}}\|Bu^i\|, \quad (90)$$

where $u_0 = \|B(\theta^0 - \theta^\star)\|/\sqrt{2}$. Following the same proof in the last last subsection (specifically, by seeing (82) as (90)), one can show that

$$t_k^2 v_{k+1} \le (u_0 + 2A_{k+1})^2,$$

where $A_{k+1} = (\sum_{i=1}^{k+1} t_{i-1}\varepsilon_i)/(\sqrt{2}\sigma_{\min})$. Finally, by applying $t_k \ge (k+2)/2$ and $t_k \le k+1$ to the left-hand and right-hand sides of the above equation, respectively, we obtain the final result in Proposition 6.

5) Proof of Proposition 7: The proof for the inexact EM scenario (57)–(60) in the case of non-convex h is as follows. First, we show that

$$\min_{i=0,\dots,k-1} \tfrac{2}{\tau}\|B(\theta^{i+1} - \theta^i)\|^2 + \delta_{i+1} \le \frac{1}{k}\left(C_1 + \sum_{i=1}^k \delta_i\right) \quad (91)$$

for $k \ge 1$ and for any $\delta_1, \dots, \delta_k \ge 0$, where $C_1 = F(\theta^0) - F(\theta^\star)$. From the expression of $G(\theta|\theta')$ in (60), we have

$$F(\theta^{i+1}) + \tfrac{\tau}{2}\|B(\theta^{i+1} - \theta^i)\|^2 \le G(\theta^{i+1}|\theta^i) \le G(\theta^i|\theta^i) \\ = F(\theta^i),$$

where the first inequality is due to $f(\theta) \le g(\theta|\theta')$, and the second inequality is due to the descent condition (59). Summing the above equation over $i = 0, \dots, k-1$ yields

$$\tfrac{\tau}{2} \sum_{i=0}^{k-1} \|B(\theta^{i+1} - \theta^i)\|^2 \le C_1,$$

and further,

$$\sum_{i=0}^{k-1}(\tfrac{\tau}{2}\|B(\theta^{i+1} - \theta^i)\|^2 + \delta_{i+1}) \le C_1 + \sum_{i=0}^{k-1} \delta_{i+1}.$$

Applying $\sum_{i=0}^{k-1} u_i \ge k \min_{i=0,\dots,k-1} u_i$ to the left-hand side of the above equation leads to the result in (91).

Second, we show that

$$\mathrm{dist}_R(0, \partial F(\theta^{k+1})) \le (2+\tau)\|B(\theta^{k+1} - \theta^k)\| + \frac{\varepsilon_{k+1}}{\sigma_{\min}}, \quad (92)$$

where $\sigma_{\min} = \sigma_{\min}(B)$. From the ε-criticality condition (58), the structure of $G(\theta|\theta')$ in (60), and the expressions of $g(\theta|\theta')$ in (76), it can be shown that

$$e^{k+1} = \nabla f(\theta^k) + (1+\tau)B^\top B(\theta^{k+1} - \theta^k) + v^{k+1},$$

for some $v^{k+1} \in \partial h(\theta^{k+1})$ and $\|e^{k+1}\| \le \varepsilon_{k+1}$. This, together with $R = (B^\top B)^{-1}$, give rise to

$$\mathrm{dist}_R(0, \partial F(\theta^{k+1})) = \mathrm{dist}_R(0, \nabla f(\theta^{k+1}) + \partial h(\theta^{k+1})) \\ \le \|\nabla f(\theta^{k+1}) - \nabla f(\theta^k) \\ - (1+\tau)R^{-1}(\theta^{k+1} - \theta^k) + e^{k+1}\|_R \\ \le \|\nabla f(\theta^{k+1}) - \nabla f(\theta^k)\|_R \\ + (1+\tau)\|R^{-1}(\theta^{k+1} - \theta^k)\|_R + \tfrac{\varepsilon_{k+1}}{\sigma_{\min}},$$

where we have used the fact that $\|x\|_R \le \sigma_{\max}(R)^{1/2}\|x\|$ and $\sigma_{\max}(R)^{1/2} = 1/\sigma_{\min}(B)$. It can be verified that

$$\|R^{-1}(\theta^{k+1} - \theta^k)\|_R = \|B(\theta^{k+1} - \theta^k)\|.$$

Let $u = \psi(B\theta^k) - \psi(B\theta^{k+1})$ for convenience. We have

$$\|\nabla f(\theta^{k+1}) - \nabla f(\theta^k)\|_R = u^\top B(B^\top B)^{-1}B^\top u \\ \le \|u\|^2 = \|\psi(B\theta^k) - \psi(B\theta^{k+1})\|^2 \\ \le \|B\theta^k - B\theta^{k+1}\|^2,$$

where the first inequality is due to the fact that $B(B^\top B)^{-1}B^\top$ is an orthogonal projection matrix, and the second inequality is due to Lemma 1. Putting the above components together, we have (92).

