Unsupervised Massive MIMO Channel Estimation with Dual-Path Knowledge-Aware Auto-Encoders

Zhiheng Guo, Yuanzhang Xiao, *Member, IEEE*, Xiang Chen, *Member, IEEE*,

***Abstract*—In this paper, an unsupervised deep learning framework based on dual-path model-driven variational auto-encoders (VAE) is proposed for channel and angle-of-arrivals (AoAs) estimation in massive MIMO systems. Specifically designed for the task of channel estimation, the proposed VAE differs from the standard VAE in two aspects. First, the encoder is a dual-path neural network, where one path uses the received signal to estimate the path gains and path angles, and the other path uses the correlation matrix of the received signal to estimate AoAs. Second, the decoder has fixed weights that implement the signal propagation model, instead of having learnable parameters. This knowledge-aware decoder forces the encoder to output meaningful physical parameters of interests (i.e., path gains, path angles, and AoAs), which cannot be achieved by standard VAE. Rigorous analysis is carried out to characterize the multiple global optima and local optima of the channel estimation problem, which motivates the design of the dual-path encoder. By alternating between the estimation of path gains, path angles and the estimation of AoAs, the dual-path encoder is proved to converge. To further improve the convergence performance, a low-complexity initialization procedure is proposed to find good initial points. Numerical results validate theoretical analysis and demonstrate the performance improvements of our proposed framework.**

***Index Terms*—Massive MIMO, Angle of Arrival, Channel Estimation, Variational Auto-Encoder, Unsupervised Learning**

I. Introduction

MASSIVE multiple-input-multiple-output (MIMO) is one of the vital technologies to address the challenges of explosive data traffic and high quality of service requirement in fifth generation (5G) and beyond wireless communication systems. Theoretically, massive MIMO can enhance the capacity of a communication system with additional antennas [1]. In practice, accurate channel estimation is key to realize the potential gain of massive MIMO. However, there are challenges in accurate channel estimation, arising from sophisticated channel modeling [2], costly channel state information (CSI) [3], and high computational complexity due to a large number of antennas.

A. Prior Works

There are several strands of works for channel and/or angle-of-arrivals (AoAs) estimation in massive MIMO. The first strand of works are subspace-based estimation methods, dating back to classic algorithms such as multiple signal classification (MUSIC) [4] and estimation of signal parameters via rotational invariant techniques (ESPRIT) [5]. Later development of subspace-based methods include algorithms with lower complexity [6], methods for expanding the degrees of freedom [7], methods for special antenna array configurations [8], and methods for millimeter wave massive MIMO [9]. However, when applied to massive MIMO, subspace-based methods may have high computational complexity due to eigendecomposition of large covariance matrices of the received signals.

The second strand of works are based on compressed sensing [10]–[16]. Under certain assumptions on the sparsity and/or structures of the received signals, compressed sensing based methods usually pose the estimation problems as optimization problems, obtain equivalent semidefinite programs, and solve them by interior-point methods. Recent works have proposed to use the alternating direction method of multipliers to lower the computational complexity [17], and have extended the methods to arbitrary antenna arrays [17], [18] and massive MIMO [19]–[26]. While compressed sensing based methods have higher estimation accuracy than classic subspace-based methods, they also have higher computational complexity because the estimation involves solving an optimization problem.

The third strand of works use variational Bayesian inference on the received signals for channel and AoA estimation [27], [28]. Variational Bayesian methods calculate the posterior distribution of the channel gains and AoAs given the received signals, and choose the estimates to maximize the posterior distribution. While this approach yields highly accurate estimation [27] and can be computationally efficient [28], it may be hard to apply to more general channel models when the posterior distribution is intractable.

Finally, the recent development of machine learning has spurred the applications of deep learning to channel and AoA estimation. Deep learning based methods can be categorized into "data-driven" and "model-driven", a terminology coined by [29], [30]. Data-driven methods follow the end-to-end principle of deep learning, where the estimation is achieved by training neural networks on a large number of training data (e.g., received signals, covariance matrices) and training labels. Data-driven methods have utilized various neural network architectures (e.g., fully-connected networks, convolutional neural networks) and have been successfully applied to different scenarios (e.g., massive MIMO, vehicular networks) [31]–[45]. In contrast, model-driven methods exploit the established physical models and properties of signal propagation in wireless channels, and combine deep

Zhiheng Guo and Xiang Chen are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China.

Yuanzhang Xiao is with the Hawaii Advanced Wireless Technologies Institute, University of Hawaii, Honolulu, HI.

Xiang Chen is the corresponding author.

This work was supported by the National Key R&D Program of China under Grant 2019YFE0196400, the National Natural Science Foundation of China under Grant 61871035.

learning with traditional signal processing schemes such as MIMO detection [46], [47], subspace-based methods [48], [49] and compressed sensing based methods [50], [51]. Compared with other methods, deep learning based methods shift the majority of the computation to the training stage and has low complexity during deployment. However, with few exceptions [50], [51], deep learning based methods [31]–[49] are usually *supervised*, namely requiring accurately labeled data samples. In this sense, unsupervised methods may be more desirable and more aligned with traditional methods, which do not require labeled data. But without labels, unsupervised deep learning based methods are susceptible to local optima, as empirically observed in [51].

B. Contributions

In this paper, we propose an *unsupervised* deep learning based channel and AoA estimation framework in massive MIMO. Like other deep learning based methods, our framework has low computational complexity during deployment. As an unsupervised method, our framework achieves similar performance as existing supervised methods, and outperforms traditional methods such as MUSIC. Moreover, our framework can be applied to more general scenarios (e.g., correlated paths) beyond typical scenarios used in the literature.

There are two major hurdles in developing unsupervised deep learning methods for channel and AoAs estimation. The first hurdle is the *uninterpretability/unidentifiability of the latent variables/features* learned [52]. Specifically, the low-dimensional latent variables learned by unsupervised methods usually has no physical meaning. Therefore, if we directly applied an unsupervised learning method to the received signals, the output would not be the parameters we want to estimate (i.e., channel gains and AoAs). This is because without the labels, the network can only take the difference between the input (i.e., the received signals) and its reconstruction based on the latent variables as the loss function. Such a loss function also results in the second hurdle of *multiple local optima*. In other words, there may be multiple sets of channel parameters that are local or global optima of the loss function [51]. Without labels, it is hard to know which set of channel parameters is correct.

Our framework overcomes the above hurdles by restructuring the canonical variational auto-encoders (VAEs), inspired by the insights obtained from our rigorous analysis. A standard VAE consists of an encoder that learns low-dimensional latent variables from the input (i.e., the received signals) and a decoder that reconstructs the input. Our first key idea is to "hardwire" the decoder to implement the signal propagation model, instead of learning the decoder from data as in standard VAEs. This knowledge-aware decoder enforces the output of the encoder to be the parameters to estimate, and therefore break the first hurdle of uninterpretable/unidentifiable latent variables. To break the second hurdle of multiple local and global optima, we rigorously analyze the loss landscape and characterize the global and local optima. Our analysis leads us to implement the encoder as a dual-path neural network, where one path estimates the AoAs from the correlation matrix of the received signals and the other path estimates the path gains and path angles from the received signals. We update the weights of both paths in an alternating fashion, which ensures the convergence of the training. To avoid local optima, a low-cost method is proposed to find good initial points prior to the training. Ablation study is performed to demonstrate the necessity of the knowledge-aware decoder, the proposed alternating update manner for the dual-path neural network, and our choices of initial points. Numerical results validate our theoretical analysis and show the performance gains over traditional methods such as MUSIC.

Compared with existing works, our work falls in the small category of unsupervised deep learning based methods [50], [51]. While existing unsupervised methods numerically illustrate the multiplicity of local optima, our work provides theoretical analysis and characterization of the global and local optima. Our work also bears some similarity to variational Bayesian inference [27], [28]. By adopting the deep learning approach, our work applies to more general scenarios in which the posterior distribution of the latent variables is not tractable.

To the best of our knowledge, an unsupervised channel and AoA estimation framework with rigorous analysis is first proposed for massive MIMO in this paper. The main contributions of this paper are as follows.

- By restructuring the standard VAE, the proposed framework overcomes two challenges in unsupervised channel and AoA estimation. It breaks the hurdle of uninterpretable/unidentifiable latent variables by the knowledge-aware decoder, and alleviates the problem of multiple (bad) local optima by the dual-path encoder.
- Rigorous analysis of global optima (Propositions 3 and 4) and in the regime of large numbers of antennas, analysis of local optima (Proposition 5) is performed.
- Informed by the analysis, a two-phase training process is proposed. The first phase initializes the network with good initial points, and the second phase trains the dual-path encoder to learn the global optima.

The remainder of this paper is organized as follows. Section II introduces the signal model. Section III describes the proposed channel and AoA estimation framework. Section IV provides in-depth performance analysis of the proposed framework. Section V presents numerical results to validate our analysis and demonstrate the performance improvement of our proposed framework. Finally, Section VI concludes the paper.

Notations: In this paper, scalars, vectors and matrices are denoted by lower-case letters, bold lower-case letters and bold capital letters, respectively. The conjugate, transpose, conjugate transpose, and the Frobenius norm of a matrix are denoted by $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^H$, and $\|\cdot\|_F$, respectively. The Kronecker product is denoted by $\odot$ and $j = \sqrt{-1}$ is the imaginary unit. Superscripts $(\cdot)^R$ and $(\cdot)^I$ denote the real and imaginary parts of a variable. The sets of m-by-n real and complex matrices are $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$. Finally, the symbol $\mathbb{E}[\cdot]$ is the expectation operator.

II. Signal Propagation Model

Consider a typical massive MIMO uplink system with one base station (BS) and K users. The base station is equipped with a uniform linear array of N antennas, and each user equipment has one antenna. The users are assumed to be stationary for M time slots, during which the BS collects M snapshots of the received signals to estimate the users' locations. Each user k's location is specified by the distance to the BS and the AoA $\theta_k \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. User k's AoA θ_k is the impinging direction of its signal to the BS, relative to the broadside of the antenna array (i.e., the line perpendicular to the antenna array). Given the AoA θ_k, user k's signal experiences different phase shifts across the antenna array, as characterized by the array response vector

$$\boldsymbol{a}(\theta_k) = \left[1, e^{-j\frac{2\pi d}{\lambda}\sin\theta_k}, \ldots, e^{-j\frac{2\pi d}{\lambda}(N-1)\sin\theta_k}\right]^T, \quad (1)$$

where λ is the wavelength and $d \geq \frac{\lambda}{2}$ is the distance between adjacent antenna elements. User k's signal is also attenuated due to channel fading. In each time slot m, the vector of channel gains from the K users to the base station is denoted by $\boldsymbol{h}_m = \boldsymbol{\alpha}_m \odot e^{j\boldsymbol{\psi}_m}$, where $\boldsymbol{\alpha}_m \in \mathbb{R}^K$ is the vector of path gains and $\boldsymbol{\psi}_m \in \mathbb{R}^K$ is the vector of path angles. Following the literature [10], [12], [13], we assume block fading channels, namely the channel gains of different time slots are independent. We also assume that the vector of channel gains $\boldsymbol{h}_m$ is jointly Gaussian distributed, namely $\boldsymbol{h}_m \sim \mathcal{CN}(\boldsymbol{\mu}_{\boldsymbol{h}_m}, \boldsymbol{\Sigma}_{\boldsymbol{h}_m})$. Moreover, the channel gains and AoAs are assumed to be independent of each other [27], [28].