Third, we use (91) and (92) to show the desired result. From (92), we have

$$\mathrm{dist}_R(0, \partial F(\theta^{i+1}))^2 \le 2\left((2+\tau)^2\|B(\theta^{i+1} - \theta^i)\|^2 + \tfrac{\varepsilon_{i+1}^2}{\sigma_{\min}^2}\right),$$

which is due to $(x+y)^2 \le 2x^2 + 2y^2$. By appropriately setting the δ_i's in (91), we have

$$\min_{i=0,\dots,k-1} \tfrac{\tau}{2}\|B(\theta^{i+1} - \theta^i)\|^2 + \tfrac{\frac{\tau}{2}\varepsilon_{k+1}^2}{(2+\tau)^2\sigma_{\min}^2} \le \frac{1}{k}\left(C_1 + \tfrac{\frac{\tau}{2}C_2}{(2+\tau)^2}\right),$$

where $C_2 = \sum_{i=1}^{k} \varepsilon_i^2 / \sigma_{\min}^2$. Combining the above two equations gives

$$\min_{i=0,\dots,k-1} \operatorname{dist}_R(\mathbf{0}, \partial F(\boldsymbol{\theta}^{i+1}))^2 \le \frac{4}{k\tau}((2+\tau)^2 C_1 + \tfrac{\tau}{2}C_2),$$

which leads to the final result of Proposition 7.

C. Proof of Proposition 3

1) The Case With PG, $h(\boldsymbol{\theta}) = \mathbb{I}_{\mathcal{X}}(\boldsymbol{\theta})$: Suppose that $\varepsilon \ge \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^k\|$. By the triangle inequality,

$$\varepsilon \ge \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^k\| \ge \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^k\|, \tag{93a}$$

$$\ge \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \sum_{i=0}^{k-1} \|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1}\|. \tag{93b}$$

We focus on $\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1}\|$. Consider the following result.

Lemma 5 [23, Theorem 6.42]: *For any convex $h: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, we have $\|\operatorname{prox}_h(\boldsymbol{z}_1) - \operatorname{prox}_h(\boldsymbol{z}_2)\| \le \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$ for any $\boldsymbol{z}_1, \boldsymbol{z}_2$.*

Also, for $h(\boldsymbol{\theta}) = \mathbb{I}_{\mathcal{X}}(\boldsymbol{\theta})$, $\mathcal{X}$ being non-empty closed convex, we have $\operatorname{prox}_{\eta h}(\boldsymbol{z}) = \arg\min_{\boldsymbol{\theta} \in \mathcal{X}} \|\boldsymbol{z} - \boldsymbol{\theta}\|^2$. From the above equation, it can be verified that the PG iterates in (38) satisfy

$$\boldsymbol{\theta}^i = \operatorname{prox}_{\eta h}(\boldsymbol{\theta}^i).$$

Applying the above results to the PG method (38), we have

$$\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1}\| = \|\operatorname{prox}_{\eta h}(\boldsymbol{\theta}^i) - \operatorname{prox}_{\eta h}(\boldsymbol{\theta}^i - \eta\nabla f(\boldsymbol{\theta}^i))\|$$
$$\le \|\eta\nabla f(\boldsymbol{\theta}^i)\|. \tag{94}$$

Also, from the expression of $\nabla f(\boldsymbol{\theta})$ in (34), we have

$$\|\nabla f(\boldsymbol{\theta})\| = \|\boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}))\| \le \sigma_1 \|\psi(\boldsymbol{B}\boldsymbol{\theta})\| \le \sigma_1 \sqrt{m}, \quad (95)$$

where we denote $\sigma_1 = \sigma_{\max}(\boldsymbol{B})$ for convenience, and the second inequality is due to $0 < \psi(z) < 1$ for all z (see (35)). By putting (95) into (94), and then (94) into (93b), and by setting $\eta = 1/\sigma_1^2$, we obtain $\varepsilon \ge \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \sigma_1^{-1}\sqrt{m}k$, and consequently the desired result

$$k \ge \sigma_1(\sqrt{m})^{-1}(\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \varepsilon).$$

2) The Case With PG, $h(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|^2/2$: We begin with $\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^0\|$. For $h(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|^2/2$, $\lambda > 0$, it can be verified that $\operatorname{prox}_{\eta h}(\boldsymbol{z}) = \alpha^{-1}\boldsymbol{z}$, where $\alpha = 1 + \eta\lambda$. Applying this result to the PG method (38) leads to

$$\boldsymbol{\theta}^k = \alpha^{-1}\boldsymbol{\theta}^{k-1} - \alpha^{-1}\eta\nabla f(\boldsymbol{\theta}^{k-1})$$
$$= \alpha^{-k}\boldsymbol{\theta}^0 - \eta\sum_{i=0}^{k-1} \alpha^{-(k-i)}\nabla f(\boldsymbol{\theta}^i),$$

and subsequently

$$\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^0\| \le (1-\alpha^{-k})\|\boldsymbol{\theta}^0\| + \eta\left(\sum_{i=0}^{k-1} \alpha^{-(k-i)}\right)\|\nabla f(\boldsymbol{\theta}^i)\|$$
$$\le (1-\alpha^{-k})\|\boldsymbol{\theta}^0\| + \eta\sigma_1\sqrt{m}\left(\sum_{i=0}^{k-1} \alpha^{-(k-i)}\right),$$

$$\le (1-\alpha^{-k})\|\boldsymbol{\theta}^0\| + \eta\sigma_1\sqrt{m}\frac{\alpha^{-1} - \alpha^{-k-1}}{1-\alpha^{-1}},$$
$$= (1-\alpha^{-k})(\|\boldsymbol{\theta}^0\| + \sigma_1\sqrt{m}\lambda^{-1}), \tag{96}$$

where the second inequality is due to (95), and the last equation is obtained by putting $\alpha = 1 + \eta\lambda$ into the equation.

Next, we apply (96) to (93a) to obtain

$$1 - \alpha^{-k} \ge \frac{\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \varepsilon}{\|\boldsymbol{\theta}^0\| + \sigma_1\sqrt{m}\lambda^{-1}} := C. \tag{97}$$

Assume $C < 1$. Then (97) can be rewritten as

$$\log(1 - C) \ge \log(\alpha^{-k}) = -k \cdot \log(\alpha). \tag{98}$$

Applying $\log(z) \le z - 1$ for $z > 0$ to both sides of (98) gives

$$-C \ge -k(\alpha - 1) = -k\eta\lambda \tag{99}$$

By setting $\eta = 1/\sigma_1^2$ and putting the expression of C in (97) into (99), we obtain the desired result

$$k \ge \frac{C}{\eta\lambda} = \sigma_1^2 \frac{\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| - \varepsilon}{\lambda\|\boldsymbol{\theta}^0\| + \sigma_1\sqrt{m}}.$$

What remains to show is that $C < 1$ is true for any $\lambda > 0$. Since $\nabla F(\boldsymbol{\theta}^\star) = \mathbf{0}$, and $\nabla F(\boldsymbol{\theta}) = \nabla f(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}$, we have

$$\|\boldsymbol{\theta}^\star\| = \lambda^{-1}\|\nabla f(\boldsymbol{\theta}^\star)\| \le \lambda^{-1}\sigma_1\sqrt{m};$$

the inequality is due to (95). Consequently,

$$\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0\| \le \|\boldsymbol{\theta}^0\| + \sigma_1\sqrt{m}\lambda^{-1}, \tag{100}$$

and it is seen that the C in (97) satisfies $C < 1$ for any $\varepsilon > 0$. The proof is complete.