Based on the propagation model, the received signals during the M snapshots, denoted by $\boldsymbol{Y} \in \mathbb{C}^{N\times M}$, can be written as [10], [12], [13], [27], [28], [51]

$$\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{\theta}) \cdot [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M] + \boldsymbol{N}, \quad (2)$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]^T$ is the collection of users' AoAs, $\boldsymbol{A}(\boldsymbol{\theta}) = [\boldsymbol{a}(\theta_1), \boldsymbol{a}(\theta_2), \ldots, \boldsymbol{a}(\theta_K)] \in \mathbb{C}^{N\times K}$ is the array response matrix, and $\boldsymbol{N} = [\boldsymbol{n}_1, \ldots, \boldsymbol{n}_M] \in \mathbb{C}^{N\times M}$ is the complex circular symmetric white Gaussian noise with $\boldsymbol{n}_m \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$ for $m = 1, \ldots, M$.

The aim of this paper is to estimate the AoAs, θ_k for $k = 1, \ldots, K$, and the path gains $\boldsymbol{\alpha}_m$ and path angles $\boldsymbol{\psi}_m$ for $m = 1, \ldots, M$, based on the received signals $\boldsymbol{Y}$.

Note that we allow the channel gains of different users to be *correlated* (i.e., the covariance $\boldsymbol{\Sigma}_{\boldsymbol{h}_m}$ may not be a diagonal matrix). This is an extension to the independent assumption made by existing works [4], [5], [10], [12], [13], [49], [51].

III. Proposed Solution

This section describes our proposed unsupervised channel and AoA estimation framework for massive MIMO systems. An overview of our framework and high-level design principles are provided first, which is followed by detailed description.

Note that the proposed framework can be applied to general 2-dimensional array configuration (e.g., circular, rectangular); but some theoretical analysis (i.e., Proposition 4 and Proposition 5) hold for uniform linear arrays only.

Another widely-used model is $\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{\theta}) \cdot [\boldsymbol{h}_1 \odot \boldsymbol{s}_1, \ldots, \boldsymbol{h}_M \odot \boldsymbol{s}_M] + \boldsymbol{N}$, where $\boldsymbol{s}_m \in \mathbb{C}^K$ is the transmit signal in time slot m [4], [5], [49]. This model is equivalent to ours when the signals are known (e.g., pilot and reference signals).

A. Overview and Design Principles

A canonical approach of unsupervised learning is to use variational auto-encoders. The VAE aims to learn the latent variables $\boldsymbol{z}$ from the data (e.g., the received signal $\boldsymbol{Y}$ in our case). The standard VAE consists of an encoder that learns the latent variables $\boldsymbol{z}$ from the data $\boldsymbol{Y}$ and an decoder that reconstructs the input data $\widehat{\boldsymbol{Y}}$. However, the standard VAE has no control over the physical meanings of the latent variables $\boldsymbol{z}$, and therefore cannot be directly applied to our problem.

For our AoA and channel estimation problem, a redesigned VAE architecture is proposed, where *the decoder is fixed and implements the signal propagation model in* (2), instead of having learnable parameters. Specifically, the latent variable $\boldsymbol{z}$ is a vector of length $K(1+M)$, namely

$$\boldsymbol{z} = \left[\boldsymbol{z}_0^T, \boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_M^T\right]^T, \quad (3)$$

where $\boldsymbol{z}_0 \in \mathbb{R}^K$ and $\boldsymbol{z}_m \in \mathbb{C}^K$ for $m = 1, \ldots, M$. Then the decoder reconstructs the received signals based on the latent variable using the following equation:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{A}(\boldsymbol{z_0}) \cdot [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M]. \quad (4)$$

Comparing (4) with the signal propagation model (2), it can be seen that $\boldsymbol{z}_0$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$ play the roles of the AoAs and the channel gains. In this way, the knowledge-aware decoder forces the latent variables to be the parameters to estimate, namely the AoAs and the channel gains. For this reason, we also write $\boldsymbol{z}_0 = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{z}_m = \widehat{\boldsymbol{h}}_m$, where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{h}}_m$ explicitly denote the estimated AoAs and channel gains.

Similar to standard VAEs, the encoder of our restructured VAE outputs the posterior distribution of the latent variable given the received signal, denoted by $p(\boldsymbol{z}|\boldsymbol{Y})$. In some special cases, the posterior distribution $p(\boldsymbol{z}|\boldsymbol{Y})$ can be expressed analytically [27], [28]. However, in the more general scenarios considered in this paper (e.g., correlated channel gains), it is hard to derive the analytical expression of the posterior distribution $p(\boldsymbol{z}|\boldsymbol{Y})$. Therefore, the encoder outputs an *approximate posterior distribution*, denoted by $q(\boldsymbol{z}|\boldsymbol{Y})$. The goal is then to minimize the difference between the approximate posterior distribution $q(\boldsymbol{z}|\boldsymbol{Y})$ and the true posterior $p(\boldsymbol{z}|\boldsymbol{Y})$, measured by the Kullback–Leibler (KL) divergence $\mathcal{D}_{KL}(q||p)$. As in existing works [27], [28], [53], we restrict to approximate posterior distributions that can be factorized as

$$q(\boldsymbol{z}|\boldsymbol{Y}) = q_0\left(\widehat{\boldsymbol{\theta}}|\boldsymbol{Y}\right) \cdot \prod_{m=1}^{M} q_m\left(\widehat{\boldsymbol{h}}_m|\boldsymbol{Y}\right). \quad (5)$$

The following proposition characterizes the optimal approximate posterior distribution that minimizes the KL divergence.

Proposition 1: For each time slot $m = 1, \ldots, M$, the optimal posterior distribution $q_m^*(\widehat{\boldsymbol{h}}_m|\boldsymbol{Y})$ of the channel gain is a circularly symmetric Gaussian distribution, namely

$$q_m^*\left(\widehat{\boldsymbol{h}}_m|\boldsymbol{Y}\right) = \frac{1}{\pi^K|\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}|} e^{-(\widehat{\boldsymbol{h}}_m - \boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})^H \boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}^{-1}(\widehat{\boldsymbol{h}}_m - \boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})}. \quad (6)$$

Proof: See Appendix A. ■

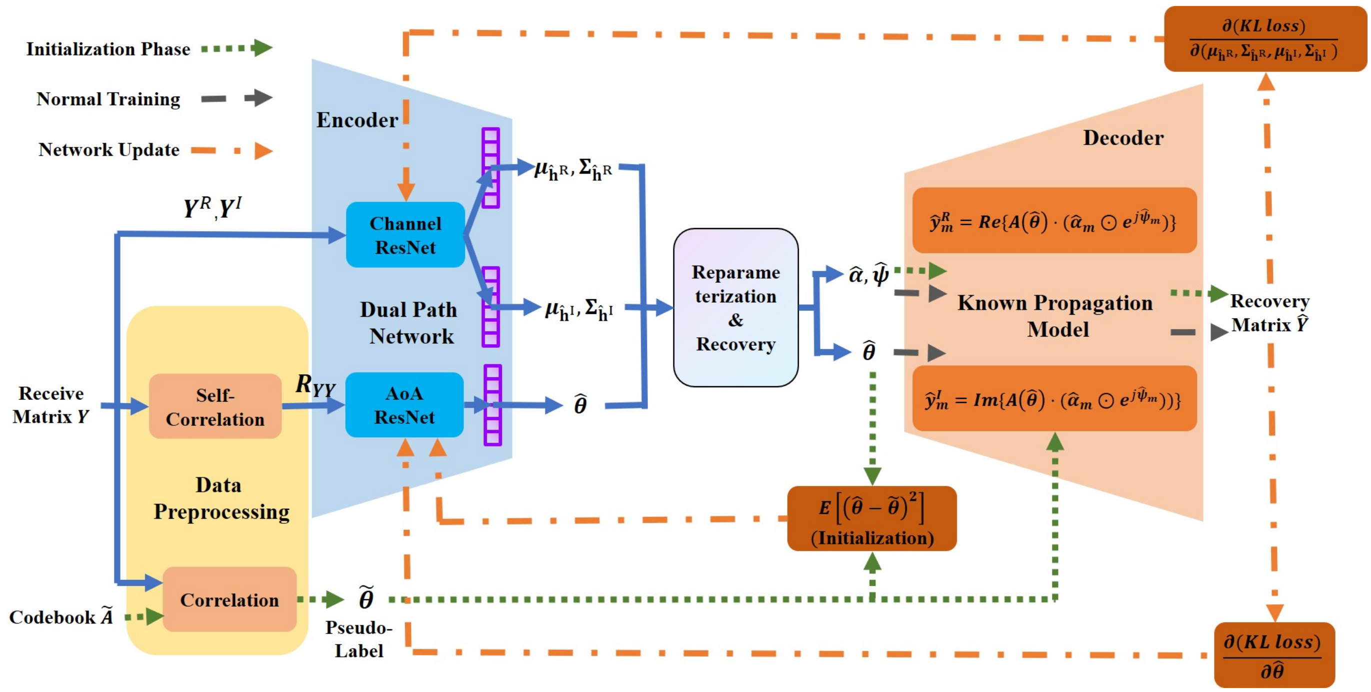


Fig. 1: Illustration of the proposed VAE with a dual-path encoder and a knowledge-aware decoder.

Proposition 1 shows that the optimal posterior distribution of channel gains is the circularly symmetric Gaussian distribution, which can be completely determined by the mean vector and the covariance matrix. Therefore, the encoder only needs to output the estimated mean $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$ and covariance $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}$ of the channel gains. For the AoAs, a commonly-used prior distribution of the AoAs is the uniform distribution, under which it is difficult to calculate the true posterior or the optimal approximate posterior distribution. Thus, we let the encoder output a point estimate $\widehat{\boldsymbol{\theta}}$ of the AoAs. This is equivalent to using the Dirac distribution $\delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$ as the approximate posterior distribution of the AoAs.

Based on the optimal posterior distribution in Proposition 1 and the knowledge-aware decoder in (4), we can derive the loss function of the proposed framework. Although the loss function is initially defined as the KL divergence, it is well know that minimizing the KL divergence is equivalent to maximizing the evidence lower bound (ELBO) [54]. Hence, we define the loss function as the negative of the ELBO and calculated in the following proposition.

Proposition 2: In the proposed framework with the knowledge-aware decoder, the loss function, defined as the negative of the ELBO, can be written analytically as in (7).

Proof: See Appendix B. ■

The ELBO (7) consists of two parts. The first part is the difference of the entropy of the prior distribution and that of the approximate posterior distribution. The second part is the mean square error (MSE) between the received signals and the reconstructed signals. When using the ELBO as the loss function during training, the restructured VAE aims to reconstruct the received signal as accurately as possible while taking into account the prior distribution of the AoAs and channel gains. This is different from the prior work [51] that use the MSE as the loss function.