3) The Case With EM, $h(\boldsymbol{\theta}) = \mathbb{I}_{\mathcal{X}}(\boldsymbol{\theta})$: To facilitate the proof, let

$$\boldsymbol{B} = \boldsymbol{U}_1 \tilde{\boldsymbol{\Sigma}} \boldsymbol{V}_1^\top \tag{101}$$

be the thin singular value decomposition of $\boldsymbol{B}$, where $\boldsymbol{U}_1 \in \mathbb{R}^{m \times r}$ and $\boldsymbol{V}_1 \in \mathbb{R}^{n \times r}$ are semi-orthogonal; $r = \operatorname{rank}(\boldsymbol{B})$; $\tilde{\boldsymbol{\Sigma}} = \operatorname{Diag}(\sigma_1, \dots, \sigma_r)$; $\sigma_1 \ge \cdots \ge \sigma_r > 0$. First we show that the EM method (41) can be expressed as

$$\boldsymbol{\theta}^{k+1} \in \operatorname{Prox}_{h, \boldsymbol{B}}(\boldsymbol{\theta}^k + \boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}^k)), \tag{102}$$

where $\boldsymbol{W} = \boldsymbol{B}^\top \boldsymbol{B}$, and

$$\operatorname{Prox}_{h, \boldsymbol{B}}(\boldsymbol{z}) = \arg\min_{\boldsymbol{\theta}} \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{z} - \boldsymbol{\theta})\|^2 + h(\boldsymbol{\theta}). \tag{103}$$

To prove (102), we put $\nabla f(\boldsymbol{\theta})$ in (34) into $g(\boldsymbol{\theta}|\boldsymbol{\theta}')$ in (46):

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}') = \tfrac{1}{2}\|\boldsymbol{B}\boldsymbol{\theta}\|^2 - \langle \boldsymbol{W}\boldsymbol{\theta}' + \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}'), \boldsymbol{\theta}\rangle + \text{const.}$$
$$= \tfrac{1}{2}\|\boldsymbol{B}\boldsymbol{\theta}\|^2 - \langle \boldsymbol{W}\boldsymbol{\theta}' + \boldsymbol{W}\boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}'), \boldsymbol{\theta}\rangle + \text{const.}$$
$$= \tfrac{1}{2}\|\boldsymbol{B}(\boldsymbol{\theta} - (\boldsymbol{\theta}' + \boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}')))\|^2 + \text{const.},$$

where the second inequality is due to $\boldsymbol{B}^\top = \boldsymbol{W}\boldsymbol{W}^\dagger \boldsymbol{B}^\top$, which can be verified to be true from (101). The proof of (102) is done.

Second, in a similar way as (93), we have

$$
\begin{aligned}
\varepsilon \geq \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^k\| &\geq \|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^k)\| \\
&\geq \|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - \|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^k)\| \quad (104a) \\
&\geq \|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - \sum_{i=0}^{k-1}\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1})\|, \quad (104b)
\end{aligned}
$$

where we have used the fact that $\|\boldsymbol{x}\| \geq \|\boldsymbol{V}_1^\top \boldsymbol{x}\|$ for any semi-orthogonal $\boldsymbol{V}_1$. For $h(\boldsymbol{\theta}) = \mathbb{I}_{\mathcal{X}}(\boldsymbol{\theta})$, $\mathcal{X}$ being non-empty closed convex, it can be verified that $\mathrm{Prox}_{h,\boldsymbol{B}}(\boldsymbol{z}) = \arg\min_{\boldsymbol{\theta} \in \mathcal{X}} \|\boldsymbol{B}(\boldsymbol{z} - \boldsymbol{\theta})\|^2$, and then, that $\boldsymbol{\theta}^i \in \mathrm{Prox}_{h,\boldsymbol{B}}(\boldsymbol{\theta}^i)$. Consider an extension of Lemma 5.

Lemma 6: *Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex, let $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^n$, and let $\boldsymbol{\theta}_1 \in \mathrm{Prox}_{h,\boldsymbol{B}}(\boldsymbol{z}_1), \boldsymbol{\theta}_2 \in \mathrm{Prox}_{h,\boldsymbol{B}}(\boldsymbol{z}_2)$. We have $\|\boldsymbol{B}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\| \leq \|\boldsymbol{B}(\boldsymbol{z}_1 - \boldsymbol{z}_2)\|$.*

Proof of Lemma 6: The proof is similar to that of Lemma 5; see, e.g., [23]. By the optimality condition of the problem in (103), $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ must satisfy

$$
\boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{z}_1 - \boldsymbol{\theta}_1) \in \partial h(\boldsymbol{\theta}_1), \ \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{z}_2 - \boldsymbol{\theta}_2) \in \partial h(\boldsymbol{\theta}_2). \quad (105)
$$

Since h is convex, we have the following property

$$
\langle \boldsymbol{g}_1 - \boldsymbol{g}_2, \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \geq 0, \text{ for any } \boldsymbol{g}_1 \in \partial h(\boldsymbol{\theta}_1), \boldsymbol{g}_2 \in \partial h(\boldsymbol{\theta}_2);
$$

see, e.g., [38]. Applying (105) to the above equation, we obtain

$$
\begin{aligned}
&\langle \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{z}_1 - \boldsymbol{\theta}_1) - \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{z}_2 - \boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \geq 0 \\
&\iff \langle \boldsymbol{B}^\top \boldsymbol{B}(\boldsymbol{z}_1 - \boldsymbol{z}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \geq \|\boldsymbol{B}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^2 \\
&\implies \|\boldsymbol{B}(\boldsymbol{z}_1 - \boldsymbol{z}_2)\|\|\boldsymbol{B}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\| \geq \|\boldsymbol{B}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^2 \\
&\iff \|\boldsymbol{B}(\boldsymbol{z}_1 - \boldsymbol{z}_2)\| \geq \|\boldsymbol{B}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|,
\end{aligned}
$$

where the Cauchy-Schwarz inequality is used to obtain the third equation. The proof is complete. ∎

Applying the above results to the equivalent form of EM in (102), we obtain

$$
\|\boldsymbol{B}(\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1})\| \leq \|\boldsymbol{B}(\boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}^i))\| \leq \sqrt{m}, \quad (106)
$$

where we have also used $\sigma_{\max}(\boldsymbol{B}\boldsymbol{W}^\dagger \boldsymbol{B}^\top) = 1$, which can be verified from (101), and $\|\psi(\boldsymbol{B}\boldsymbol{\theta})\| \leq \sqrt{m}$ (see (95)). Applying (106) to (104b), and using $\|\boldsymbol{B}\boldsymbol{x}\| \geq \sigma_r \|\boldsymbol{V}_1^\top \boldsymbol{x}\|$, we obtain $\varepsilon \geq \|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - k\sigma_r^{-1}\sqrt{m}$, and consequently

$$
k \geq \sigma_r(\sqrt{m})^{-1}\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\|. \quad (107)
$$

Finally, it should be noted that σ_r is the smallest positive singular value of $\boldsymbol{B}$; and that $\|\boldsymbol{V}_1^\top \boldsymbol{x}\| = \|\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^\dagger \boldsymbol{B}\boldsymbol{x}\|$, which can be verified from (101). Applying the above results to (107) leads to the desired result.

4) The Case With EM, $h(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|^2/2$: Our analysis starts with $\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^0 - \boldsymbol{\theta}^k)\|$. For $h(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|^2/2$, $\lambda > 0$, it can be shown from (102)–(103) that

$$
\begin{aligned}
\boldsymbol{\theta}^k &= \boldsymbol{\Phi}(\boldsymbol{\theta}^{k-1} + \boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}^{k-1})) \\
&= \boldsymbol{\Phi}^k \boldsymbol{\theta}^0 + \sum_{i=0}^{k-1} \boldsymbol{\Phi}^{k-i}(\boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}^i)),
\end{aligned}
$$

where $\boldsymbol{\Phi} = (\boldsymbol{W} + \lambda\boldsymbol{I})^{-1}\boldsymbol{W}$. From (101), one can show that

$$
\boldsymbol{\Phi} = \boldsymbol{V}_1 \tilde{\boldsymbol{D}} \boldsymbol{V}_1^\top,
$$

where $\tilde{\boldsymbol{D}} = \mathrm{Diag}(\alpha_1^{-1}, \ldots, \alpha_r^{-1})$, $\alpha_i = 1 + \lambda\sigma_i^{-2}$; and then

$$
\boldsymbol{V}_1^\top \boldsymbol{\Phi}^i = \boldsymbol{V}_1^\top(\boldsymbol{V}_1 \tilde{\boldsymbol{D}}^i \boldsymbol{V}_1^\top) = \tilde{\boldsymbol{D}}^i \boldsymbol{V}_1^\top
$$