In sum, the encoder of the proposed VAE takes the received signal $\boldsymbol{Y}$ as input and outputs the estimated parameters $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$ and $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}$ of the optimal posterior distribution of the channel gains and the estimated AoAs $\widehat{\boldsymbol{\theta}}$. Given these estimates, random samples of the channel gains are drawn and fed into the decoder, who reconstructs the received signal $\widehat{\boldsymbol{Y}}$ by (4).

B. Implementation Details of the Proposed framework

Fig. 1 shows a detailed diagram of the proposed VAE. Since neural networks generally use real numbers, all the variables in the diagram are real-valued. We denote the real and imaginary parts of a complex variable by superscripts R and I, respectively. The proposed framework consists of a data pre-processing module, a dual-path encoder, a reparameterization and recovery module, and a decoder, which will be described in details.

$$\mathcal{L}^{(train)} = \underbrace{\sum_{m=1}^{M} \left\{ -\log\left|\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\right| + \operatorname{tr}\left(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{h}_m}\right) + (\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m} - \boldsymbol{\mu}_{\boldsymbol{h}_m})^H \boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1} (\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m} - \boldsymbol{\mu}_{\boldsymbol{h}_m}) \right\}}_{\text{KL divergence between the approximate posterior distribution and the prior distribution}} + \underbrace{\mathbb{E}_{q(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{h}}_1, \ldots, \widehat{\boldsymbol{h}}_M | \boldsymbol{Y})} \left[\frac{1}{\sigma^2} \left\| \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \right\|_F^2 \right]}_{\text{mean squared reconstruction error}}, \tag{7}$$

1) Data Preprocessing: The data preprocessing module has two functionalities. First, it calculates the empirical covariance matrix from the received signals as

$$\boldsymbol{R}_{\boldsymbol{YY}} = \frac{1}{M}\boldsymbol{Y}\boldsymbol{Y}^H, \tag{8}$$

which will be used for AoA estimation.

Second, it calculates pseudo-labels $\widetilde{\boldsymbol{\theta}}$ of the AoAs, which will be used in the initial stage of the training. The pseudo-labels take values from a predefined set

$$\mathcal{S}_\theta = \left\{\underline{\theta}, \underline{\theta} + \Delta_\theta, \ldots, \bar{\theta}\right\}, \tag{9}$$

where $\underline{\theta}$ is the minimum AoA, $\bar{\theta}$ is the maximum AoA, and Δ_θ is the precision of the pseudo-labels. The received signal $\boldsymbol{Y}$ is then compared to the array response vectors $\boldsymbol{a}(\theta)$ at all the angle θ in the set $\mathcal{S}_\theta$, by calculating the empirical correlation

$$r(\theta, \boldsymbol{Y}) = \frac{1}{M}\left|\boldsymbol{1}_M^T \boldsymbol{Y}^H \boldsymbol{a}(\theta)\right|, \tag{10}$$

where $\boldsymbol{1}_M$ is a M-dimensional column vector of all ones. Finally, the set $\widetilde{\Theta}$ of K angles corresponding to the K largest correlations are selected, namely

$$\widetilde{\Theta} = \arg \max_{\widetilde{\Theta} \subset \mathcal{S}_\theta, |\widetilde{\Theta}|=K} \sum_{\theta \in \widetilde{\Theta}} r(\theta, \boldsymbol{Y}). \tag{11}$$

The pseudo-labels $\widetilde{\boldsymbol{\theta}}$ are obtained from the set $\widetilde{\Theta}$.

2) The Dual-Path Encoder: As shown in Fig. 1, the encoder is a dual-path neural network, which is divided into the channel Residual Neural Network (ResNet) and the AoA ResNet. The channel ResNet takes the real and imaginary parts of the received signals matrix $\boldsymbol{Y}_R$ and $\boldsymbol{Y}_I$, and outputs the estimated means and covariance matrices of the channel gains, $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$ and $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}$ for $m = 1, \ldots, M$. The AoA ResNet takes the empirical correlation matrix $\boldsymbol{R}_{\boldsymbol{YY}}$ calculated by the preprocessing module, and outputs the estimated AoAs $\widehat{\boldsymbol{\theta}}$.

The proposed dual-path neural network separates the estimation of AoAs and channel gains, and enables the proposed training process that alternates between updating the weights of the channel ResNet and the AoA ResNet. As we will show in Proposition 3, the alternating training process facilitates the convergence to the optimal estimates.

3) Reparameterization and Recovery: As seen from the loss function (7), the expectation is taken over the estimated parameters. Standard Monte Carlo simulations to calculate the empirical expectation would result in a high variance of the gradient [54]. To avoid such higher variance, a reparameterization trick has been introduced in the literature [54].

Since it has been proved that the optimal posterior distribution of the channel gains is Gaussian, the reparameterization trick generates standard Gaussian random vectors $\boldsymbol{\epsilon}_{\widehat{\boldsymbol{h}}_m^R}, \boldsymbol{\epsilon}_{\widehat{\boldsymbol{h}}_m^I} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ and samples of the estimated parameters by

$$\widehat{\boldsymbol{h}}_m^R = \boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m^R} + \boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m^R} \cdot \boldsymbol{\epsilon}_{\widehat{\boldsymbol{h}}_m^R} \text{ and } \widehat{\boldsymbol{h}}_m^I = \boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m^I} + \boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m^I} \cdot \boldsymbol{\epsilon}_{\widehat{\boldsymbol{h}}_m^I}. \tag{12}$$

Finally, samples of the path gain and the path angle can be recovered from the estimated parameters by:

$$\begin{cases} \widehat{\boldsymbol{\alpha}}_m = \text{abs}\left(\widehat{\boldsymbol{h}}_m\right) \\ \widehat{\boldsymbol{\psi}}_m = \arctan\left[\text{diag}\left(\widehat{\boldsymbol{h}}_m^R\right)^{-1} \cdot \widehat{\boldsymbol{h}}_m^I\right] \end{cases}, \tag{13}$$

where $\text{abs}(\cdot)$ returns the magnitudes of all elements in a vector, and $\text{diag}(\widehat{\boldsymbol{h}}_m^R)$ is the diagonal matrix with $\widehat{\boldsymbol{h}}_m^R$ as its diagonal elements.

4) The Decoder: As discussed before, the knowledge-aware decoder takes the samples of path gains $\widehat{\boldsymbol{\alpha}}_m$ and path angles $\widehat{\boldsymbol{\psi}}_m$ and the AoAs $\widehat{\boldsymbol{\theta}}$, and outputs the reconstructed signals based on the propagation model (2).

C. Training Process

As will be proved in Proposition 4 and Proposition 5, the loss function (7) has multiple global and local optima. Hence, it is important to start the training process from good initial points that are far away from local minima. To obtain good initial points, the training process is divided into two phases – the initialization phase and the normal training phase.

In the initialization phase, the network uses the pseudo-labels $\widetilde{\boldsymbol{\theta}}$ calculated by the preprocessing module, and updates the weights to minimize the modified loss function as follows:

$$\mathcal{L}^{(init)} = \widetilde{\mathcal{L}}^{(train)} + \gamma \cdot \mathbb{E}\left[\left\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right\|_2^2\right], \tag{14}$$

where $\widetilde{\mathcal{L}}^{(train)}$ replaces the estimates $\widehat{\boldsymbol{\theta}}$ with the pseudo-labels $\widetilde{\boldsymbol{\theta}}$ in the loss function $\mathcal{L}^{(train)}$ in (7), and γ is a scaling factor to ensure that the two terms in (14) are on the same order. After the initialization, the encoder neural network is trained in the training phase with the loss function $\mathcal{L}^{(train)}$ in (7).

As seen from (14), at the end of the initialization phase, the AoA ResNet learns to predict the pseudo-labels $\widetilde{\boldsymbol{\theta}}$. Then during the training phase, the AoA ResNet is more likely to stay in the neighborhood of the pseudo-labels. Our presumption is that the pseudo-labels are in close vicinity of the true AoAs. Under this presumption, the initialization phase with the modified loss function (14) can prevent the model from converging to bad local optima.

IV. Performance Analysis

This section analyzes the global optima and local optima of the loss function (7), which sheds light on the design of the proposed framework and the training process.

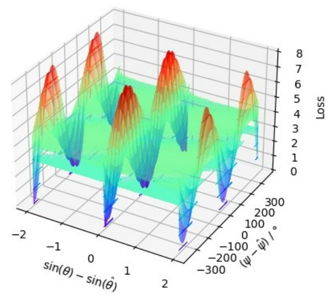
The multiplicity of global optima occurs when there are multiple sets of AoAs, path gains and path angles that result in the same received signal. These multiple sets of parameters all minimize the loss function (7). Therefore, it might seem impossible to get accurate channel estimation without supervision of true labels. However, the following proposition shows that we can alleviate the problem of multiple global optima by separating the channel estimation and the AoA estimation.

Proposition 3: When the AoA estimation is accurate (i.e., $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$), the optimal estimates of the channel gains that minimize the loss function in (7) are

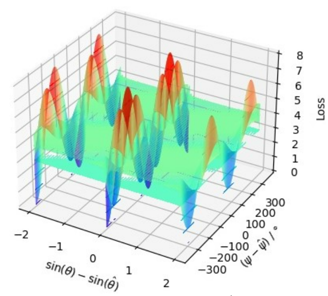
$$\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m} = \left(\boldsymbol{A}^H\boldsymbol{A} + \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\right)^{-1}\left(\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{h}_m + \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\boldsymbol{\mu}_{\boldsymbol{h}_m}\right). \tag{15}$$

Proof: See Appendix C. ■

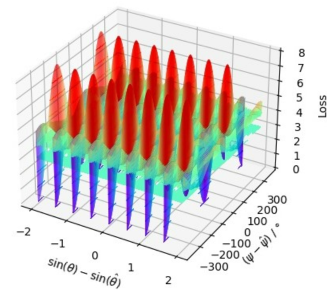
Proposition 3 indicates that once the accurate AoA estimation is obtained, there is a *unique* set of channel gains that minimize the loss function. This motivates us to design



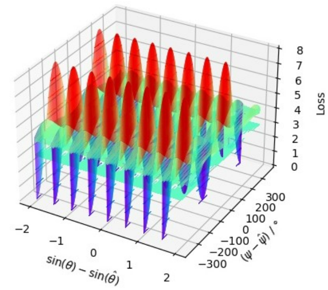
(a) 32 antennas with $d/\lambda = 0.5$



(b) 64 antennas with $d/\lambda = 0.5$

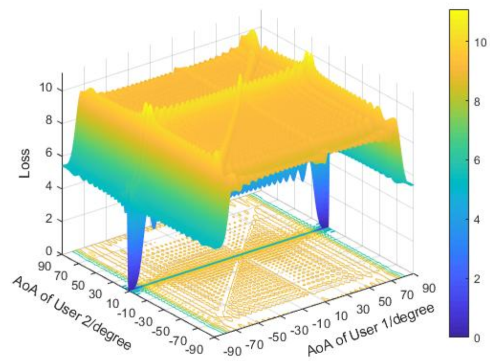


(c) 32 antennas with $d/\lambda = 2$

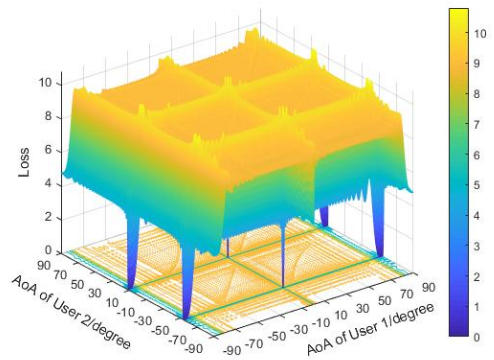


(d) 64 antennas with $d/\lambda = 2$

Fig. 2: Illustration of the loss landscape with respect to the estimates of AoAs and path angles (estimation of path gains is assumed to be accurate).



(a) True AoAs $= (-90.0°, 11.0°)$ with $d/\lambda = 0.5$



(b) True AoAs $= (-90.0°, 11.0°)$ with $d/\lambda = 1$

Fig. 3: Illustration of the loss landscape with respect to AoA estimates (channel estimation is assumed to be accurate).

the encoder as a dual-path neural network, where the AoA estimation is separated from the estimation of path gains and path angles. It allows us to update the weights of the two networks in an alternating fashion. In this way, if the AoA estimation network converges to the optimal solution, the channel estimation network has a unique global optimum.

Note that the optimal estimates of channel gains in (15) maximize the posterior distribution of the channel gains given the signals, and therefore may not equal to the true means of the channel gains. They converge to the true means of the channel gains when the SNR goes to infinity.

Now that we have characterized the optimal estimation of the channel gains, we turn to the analysis of AoA estimation. In contrast to channel estimation, the optimal AoA estimates that minimize the loss function are not unique. This is because the phases of the received signals and those of the reconstructed signals are the same if for all $m = 1, \ldots, M$ and $n = 0, \ldots, N-1$, there exists an integer $l_{m,n} \in \mathbb{Z}$ such that

$$\left(\phi_{k,m} - \widehat{\phi}_{k,m}\right) - \frac{2\pi n d}{\lambda}\left(\sin\theta_k - \sin\widehat{\theta}_k\right) = 2\pi l_{m,n}.$$

The following proposition characterizes the multiplicity of globally optimal AoA estimates.

Proposition 4: Suppose that the channel estimation is accurate, namely $\widehat{\boldsymbol{h}}_m = \boldsymbol{h}_m$ for $m = 1, \ldots, M$. Then the AoA estimation minimizes the loss function if each user k's AoA estimation takes one of the values:

$$\widehat{\theta}_k = \arcsin\left(\sin\theta_k - l \cdot \frac{\lambda}{d}\right), \tag{16}$$

where l is an integer that satisfies

$$\left\lceil \frac{d}{\lambda}\left(\sin\theta_k - 1\right)\right\rceil \le l \le \left\lfloor \frac{d}{\lambda}\left(\sin\theta_k + 1\right)\right\rfloor, \tag{17}$$

where $\lceil x \rceil$ is the minimum integer no smaller than x and $\lfloor x \rfloor$ is the maximum integer no larger than x.

Proof: See Appendix D. ■

Proposition 4 indicates that even if the estimation of channel gains is accurate, there are multiple sets of AoAs that minimize the loss function. It also shows that the number of global optima increases with the carrier frequency, suggesting that the problem is more severe in millimeter wave systems due to high frequencies. Another important observation is that the number of global optima does not depend on the number of antenna elements. Therefore, simply increasing the number of antenna elements may not solve the problem. A straightforward way to alleviate the problem of multiple global optima is to apply a spatial filter [55] or to divide the cell into sectors [56].

Fig. 2 and Fig. 3 illustrate the multiple global optima in AoA estimation under different numbers of antennas and antenna spacing. Fig. 2 shows the loss landscape when we vary AoA and path angle estimates of one user with other users' AoA and channel estimates being accurate. Fig. 3 shows the loss landscape when we vary the AoA estimates of two users with all other estimates being accurate. Both figures validate the observations from Proposition 4 that the number of global optima increases with the carrier frequency (i.e., d/λ) and does not decrease with the number of antennas.

The loss landscapes in Fig. 2 and Fig. 3 exhibit a plethora of stationary points, which is a more serious issue compared to the multiplicity of global optima. This is because the estima-

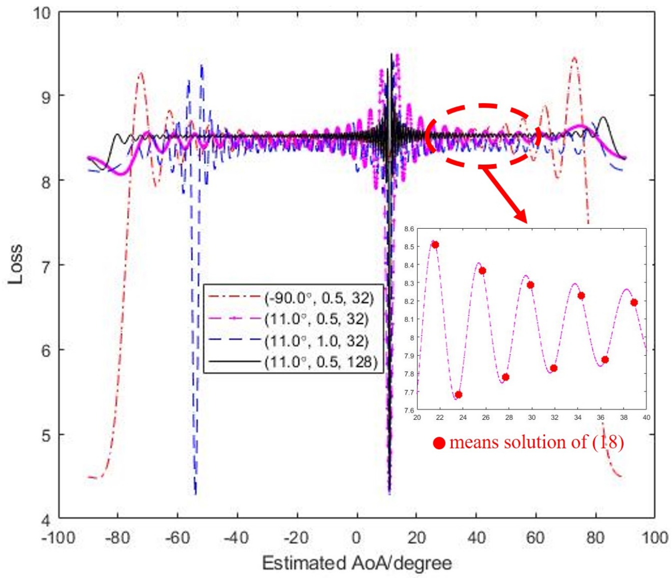


Fig. 4: Illustration of the loss landscape with respect to AoA estimates. The triplets in the legend indicate (true AoA, d/λ, number of antennas). The red dots are the stationary points calculated from Proposition 5.

tion at stationary points can be highly inaccurate and because neural networks, typically trained by first-order methods, are susceptible to stationary points. Hence, it is important to analyze stationary points.

Proposition 5: Suppose that the channel estimation is accurate, namely $\widehat{\boldsymbol{h}}_m = \boldsymbol{h}_{\boldsymbol{m}}$ for all m, and that the AoA estimation is accurate except for user k, namely $\widehat{\theta}_i = \theta_i$ for all $i \neq k$. Given an arbitrary small number $\epsilon > 0$, there exists a threshold $\underline{N}(\epsilon)$ such that for any antenna array that has more than $\underline{N}(\epsilon)$ antennas (i.e., $N > \underline{N}(\epsilon)$), the stationary points of the loss function are within ϵ of the solutions to the following equation:

$$\cos\widehat{\theta}_k \cdot \left[\frac{\cos\left(\zeta_{\widehat{\theta}_k}(N-1)\right)}{\zeta_{\widehat{\theta}_k}} - \frac{\cos\left(\eta_{\widehat{\theta}_k}(N-1)\right)}{\eta_{\widehat{\theta}_k}}\right] = 0, \tag{18}$$

where $\eta_{\widehat{\theta}_k} = \frac{2\pi d}{\lambda}(\sin\theta_k + \sin\widehat{\theta}_k)$ and $\zeta_{\widehat{\theta}_k} = \frac{2\pi d}{\lambda} 2\sin\widehat{\theta}_k$.

Proof: See Appendix E. ■

Proposition 5 characterizes the stationary points in the regime of large numbers of antennas, which is especially relevant for massive MIMO. It stresses the importance of good initial points for AoA estimation.

From (18), we can see that there is always a stationary point around $\hat{\theta}_k = \pm 90°$, namely when the impinging signal is parallel to the antenna array. This stationary point can usually be avoided due to sectorization. Other stationary points $\hat{\theta}_k$ roughly satisfy

$$\frac{\cos\left(\zeta_{\widehat{\theta}_k}(N-1)\right)}{\zeta_{\widehat{\theta}_k}} = \frac{\cos\left(\eta_{\widehat{\theta}_k}(N-1)\right)}{\eta_{\widehat{\theta}_k}},$$

which depends on the true AoA θ_k, the spacing between antennas d/λ, and the number N of antennas.

Fig. 4 illustrates the loss landscape and the stationary points calculated from Proposition 5. It can be observed that there is always a stationary point around $\pm 90°$. It is also observed that the characterization of stationary points from Proposition 5 is close to simulation results. These observations validate Proposition 5. Moreover, Fig. 4 shows that the density of stationary points is higher around the true AoA, which underscores the importance of good initialization.

V. Numerical Simulation

In this section, we first evaluate the performance of the proposed framework against representative methods. Then we conduct ablation study to demonstrate how the proposed framework overcomes the challenges of uninterpretable latent variables and multiple local optima in unsupervised learning.

We set the number of users $K = 3$, the number of snapshots $M = 40$, the number of antennas $N = 32$, and the antenna spacing $d/\lambda = 0.5$. The cell is divided into three 120-degree sectors. The neural network is trained using the adaptive moment estimation (Adam) algorithm with a fixed learning rate of 10^{-4}. Both AoAs and channel ResNets use the ResNet-18 architecture. The batch size of the input data is 128. The training data set has 40000 samples and the test set has 10000 samples, all generated from the signal propagation model (2). The maximum number of epochs of the initialization phase is 200, and that of the training phase is 800. Training is done separately for each SNR.

A. Performance Comparison

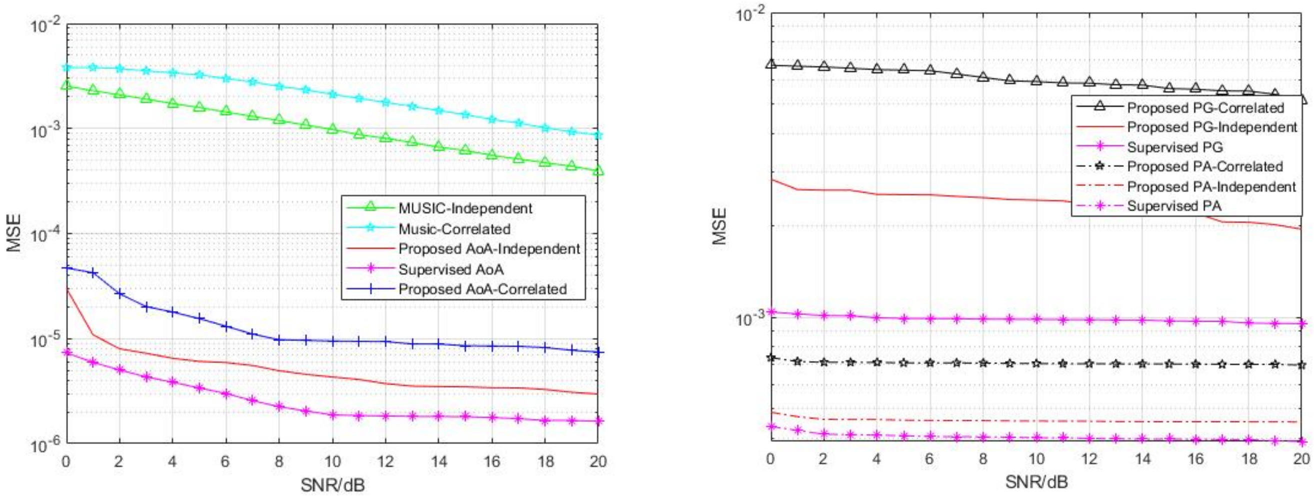
We first compare the estimation accuracy of the proposed framework with the canonical MUSIC algorithm [4] and the supervised learning method. The supervised learning method uses the labels of true AoAs, path gains, and path angles and minimizes the MSE between the estimates and the labels. The performance comparison is done in two cases where the users have independent and correlated channels, respectively. The case of correlated channels demonstrates that the proposed framework can work beyond the typical independent assumption made in some existing works.