The above results lead to

$$
\begin{aligned}
\boldsymbol{V}_1^\top(\boldsymbol{\theta}^k - \boldsymbol{\theta}^0) &= (\tilde{\boldsymbol{D}}^k - \boldsymbol{I})\boldsymbol{V}_1^\top \boldsymbol{\theta}^0 \\
&\quad + \sum_{i=0}^{k-1} \tilde{\boldsymbol{D}}^{k-i} \boldsymbol{V}_1^\top \boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}^i).
\end{aligned}
$$

Subsequently we have

$$
\begin{aligned}
&\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^k - \boldsymbol{\theta}^0)\| \\
&\leq (1 - \alpha_r^{-k})\|\boldsymbol{V}_1^\top \boldsymbol{\theta}^0\| \\
&\quad + \sum_{i=0}^{k-1} \alpha_1^{-(k-i)}\|\boldsymbol{V}_1^\top \boldsymbol{W}^\dagger \boldsymbol{B}^\top \psi(\boldsymbol{B}\boldsymbol{\theta}^i)\| \\
&\leq (1 - \alpha_r^{-k})\|\boldsymbol{V}_1^\top \boldsymbol{\theta}^0\| + \sum_{i=0}^{k-1} \alpha_1^{-(k-i)}\sigma_r^{-1}\sqrt{m} \\
&= (1 - \alpha_r^{-k})\underbrace{\|\boldsymbol{V}_1^\top \boldsymbol{\theta}^0\|}_{:=a_1} + (1 - \alpha_1^{-k})\underbrace{\sigma_1^2 \lambda^{-1}\sigma_r^{-1}\sqrt{m}}_{:=a_2}, \quad (108)
\end{aligned}
$$

where the second inequality is due to $\boldsymbol{V}_1^\top \boldsymbol{W}^\dagger \boldsymbol{B}^\top = \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}_1^\top$, which can be verified from (101), and $\|\psi(\boldsymbol{B}\boldsymbol{\theta})\| \leq \sqrt{m}$.

Next, by applying (108) to (104a), we obtain

$$
\frac{(1 - \alpha_r^{-k})a_1 + (1 - \alpha_1^{-k})a_2}{a_1 + a_2} \geq \frac{\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - \varepsilon}{a_1 + a_2} := C. \quad (109)
$$

Assume $C < 1$. Then, from (109),

$$
\begin{aligned}
\log(1 - C) &\geq \log\left(\frac{\alpha_r^{-k}a_1 + \alpha_1^{-k}a_2}{a_1 + a_2}\right) \\
&\geq \frac{a_1 \log(\alpha_r^{-k}) + a_2 \log(\alpha_1^{-k})}{a_1 + a_2}, \quad (110)
\end{aligned}
$$

where we use Jensen's inequality to obtain the second inequality. Applying $\log(z) \leq z - 1$ to both sides of (110) gives

$$
\begin{aligned}
&-C \geq -k\frac{a_1(\alpha_r - 1) + a_2(\alpha_1 - 1)}{a_1 + a_2} = -k\frac{a_1 \lambda\sigma_r^{-2} + a_2 \lambda\sigma_1^{-2}}{a_1 + a_2} \\
&\implies -(\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - \varepsilon) \geq -k(\lambda\sigma_r^{-2}\|\boldsymbol{V}_1^\top \boldsymbol{\theta}^0\| + \sigma_r^{-1}\sqrt{m}) \\
&\implies k \geq \sigma_r^2 \frac{\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| - \varepsilon}{\lambda\|\boldsymbol{V}_1^\top \boldsymbol{\theta}^0\| + \sigma_r\sqrt{m}},
\end{aligned}
$$

which is the desired result. We also need to show $C < 1$. As a variant of (100), we have

$$\|\boldsymbol{V}_1^\top(\boldsymbol{\theta}^\star - \boldsymbol{\theta}^0)\| \leq \|\boldsymbol{V}_1^\top\boldsymbol{\theta}^0\| + \|\boldsymbol{\theta}^\star\| \leq \|\boldsymbol{V}_1^\top\boldsymbol{\theta}^0\| + \sigma_1\sqrt{m}\lambda^{-1}$$
$$\leq \|\boldsymbol{V}_1^\top\boldsymbol{\theta}^0\| + \sigma_1(\sigma_1/\sigma_r)\sqrt{m}\lambda^{-1} = a_1 + a_2.$$

It follows that the C in (109) satisfies $C < 1$ for $\varepsilon > 0$. The proof is complete.

REFERENCES

[1] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[2] H. Jedda, A. Mezghani, A. L. Swindlehurst, and J. A. Nossek, "Quantized constant envelope precoding with PSK and QAM signaling," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8022–8034, Dec. 2018.

[3] F. Sohrabi, Y.-F. Liu, and W. Yu, "One-bit precoding and constellation range design for massive MIMO with QAM signaling," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 557–570, Jun. 2018.

[4] M. Shao, Q. Li, W.-K. Ma, and A. M.-C. So, "A framework for one-bit and constant-envelope precoding over multiuser massive MISO channels," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5309–5324, Oct. 2019.

[5] M. Shao, W.-K. Ma, Q. Li, and A. L. Swindlehurst, "One-bit sigma-delta MIMO precoding," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1046–1061, Sep. 2019.

[6] S. Rao, G. Seco-Granados, H. Pirzadeh, J. A. Nossek, and A. L. Swindlehurst, "Massive MIMO channel estimation with low-resolution spatial sigma-delta ADCs," *IEEE Access*, vol. 9, pp. 109320–109334, 2021.

[7] A. Mezghani, M.-S. Khoufi, and J. A. Nossek, "A modified MMSE receiver for quantized MIMO systems," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, 2007.

[8] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," 2014, *arXiv:1404.7736*.

[9] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[10] D. Plabst, J. Munir, A. Mezghani, and J. A. Nossek, "Efficient non-linear equalization for 1-bit quantized cyclic prefix-free massive MIMO systems," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 1–6.

[11] J. Choi, J. Mo, and R. W. Heath, "Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2005–2018, May 2016.

[12] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[13] M. Shao and W.-K. Ma, "Binary MIMO detection via homotopy optimization and its deep adaptation," *IEEE Trans. Signal Process.*, vol. 69, pp. 781–796, 2020.

[14] L. V. Nguyen, A. L. Swindlehurst, and D. H. Nguyen, "Linear and deep neural network-based receivers for massive MIMO systems with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7333–7345, Nov. 2021.

[15] S. Khobahi, N. Shlezinger, M. Soltanalian, and Y. C. Eldar, "LoRD-Net: Unfolded deep detection network with low-resolution receivers," *IEEE Trans. Signal Process.*, vol. 69, pp. 5651–5664, 2021.

[16] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

[17] S. S. Thoota and C. R. Murthy, "Massive MIMO-OFDM systems with low resolution ADCs: Cramér-Rao bound, sparse channel estimation, and soft symbol decoding," *IEEE Trans. Signal Process.*, vol. 70, pp. 4835–4850, 2022.

[18] S. H. Mirfarshbafan, M. Shabany, S. A. Nezamalhosseini, and C. Studer, "Algorithm and VLSI design for 1-bit data detection in massive MIMO-OFDM," *IEEE Open J. Circuits Syst.*, vol. 1, pp. 170–184, 2020.

[19] J. García, J. Munir, K. Roth, and J. A. Nossek, "Channel estimation and data equalization in frequency-selective MIMO systems with one-bit quantization," 2021, *arXiv:1609.04536*.

[20] Y. Nesterov, "Gradient methods for minimizing composite functions," CORE Discussion Paper, 2007/76, Université catholique de Louvain, Belgium, 2007.

[21] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[22] M. Schmidt, N. Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 24, 2011.

[23] A. Beck, *First-Order Methods in Optimization*, vol. 25. Philadelphia, PA, USA: SIAM, 2017.

[24] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, May 2019.

[25] V. Corlay, J. J. Boutros, P. Ciblat, and L. Brunel, "Multilevel MIMO detection with deep learning," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, 2018, pp. 1805–1809.

[26] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Trans. Signal Process.*, vol. 68, pp. 1702–1715, 2020.

[27] N. Shlezinger, R. Fu, and Y. C. Eldar, "DeepSIC: Deep soft interference cancellation for multiuser MIMO detection," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1349–1362, Feb. 2021.

[28] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 399–406.

[29] M. T. Ivrlac and J. A. Nossek, "On MIMO channel estimation with single-bit signal-quantization," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, 2007.

[30] A. Mezghani, F. Antreich, and J. A. Nossek, "Multiple parameter estimation with quantized channel output," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, Piscataway, NJ, USA: IEEE Press, 2010, pp. 143–150.