Fig. 5 shows the MSEs of AoA estimation and channel estimation. For the MUSIC algorithm, the MSEs of path gains and path angles are not shown because the algorithm does not produce these estimates. For the supervised learning method, only the case of independent channels is shown because the MSE under correlated channels is similar.

It is observed that the proposed framework achieves almost identical performance as the supervised learning method. This demonstrates the advantage of the proposed method: the removal of labels in our method comes at almost no cost. This achievement is not trivial, because the classic unsupervised MUSIC algorithm has much higher AoA estimation errors.

B. Ablation Study

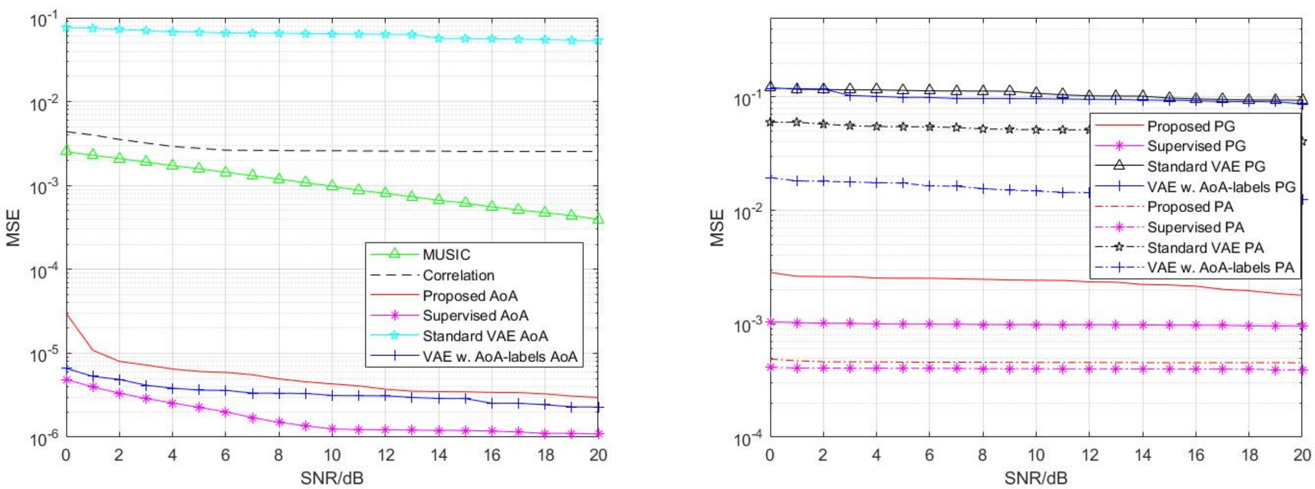
1) Dealing with uninterpretable latent variables: The proposed framework uses the knowledge-aware decoder, which implements the signal propagation model, to induce the encoder to output the parameters of interest. We perform ablation study to demonstrate the effectiveness of the knowledge-aware



(a) Estimation of AoAs

(b) Estimation of path gains (PG) and path angles (PA)

Fig. 5: Comparison of estimation accuracy against MUSIC and the supervised learning method. The comparison is performed in the cases of independent channels and correlated channels.



(a) Estimation of AoAs

(b) Estimation of path gains (PG) and path angles (PA)

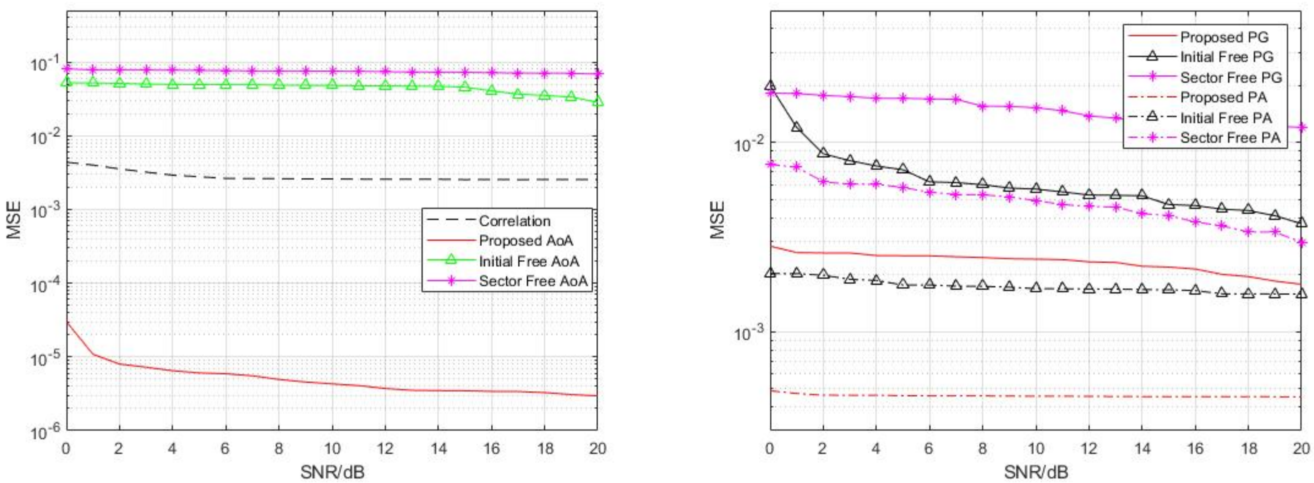
Fig. 6: Ablation study to show how the proposed framework overcomes the challenge of uninterpretable latent variables.

decoder. Specifically, we provide the standard VAE with labels of true AoAs and evaluate its performance. This variant of the standard VAE uses the MSE between the estimated AoAs and the true AoAs as part of the loss function. In this way, some components of the encoder output are enforced to be the AoAs, while the other components have no physical meaning.

Fig. 6 shows the estimation accuracy of the proposed framework, in comparison with the standard VAE and its variant with AoA labels. We can see that the standard VAE has extremely high MSEs in all estimation tasks, and that its variant with AoA labels has slightly higher MSEs than the proposed framework in AoA estimation, and much higher MSE in channel estimation. In other words, the standard VAE cannot enforce physical meaning of the encoder output, unless labels are provided. This ablation study shows that the proposed knowledge-aware decoder solves the issue of uninterpretable latent variables in unsupervised learning.

2) Dealing with multiple local optima: Our theoretical analysis identifies multiple global optima and indicate that the issue can be resolved by dividing the cell into sectors (Proposition 4). Our analysis also stresses the importance of finding good initial points due to local optima around the true AoAs (Proposition 5).

To demonstrate how the proposed framework deals with multiple local optima, we evaluate the performance of the proposed framework when there is no sectorization and when there is no initialization phase. As shown in Fig. 7, the proposed framework would have much higher MSEs if there were no sectorization. It also shows that the proposed initialization phase greatly helps to improve the estimation accuracy over



(a) Estimation of AoAs

(b) Estimation of path gains (PG) and path angles (PA)

Fig. 7: Ablation study to show how the proposed framework overcomes the challenge of multiple local optima.

the case where the initial points are randomly chosen.

VI. Conclusions

In this paper, an unsupervised learning framework based on a redesigned VAE is proposed for joint channel and AoA estimation for massive MIMO. The proposed framework solves two challenges of unsupervised learning in the context of channel and AoA estimation. It solves the first challenge of uninterpretable/unidentifiable latent variables by using a knowledge-aware decoder. The knowledge-aware decoder implements the known signal propagation model and enforces the encoder output to be the parameters to estimate. The proposed framework solves the second challenge of multiple local optima by using a dual-path encoder and adopting a two-phase training process. We analytically derive the optimal approximate posterior distributions of the estimated parameters and the ELBO of the proposed framework as the loss function. Theoretical analysis of the loss landscape is carried out to inform our design of the framework and is validated by numerical simulations. Numerical experiments demonstrate that our proposed framework achieves almost the same performance as the supervised learning method and outperforms the traditional unsupervised method. simulations validate the theoretical analysis. Ablation study is performed to evaluate the contributions of the key components of the proposed framework in overcoming the aforementioned challenges of unsupervised learning.

Appendix A

According to [53, Lemma 4.1], the optimal approximate posterior distribution satisfies

$$q_m^*\left(\boldsymbol{z}_m|\boldsymbol{Y}\right)=\frac{\exp \mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_m|\boldsymbol{z}_{-m},\boldsymbol{Y}\right)\right]}{\int \exp \mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_m|\boldsymbol{z}_{-m},\boldsymbol{Y}\right)\right]\, d\boldsymbol{z}_m},$$

where $\boldsymbol{z}_{-m}$ is the collection of all the latent variables except $\boldsymbol{z}_m$, and $q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y}) = \prod_{i=1,i\neq m}^{M} q_i^*(\boldsymbol{z}_i|\boldsymbol{Y})$. For $m = 1,\ldots,M$, the latent variable $\boldsymbol{z}_m$ is the channel gain $\boldsymbol{h}_m$. Hence, we have

$$\begin{aligned}
&\log q_m^*\left(\boldsymbol{z}_m|\boldsymbol{Y}\right) \qquad (19)\\
=\;& \mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_m|\boldsymbol{z}_{-m},\boldsymbol{Y}\right)\right]+C_1\\
=\;& \mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_m,\boldsymbol{z}_{-m},\boldsymbol{Y}\right)\right]\\
&-\mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_{-m},\boldsymbol{Y}\right)\right]+C_1\\
=\;& \mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_m,\boldsymbol{z}_{-m},\boldsymbol{Y}\right)\right]+C_2\\
=\;& \mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{Y}|\boldsymbol{z}_m,\boldsymbol{z}_{-m}\right)\right]\\
&+\mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\sum_{i=0}^{M}\log p\left(\boldsymbol{z}_i\right)\right]+C_2\\
=\;& \mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\sum_{i=1}^{M}\log p\left(\boldsymbol{y}_i|\boldsymbol{z}_0,\boldsymbol{z}_i\right)\right]\\
&+\mathbb{E}_{q^*_{-m}(\boldsymbol{z}_{-m}|\boldsymbol{Y})}\left[\sum_{i=0}^{M}\log p\left(\boldsymbol{z}_i\right)\right]+C_2\\
=\;& \log p\left(\boldsymbol{y}_m|\boldsymbol{z}_0,\boldsymbol{z}_m\right)+\log p\left(\boldsymbol{z}_m\right)+C_3\\
=\;& -\frac{1}{\sigma^2}\left[\boldsymbol{y}_m-\boldsymbol{A}(\boldsymbol{z}_0)\cdot\boldsymbol{z}_m\right]^H\left[\boldsymbol{y}_m-\boldsymbol{A}(\boldsymbol{z}_0)\cdot\boldsymbol{z}_m\right]\\
&-\left(\boldsymbol{z}_m-\boldsymbol{\mu}_{\boldsymbol{h}_m}\right)^H\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\left(\boldsymbol{z}_m-\boldsymbol{\mu}_{\boldsymbol{h}_m}\right)+C_4,
\end{aligned}$$

where C_1, C_2, C_3, C_4 are constant with respect to $q_m^*\left(\boldsymbol{z}_m\right)$. The last term in the above equation is a quadratic function of $\boldsymbol{z}_m$. Therefore, the optimal approximate posterior distribution $q_m^*\left(\boldsymbol{z}_m|\boldsymbol{Y}\right)$ is circularly symmetric Gaussian.

Appendix B

According to [54, Sec. 2.2], the ELBO is defined as

$$-\mathcal{D}_{KL}\left(q\left(\boldsymbol{z}|\boldsymbol{Y}\right)||p\left(\boldsymbol{z}\right)\right)+\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{Y}|\boldsymbol{z}\right)\right]. \qquad (20)$$

Therefore, the loss function, defined as the negative of the ELBO, is

$$\begin{aligned}
&\mathcal{D}_{KL}\left(q\left(\boldsymbol{z}|\boldsymbol{Y}\right)||p\left(\boldsymbol{z}\right)\right)-\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{Y}|\boldsymbol{z}\right)\right] \quad (21)\\
=\;& \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log\frac{q\left(\boldsymbol{z}|\boldsymbol{Y}\right)}{p\left(\boldsymbol{z}\right)}\right]-\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{Y}|\boldsymbol{z}\right)\right].
\end{aligned}$$

Next, we look at each term in the above expression.

The term $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log q\left(\boldsymbol{z}|\boldsymbol{Y}\right)\right]$ is the negative of the differential entropy of the approximate posterior distribution $q\left(\boldsymbol{z}|\boldsymbol{Y}\right)$. Since the posterior distribution $q_0\left(\boldsymbol{z}_0|\boldsymbol{Y}\right)$ of the AoA $\boldsymbol{z}_0=\boldsymbol{\theta}$ is assumed to be a Dirac distribution, its differential entropy is zero. Since the optimal approximate posterior distribution $q_m\left(\boldsymbol{z}_m|\boldsymbol{Y}\right)$ of the channel gains $\boldsymbol{z}_m=\boldsymbol{h}_m$ for $m=1,\ldots,M$ is proved to be circularly symmetric Gaussian, its differential entropy is $\log\left[(\pi e)^K|\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}|\right]$. Hence, we have

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log q\left(\boldsymbol{z}|\boldsymbol{Y}\right)\right] &= \textstyle\sum_{m=0}^{M}\mathbb{E}_{q_m(\boldsymbol{z}_m|\boldsymbol{Y})}\left[\log q_m\left(\boldsymbol{z}_m|\boldsymbol{Y}\right)\right]\\ &= -\textstyle\sum_{m=1}^{M}\log|\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}|+C_1, \end{aligned} \tag{22}$$

where C_1 is a constant.

The term $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}\right)\right]$ can be calculated as

$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}\right)\right] \qquad (23)\\ =\;& \mathbb{E}_{q_0(\boldsymbol{z}_0|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_0\right)\right]+\sum_{m=1}^{M}\mathbb{E}_{q_m(\boldsymbol{z}_m|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{z}_m\right)\right]\\ =\;& \sum_{m=1}^{M}\mathbb{E}_{q_m(\boldsymbol{z}_m|\boldsymbol{Y})}\left[-\left(\boldsymbol{z}_m-\boldsymbol{\mu}_{\boldsymbol{h}_m}\right)^H\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\left(\boldsymbol{z}_m-\boldsymbol{\mu}_{\boldsymbol{h}_m}\right)\right]\\ &+C_2,\\ =\;& -\sum_{m=1}^{M}\left(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m}\right)^H\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\left(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m}\right)\\ &-\sum_{m=1}^{M}\operatorname{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\right)+C_3,\end{aligned}$$

where C_2 and C_3 are constants.

The term $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{Y}|\boldsymbol{z}\right)\right]$ can be calculated as

$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{Y}|\boldsymbol{z}\right)\right] \qquad (24)\\ =\;& \sum_{m=1}^{M}\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\log p\left(\boldsymbol{y}_m|\boldsymbol{z}_0,\boldsymbol{z}_m\right)\right]\\ =\;& \sum_{m=1}^{M}\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[-\frac{\left[\boldsymbol{y}_m-\boldsymbol{A}(\boldsymbol{z}_0)\boldsymbol{z}_m\right]^H\left[\boldsymbol{y}_m-\boldsymbol{A}(\boldsymbol{z}_0)\boldsymbol{z}_m\right]}{\sigma^2}\right]\\ &+C_4,\\ =\;& \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[-\frac{\|\boldsymbol{Y}-\boldsymbol{A}(\boldsymbol{z}_{\mathbf{0}})\cdot[\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M]\|_F^2}{\sigma^2}\right]+C_4\\ =\;& \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[-\frac{1}{\sigma^2}\left\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\right\|_F^2\right]+C_4,\end{aligned}$$

where C_4 is a constant, and the last equality comes from (4).

Combining (22)–(24), we obtain the loss function in (7).

Appendix C

To get the optimal estimates of channel gains, we calculate the partial derivative of the loss function (7) with respect to the estimates $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$ of channel gains for $m=1,\ldots,M$.

We consider only the terms affected by the estimates $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$, namely $(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m})^H\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m})$ and $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{Y})}\left[\frac{1}{\sigma^2}\|\boldsymbol{y}_m-\widehat{\boldsymbol{y}}_m\|_2^2\right]$. We look at the MSE term first.

The received signal over mth snapshot can be written as

$$\boldsymbol{y}_m=\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{h}_m+\boldsymbol{n}_m\sim\mathcal{CN}\left(\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{h}_m,\sigma^2\boldsymbol{I}_N\right). \tag{25}$$

The reconstructed signal can be written as $\widehat{\boldsymbol{y}}_m=\boldsymbol{A}(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{h}}_m$ with $\widehat{\boldsymbol{h}}_m\sim\mathcal{CN}\left(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m},\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\right)$.

When the AoA estimate is accurate, (i.e., $\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}$), we write $\boldsymbol{A}=\boldsymbol{A}(\boldsymbol{\theta})=\boldsymbol{A}(\widehat{\boldsymbol{\theta}})$ for notational simplicity, and have

$$\widehat{\boldsymbol{y}}_m\sim\mathcal{CN}\left(\boldsymbol{A}\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m},\boldsymbol{A}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\boldsymbol{A}^H\right). \tag{26}$$

Since the noise $\mathbf{n}_m$ and the channel gains $\widehat{\boldsymbol{h}}_m$ are independent, $\boldsymbol{y}_m$ and $\widehat{\boldsymbol{y}}_m$ are independent. Therefore, we have

$$\boldsymbol{y}_m-\widehat{\boldsymbol{y}}_m\sim\mathcal{CN}\left(\boldsymbol{A}(\boldsymbol{h}_m-\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}),\sigma^2\boldsymbol{I}_N+\boldsymbol{A}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\boldsymbol{A}^H\right).$$

As a result, the MSE term can be calculated as

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{y}_m-\widehat{\boldsymbol{y}}_m\|_2^2\right] &= (\boldsymbol{h}_m-\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})^H\boldsymbol{A}^H\boldsymbol{A}(\boldsymbol{h}_m-\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})\\ &\quad+\sigma^2N+\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\boldsymbol{A}^H\right).\end{aligned} \tag{27}$$

We collect the terms that depend on the estimate $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$ into

$$\begin{aligned}f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}) &= (\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m})^H\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m})\\ &\quad+\frac{1}{\sigma^2}(\boldsymbol{h}_m-\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})^H\boldsymbol{A}^H\boldsymbol{A}(\boldsymbol{h}_m-\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}),\end{aligned} \tag{28}$$

which is a concave quadratic function of the estimate $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$. Therefore, the estimate $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}$ minimizes $f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})$ if and only if the derivatives with respect to the real part $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^R$ and the imaginary part $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^I$ are zero. According to the $\mathbb{CR}$-calculus [57], the derivatives can be calculated as

$$\begin{aligned}\frac{\partial f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})}{\partial\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^R} &= \frac{\partial f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m},\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^*)}{\partial\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}}+\frac{\partial f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m},\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^*)}{\partial\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^*} \qquad (29)\\ &= 2\cdot\mathfrak{Re}\left\{\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m})-\frac{1}{\sigma^2}\boldsymbol{A}^H\boldsymbol{A}(\boldsymbol{h}_m-\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})\right\}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})}{\partial\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^I} &= j\cdot\left[\frac{\partial f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m},\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^*)}{\partial\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}}-\frac{\partial f(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m},\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^*)}{\partial\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^*}\right] \qquad (30)\\ &= 2\cdot\mathfrak{Im}\left\{\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}(\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}-\boldsymbol{\mu}_{\boldsymbol{h}_m})-\frac{1}{\sigma^2}\boldsymbol{A}^H\boldsymbol{A}(\boldsymbol{h}_m-\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})\right\},\end{aligned}$$

where $\mathfrak{Re}\left\{\cdot\right\}$ and $\mathfrak{Im}\left\{\cdot\right\}$ are the real and imaginary parts.

Combining (29) and (30), we know that the optima estimate should satisfy

$$\left(\boldsymbol{A}^H\boldsymbol{A}+\sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\right)\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}=\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{h}_m+\sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\boldsymbol{\mu}_{\boldsymbol{h}_m}. \tag{31}$$

Since the matrix $\left(\boldsymbol{A}^H\boldsymbol{A}+\sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\right)$ is positive definitive and hence invertible, we have

$$\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}=\left(\boldsymbol{A}^H\boldsymbol{A}+\sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\right)^{-1}\left(\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{h}_m+\sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}_m}^{-1}\boldsymbol{\mu}_{\boldsymbol{h}_m}\right). \tag{32}$$

Appendix D

Under the assumption that the channel estimation is accurate (i.e., $\widehat{\boldsymbol{h}}_m=\boldsymbol{h}_m$ for $m=1,\ldots,M$, one type of globally optimal AoA estimates $\widehat{\boldsymbol{\theta}}$ are the ones that yield the same phase shifts of the received signals as the true AoAs $\boldsymbol{\theta}$ do. Mathematically, this condition is that for all $n=1,\ldots,N$, there exists an integer $l_n\in\mathbb{Z}$ such that

$$\frac{2\pi(n-1)d}{\lambda}(\sin\theta_k-\sin\widehat{\theta}_k)=2\pi l_n. \tag{33}$$

From (33), we have $l_1 = 0$, $l_2 = \frac{d}{\lambda}(\sin\theta_k - \sin\widehat{\theta}_k)$, and $l_n = (n-1)l_2$ for $n = 3, 4, \ldots, N$. So it boils down to finding $\widehat{\theta}_k$ such that $\frac{d}{\lambda}(\sin\theta_k - \sin\widehat{\theta}_k)$ is an integer.