[31] A. Zymnis, S. Boyd, and E. Candes, "Compressed sensing with quantized measurements," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 149–152, Feb. 2010.

[32] C. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[33] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[34] J. Mairal, "Optimization with first-order surrogate functions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, 2013, pp. 783–791.

[35] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo, "Iteration complexity analysis of block coordinate descent methods," *Math. Program.*, vol. 163, no. 1, pp. 85–114, 2017.

[36] M. Shao and W.-K. Ma, "Divide and conquer: One-bit MIMO-OFDM detection by inexact expectation maximization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 4890–4894.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[38] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[39] V. Hasselblad, A. G. Stead, and W. Galke, "Analysis of coarsely grouped data from the lognormal distribution," *J. Amer. Stat. Assoc.*, vol. 75, no. 372, pp. 771–778, 1980.

[40] A. Goldsmith, *Wireless Communications*. New York, NY, USA: Cambridge Univ. Press, 2005.

[41] J. Li, A. M.-C. So, and W.-K. Ma, "Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions," *IEEE Signal Process. Mag.*, vol. 37, no. 5, pp. 18–31, Sep. 2020.

[42] M. Shao, W.-K. Ma, and J. Liu, "An explanation of deep MIMO detection from a perspective of homotopy optimization," *IEEE Open J. Signal Process.*, vol. 4, pp. 108–116, 2023.

[43] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[44] M. R. Sampford, "Some inequalities on Mill's ratio and related functions," *Ann. Math. Statist.*, vol. 24, no. 1, pp. 130–132, 1953.

Mingjie Shao (Member, IEEE) received the B.S. degree from Xidian University, Xi'an, China, in 2015, and the Ph.D. degree from the Chinese University of Hong Kong (CUHK), in 2020. He was a Postdoctoral Fellow with the Department of Electronic Engineering, CUHK from 2020 to 2023. He is currently a Research Professor ("Qilu Young Scholar") with the School of Information Science and Engineering, Shandong University, Qingdao, China. He was the Recipient of the Hong Kong Ph.D. Fellowship Scheme (HKPFS) from August 2015. He was listed in the Student Best Paper Finalists in ICASSP 2017. His research interests are mainly on convex and non-convex optimization, statistical signal processing, and machine learning for wireless communication.

Wing-Kin Ma (Fellow, IEEE) is currently a Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. His research interests include signal processing, optimization, and communications, with recent focus on i) optimization and statistical aspects with structured matrix factorization, with application to remote sensing and data science; and ii) coarsely quantized MIMO transceiver designs. He has rich experience in editorial service, such as an Associate Editor, and then later, a Senior Area Editor, and then from 2021 to 2023, the Editor-in-Chief of IEEE TRANSACTIONS ON SIGNAL PROCESSING, and many others. He was a Tutorial Speaker in EUSIPCO 2011 and ICASSP 2014, and was an IEEE Signal Processing Society (SPS) Distinguished Lecturer in 2018–2019. He was the Recipient of the Research Excellence Award 2013–2014 by CUHK, the 2015 IEEE Signal Processing Magazine Best Paper Award, the 2016 IEEE Signal Processing Letters Best Paper Award, and the 2018 IEEE SPS Best Paper Award. He served as a Member of the Signal Processing for Communications and Networking Technical Committee (SPCOM-TC) in 2015–2020, a Member of the Signal Processing Theory and Methods Technical Committee (SPTM-TC) in 2012–2017, the SPS Regional Director-at-Large for Region 10 in 2020–2021, and a Technical Program Co-Chair of ICASSP 2023. He co-founded and co-organized One World Signal Processing in 2020, a virtual seminar series for signal processing.

Junbin Liu (Graduate Student Member, IEEE) received the B.S. degree from the South China University of Technology, Guangzhou, China, in 2017, and the M.S. degree from the University of Chinese Academy of Sciences, in 2020. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, The Chinese University of Hong Kong, under the supervision of Professor Wing-Kin Ma. His research focuses on statistical signal processing and optimization theories, and their broad applications.

Zihao Huang received the B.S. degree from Zhejiang University, Zhejiang, China, in 2018, and the M.Phil. degree from the Chinese University of Hong Kong, in 2022. His research focuses on deep learning for signal processing.