Since $\sin\theta_k - \sin\widehat{\theta}_k \in [\sin\theta_k - 1, \sin\theta_k + 1]$, we have

$$\left\lceil \frac{d}{\lambda}\left(\sin\theta_k - 1\right) \right\rceil \le l_2 \le \left\lfloor \frac{d}{\lambda}\left(\sin\theta_k + 1\right) \right\rfloor, \tag{34}$$

where $\lceil x \rceil$ is the minimum integer no smaller than x and $\lfloor x \rfloor$ is the maximum integer no larger than x.

Therefore, the solutions to (33) are

$$\widehat{\theta}_k = \arcsin\left(\sin\theta_k - l_2 \cdot \frac{\lambda}{d}\right) \tag{35}$$

for integer l_2 that satisfies (34).

Appendix E

A. Calculating the Gradient

To study the stationary points, we first calculate the gradient of the loss function with respect to the AoA estimates $\widehat{\boldsymbol{\theta}}$. Note that the AoA estimates only affect the MSE term $\mathbb{E}\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_F^2$ in the loss function (7).

Since the received signal in the mth snapshot is

$$\boldsymbol{y}_m = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{h}_m + \boldsymbol{n}_m \sim \mathcal{CN}\left(\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{h}_m, \sigma^2 \boldsymbol{I}_N\right), \tag{36}$$

and since the reconstructed signal is

$$\widehat{\boldsymbol{y}}_m = \boldsymbol{A}(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{h}}_m \sim \mathcal{CN}\left(\boldsymbol{A}(\widehat{\boldsymbol{\theta}})\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}, \boldsymbol{A}(\widehat{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\boldsymbol{A}(\widehat{\boldsymbol{\theta}})^H\right), \tag{37}$$

we have

$$\boldsymbol{y}_m - \widehat{\boldsymbol{y}}_m \sim \mathcal{CN}\left(\boldsymbol{A}\boldsymbol{h}_m - \widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}, \sigma^2\boldsymbol{I}_N + \widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\widehat{\boldsymbol{A}}^H\right),$$

where we define $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{A}} = \boldsymbol{A}(\widehat{\boldsymbol{\theta}})$ for simplicity.

Thus, the MSE can be calculated as

$$\begin{aligned} \mathbb{E}\left[\|\boldsymbol{y}_m - \widehat{\boldsymbol{y}}_m\|_2^2\right] &= (\boldsymbol{A}\boldsymbol{h}_m - \widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m})^H(\boldsymbol{A}\boldsymbol{h}_m - \widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}) \\ &+ \sigma^2 N + \operatorname{tr}\left(\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\widehat{\boldsymbol{A}}^H\right). \end{aligned} \tag{38}$$

Since the AoA estimates $\widehat{\boldsymbol{\theta}}$ affect $\widehat{\boldsymbol{A}}$ only, we collect the terms in (38) that depend on $\widehat{\boldsymbol{A}}$ as follows:

$$\begin{aligned} f_m(\widehat{\boldsymbol{A}}) &\triangleq \underbrace{\left(-\boldsymbol{h}_m^H\boldsymbol{A}^H\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m} - \boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^H\widehat{\boldsymbol{A}}^H\boldsymbol{A}\boldsymbol{h}_m\right)}_{\triangleq f_{m,1}(\widehat{\boldsymbol{A}})} \\ &+ \underbrace{\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}^H\widehat{\boldsymbol{A}}^H\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m}}_{\triangleq f_{m,2}(\widehat{\boldsymbol{A}})} + \underbrace{\operatorname{tr}\left(\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{h}}_m}\widehat{\boldsymbol{A}}^H\right)}_{\triangleq f_{m,3}(\widehat{\boldsymbol{A}})} \end{aligned} \tag{39}$$

Note that the matrix $\widehat{\boldsymbol{A}} = [\widehat{\boldsymbol{a}}_1, \ldots, \widehat{\boldsymbol{a}}_K]$ has K columns, where the k-th column $\widehat{\boldsymbol{a}}_k$ is a function of user k's AoA estimate $\widehat{\theta}_k$, namely $\widehat{\boldsymbol{a}}_k = e^{-j\boldsymbol{\Psi}(\widehat{\theta}_k)}$.

Next, we calculate the partial derivative of the function $f(\widehat{\boldsymbol{A}})$ in (39) with respect to user k's AoA estimate $\widehat{\theta}_k$ using the $\mathbb{CR}$-calculus [57]:

$$\frac{\partial f(\widehat{\boldsymbol{A}})}{\partial \widehat{\theta}_k} = \left[\frac{\partial f(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k}\right]^T \cdot \frac{\partial \widehat{\boldsymbol{a}}_k}{\partial \widehat{\theta}_k} + \left[\frac{\partial f(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k^*}\right]^T \cdot \frac{\partial \widehat{\boldsymbol{a}}_k^*}{\partial \widehat{\theta}_k}. \tag{40}$$

To calculate $\frac{\partial f(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k}$ and $\frac{\partial f(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k^*}$, it is useful to note that

$$\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m} = \sum_{k=1}^{K} \mu_{\widehat{\mathbf{h}}_m,k}\widehat{\boldsymbol{a}}_k \text{ and } \widehat{\boldsymbol{A}}^*\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}^* = \sum_{k=1}^{K} \mu_{\widehat{\mathbf{h}}_m,k}^*\widehat{\boldsymbol{a}}_k^*. \tag{41}$$

For the first term in (39), we have

$$\frac{\partial f_{m,1}(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k} = \frac{\partial\left(-\boldsymbol{h}_m^H\boldsymbol{A}^H\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}\right)}{\partial \widehat{\boldsymbol{a}}_k} = -\mu_{\widehat{\mathbf{h}}_m,k}\boldsymbol{A}^*\boldsymbol{h}_m^*, \tag{42}$$

and

$$\frac{\partial f_{m,1}(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k^*} = \frac{\partial\left(-\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}^H\widehat{\boldsymbol{A}}^H\boldsymbol{A}\boldsymbol{h}_m\right)}{\partial \widehat{\boldsymbol{a}}_k^*} = -\mu_{\widehat{\mathbf{h}}_m,k}^*\boldsymbol{A}\boldsymbol{h}_m. \tag{43}$$

For the second term in (39), we have

$$\frac{\partial f_{m,2}(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k} = \frac{\partial\left(\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}^H\widehat{\boldsymbol{A}}^H\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}\right)}{\partial \widehat{\boldsymbol{a}}_k} = \mu_{\widehat{\mathbf{h}}_m,k}\widehat{\boldsymbol{A}}^*\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}^*, \tag{44}$$

and

$$\frac{\partial f_{m,2}(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k^*} = \frac{\partial\left(\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}^H\widehat{\boldsymbol{A}}^H\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}\right)}{\partial \widehat{\boldsymbol{a}}_k^*} = \mu_{\widehat{\mathbf{h}}_m,k}^*\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}. \tag{45}$$

For the third term in (39), we have

$$\frac{\partial f_{m,3}(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k} = \frac{\partial \operatorname{tr}\left(\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_{\widehat{\mathbf{h}}_m}\widehat{\boldsymbol{A}}^H\right)}{\partial \widehat{\boldsymbol{a}}_k} = \left(\widehat{\boldsymbol{A}}^*\boldsymbol{\Sigma}_{\widehat{\mathbf{h}}_m}^*\right)_{:,k}, \tag{46}$$

and

$$\frac{\partial f_{m,3}(\widehat{\boldsymbol{A}})}{\partial \widehat{\boldsymbol{a}}_k^*} = \frac{\partial \operatorname{tr}\left(\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_{\widehat{\mathbf{h}}_m}\widehat{\boldsymbol{A}}^H\right)}{\partial \widehat{\boldsymbol{a}}_k^*} = \left(\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_{\widehat{\mathbf{h}}_m}\right)_{:,k}, \tag{47}$$

where $(\cdot)_{:,k}$ is the k-th column of a matrix.

In addition, we have

$$\frac{\partial \widehat{\boldsymbol{a}}_k}{\partial \widehat{\theta}_k} = -j \cdot \widehat{\boldsymbol{a}}_k \odot \boldsymbol{\Phi}(\widehat{\theta}_k) \text{ and } \frac{\partial \widehat{\boldsymbol{a}}_k^*}{\partial \widehat{\theta}_k} = j \cdot \widehat{\boldsymbol{a}}_k^* \odot \boldsymbol{\Phi}(\widehat{\theta}_k), \tag{48}$$

where $\boldsymbol{\Phi}(\widehat{\theta}_k) = \frac{2\pi d\cos(\widehat{\theta}_k)}{\lambda}[0, 1, \ldots, N-1]^T$.

Combining (42)–(48), we have

$$\begin{aligned} &\frac{\partial f_m(\widehat{\boldsymbol{A}})}{\partial \widehat{\theta}_k} \\ &= \left[-\mu_{\widehat{\mathbf{h}}_m,k}\boldsymbol{A}^*\boldsymbol{h}_m^* + \mu_{\widehat{\mathbf{h}}_m,k}\widehat{\boldsymbol{A}}^*\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}^* + \left(\widehat{\boldsymbol{A}}^*\boldsymbol{\Sigma}_{\widehat{\mathbf{h}}_m}^*\right)_{:,k}\right]^T \\ &\quad \cdot \left[-j \cdot \widehat{\boldsymbol{a}}_k \odot \boldsymbol{\Phi}(\widehat{\theta}_k)\right] \\ &+ \left[-\mu_{\widehat{\mathbf{h}}_m,k}^*\boldsymbol{A}\boldsymbol{h}_m + \mu_{\widehat{\mathbf{h}}_m,k}^*\widehat{\boldsymbol{A}}\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m} + \left(\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_{\widehat{\mathbf{h}}_m}\right)_{:,k}\right]^T \\ &\quad \cdot \left[j \cdot \widehat{\boldsymbol{a}}_k^* \odot \boldsymbol{\Phi}(\widehat{\theta}_k)\right] \\ &= \mathfrak{Im}\left\{\left[\boldsymbol{J}_m(\widehat{\boldsymbol{A}})\right]^T \cdot \left[\widehat{\boldsymbol{a}}_k \odot \boldsymbol{\Phi}(\widehat{\theta}_k)\right]\right\}, \end{aligned} \tag{49}$$

where

$$\boldsymbol{J}_m(\widehat{\boldsymbol{A}}) \triangleq \mu_{\widehat{\mathbf{h}}_m,k}\left(\widehat{\boldsymbol{A}}^*\boldsymbol{\mu}_{\widehat{\mathbf{h}}_m}^* - \boldsymbol{A}^*\boldsymbol{h}_m^*\right) + \left(\widehat{\boldsymbol{A}}^*\boldsymbol{\Sigma}_{\widehat{\mathbf{h}}_m}^*\right)_{:,k}. \tag{50}$$

Under the assumption that the channel estimation is accurate (i.e., $\boldsymbol{\mu}_{\widehat{\boldsymbol{h}}_m} = \boldsymbol{h}_m$ for $m = 1, \ldots, M$), the derivative can be finally written as

$$\frac{\partial \mathcal{L}^{(train)}}{\partial \widehat{\theta}_k} = \sum_{m=1}^{M} \frac{\partial f_m(\widehat{\boldsymbol{A}})}{\partial \widehat{\theta}_k} \tag{51}$$

$$= \sum_{m=1}^{M} \mathfrak{Im}\left\{ \boldsymbol{J}_m(\widehat{\boldsymbol{A}})^T \cdot \left[\widehat{\boldsymbol{a}}_k \odot \boldsymbol{\Phi}(\widehat{\theta}_k) \right] \right\}$$

$$= \sum_{m=1}^{M} |h_{m,k}|^2 \sum_{n=0}^{N-1} \frac{2\pi d n}{\lambda} \cos \widehat{\theta}_k \left[\sin\left(\eta_{\widehat{\theta}_k} n\right) - \sin\left(\zeta_{\widehat{\theta}_k} n\right) \right],$$

where $\eta_{\widehat{\theta}_k} = \frac{2\pi d}{\lambda}(\sin\theta_k + \sin\widehat{\theta}_k)$ and $\zeta_{\widehat{\theta}_k} = \frac{2\pi d}{\lambda} 2 \sin\widehat{\theta}_k$.

B. Analyzing the stationary points

In the gradient (51), the terms $\sum_{m=1}^{M} |h_{m,k}|^2$ and $\frac{2\pi d}{\lambda}$ are always positive. Therefore, we focus on

$$g(\widehat{\theta}_k) \triangleq \cos\widehat{\theta}_k \cdot \sum_{n=0}^{N-1} n \left[\sin\left(\eta_{\widehat{\theta}_k} n\right) - \sin\left(\zeta_{\widehat{\theta}_k} n\right) \right]. \tag{52}$$

The stationary points are the solutions to $g(\widehat{\theta}_k) = 0$.

Since $g(\widehat{\theta}_k)$ is a summation of N terms, we aim to find its limit when $N \to \infty$. Specifically, define

$$t(x) \triangleq x \left[\sin\left(\eta_{\widehat{\theta}_k}(N-1)x \right) - \sin\left(\zeta_{\widehat{\theta}_k}(N-1)x \right) \right]. \tag{53}$$

Note that the relationship between $g(\widehat{\theta}_k)$ and $t(x)$ is

$$g(\widehat{\theta}_k) = \cos\widehat{\theta}_k \cdot (N-1) \cdot \sum_{n=0}^{N-1} t\left(\frac{n}{N-1} \right). \tag{54}$$

Meanwhile, we have

$$\int_0^1 t(x)dx = \lim_{N\to\infty} \frac{1}{N-1} \sum_{n=0}^{N-1} t\left(\frac{n}{N-1} \right). \tag{55}$$

Note that the integral can be calculated analytically as

$$\int_0^1 t(x)dx = \frac{\cos\left(\zeta_{\widehat{\theta}_k}(N-1) \right)}{\zeta_{\widehat{\theta}_k}} - \frac{\cos\left(\eta_{\widehat{\theta}_k}(N-1) \right)}{\eta_{\widehat{\theta}_k}}. \tag{56}$$

Finally, we have

$$\lim_{N\to\infty} g(\widehat{\theta}_k) = 0 \tag{57}$$

$$\Leftrightarrow \lim_{N\to\infty} \cos\widehat{\theta}_k \cdot \sum_{n=0}^{N-1} t\left(\frac{n}{N-1} \right) = 0$$

$$\Leftrightarrow \cos\widehat{\theta}_k \cdot \int_0^1 t(x)dx = 0$$

$$\Leftrightarrow \cos\widehat{\theta}_k \cdot \left[\frac{\cos\left(\zeta_{\widehat{\theta}_k}(N-1) \right)}{\zeta_{\widehat{\theta}_k}} - \frac{\cos\left(\eta_{\widehat{\theta}_k}(N-1) \right)}{\eta_{\widehat{\theta}_k}} \right] = 0,$$

which concludes the proof.

REFERENCES

[1] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[2] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. D. Doncker, “The COST 2100 MIMO channel model,” *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, 2012.

[3] J. Li, Q. Zhang, X. Xin, Y. Tao, Q. Tian, F. Tian, D. Chen, Y. Shen, G. Cao, Z. Gao, and J. Qian, “Deep learning-based massive MIMO CSI feedback,” in *2019 18th International Conference on Optical Communications and Networks (ICOCN)*, 2019, pp. 1–3.

[4] P. Stoica and A. Nehorai, “MUSIC, maximum likelihood and cramer-rao bound: further results and comparisons,” in *International Conference on Acoustics, Speech, and Signal Processing,*, 1989, pp. 2605–2608 vol.4.

[5] R. Roy and T. Kailath, “ESPRIT – estimation of signal parameters via rotational invariance techniques,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[6] S. Marcos, A. Marsal, and M. Benidir, “The propagator method for source bearing estimation,” vol. 42, no. 2, 1995.

[7] P. Pal and P. P. Vaidyanathan, “Nested arrays: A novel approach to array processing with enhanced degrees of freedom,” *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4167–4181, 2010.

[8] X. Dai, X. Zhang, and Y. Wang, “Extended DOA-matrix method for DOA estimation via two parallel linear arrays,” *IEEE Communications Letters*, vol. 23, no. 11, pp. 1981–1984, 2019.

[9] B. Li, S. Wang, J. Zhang, X. Cao, and C. Zhao, “Fast randomized-MUSIC for mm-Wave massive MIMO radars,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1952–1956, 2021.

[10] Z. Yang and L. Xie, “Enhancing sparsity and resolution via reweighted atomic norm minimization,” *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 995–1006, 2015.

[11] X. Xu, X. Rao, and V. K. Lau, “Active user detection and channel estimation in uplink CRAN systems,” in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 2727–2732.

[12] Z. Yang and L. Xie, “Exact joint sparse frequency recovery via optimization methods,” *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5145–5157, 2016.

[13] H.-H. Chao and L. Vandenberghe, “Semidefinite representations of gauge functions for structured low-rank matrix decomposition,” *SIAM Journal on Optimization*, vol. 27, no. 3, pp. 1362–1389, 2017.

[14] F. Roos, P. Hügler, L. L. T. Torres, C. Knill, J. Schlichenmaier, C. Vasanelli, and et al, “Compressed sensing based single snapshot DoA estimation for sparse MIMO radar arrays,” in *2019 12th German Microwave Conference (GeMiC)*, March 2019, pp. 75–78.

[15] H. Liu, X. Yuan, and Y. J. Zhang, “Super-resolution blind channel-and-signal estimation for massive MIMO with one-dimensional antenna array,” *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4433–4448, 2019.

[16] P. Chen, Z. Chen, Z. Cao, and X. Wang, “A new atomic norm for DOA estimation with gain-phase errors,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 4293–4306, 2020.

[17] Y. Wang, Y. Zhang, Z. Tian, G. Leus, and G. Zhang, “Super-resolution channel estimation for arbitrary arrays in hybrid millimeter-wave massive MIMO systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 947–960, 2019.

[18] M. Wagner, Y. Park, and P. Gerstoft, “Gridless DOA estimation and root-MUSIC for non-uniform linear arrays,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 2144–2157, 2021.

[19] A. Garcia-Rodriguez and C. Masouros, “Low-complexity compressive sensing detection for spatial modulation in large-scale multiple access channels,” *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2565–2579, 2015.

[20] Y. Peng, Y. Li, and P. Wang, “An enhanced channel estimation method for millimeter wave systems with massive antenna arrays,” *IEEE Communications Letters*, vol. 19, no. 9, pp. 1592–1595, 2015.

[21] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, 2016.

[22] Z. Gao, L. Dai, W. Dai, B. Shim, and Z. Wang, “Structured compressive sensing-based spatio-temporal joint channel estimation for FDD massive MIMO,” *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 601–617, 2016.

[23] Z. Zhou, J. Fang, L. Yang, H. Li, Z. Chen, and R. S. Blum, “Low-rank tensor decomposition-aided channel estimation for millimeter wave

MIMO-OFDM systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1524–1538, 2017.

[24] Z. Gao, L. Dai, S. Han, C.-L. I, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 144–153, 2018.

[25] S. Park and R. W. Heath, "Spatial channel covariance estimation for the hybrid MIMO architecture: A compressive sensing-based approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8047–8062, 2018.

[26] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Transactions on Signal Processing*, vol. 68, pp. 764–779, 2020.

[27] M.-A. Badiu, T. L. Hansen, and B. H. Fleury, "Variational Bayesian inference of line spectra," *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2247–2261, 2017.

[28] T. L. Hansen, B. H. Fleury, and B. D. Rao, "Superfast line spectral estimation," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2511–2526, 2018.

[29] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 93–99, 2019.

[30] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, 2019.

[31] Z.-M. Liu, C. Zhang, and S. Y. Philip, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, pp. 7315–7327, 2018.

[32] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 852–855, 2018.

[33] H. Huang, J. Yang, H. Huang, Y. Song, and G. Gui, "Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8549–8560, 2018.

[34] G. Izacard, S. Mohan, and C. Fernandez-Granda, "Data-driven estimation of sinusoid frequencies," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 5127–5137.

[35] L. Wu, Z.-M. Liu, and Z.-T. Huang, "Deep convolution network for direction of arrival estimation with sparse prior," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1688–1692, 2019.

[36] Y. Zhang, Y. Mu, Y. Liu, T. Zhang, and Y. Qian, "Deep learning-based beamspace channel estimation in mmWave massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 9, no. 12, pp. 2212–2215, 2020.

[37] L. Wan, Y. Sun, L. Sun, Z. Ning, and J. J. Rodrigues, "Deep learning based autonomous vehicle super resolution DOA estimation for safety driving," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[38] Y. Guo, Z. Zhang, Y. Huang, and P. Zhang, "DOA estimation method based on cascaded neural network for two closely spaced sources," *IEEE Signal Processing Letters*, vol. 27, pp. 570–574, 2020.

[39] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3133–3143, 2020.

[40] R. Akter, V.-S. Doan, T. Huynh-The, and D.-S. Kim, "RFDOA-Net: An efficient ConvNet for RF-based DOA estimation in UAV surveillance systems," *IEEE Transactions on Vehicular Technology*, 2021.

[41] G. K. Papageorgiou, M. Sellathurai, and Y. C. Eldar, "Deep networks for direction-of-arrival estimation in low SNR," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3714–3729, 2021.

[42] A. Barthelme and W. Utschick, "A machine learning approach to DoA estimation and model order selection for antenna arrays with subarray sampling," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3075–3087, 2021.

[43] P. Pan, Y. Zhang, Z. Deng, and W. Qi, "Deep learning-based 2-D frequency estimation of multiple sinusoidals," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[44] P. Pan, Y. Zhang, Z. Deng, and G. Wu, "Complex-valued frequency estimation network and its applications to superresolution of radar range profiles," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[45] X. Su, P. Hu, Z. Liu, T. Liu, B. Peng, and X. Li, "Mixed near-field and far-field source localization based on convolution neural networks via symmetric nested array," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 7908–7920, 2021.

[46] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "ComNet: Combination of deep learning and expert knowledge in OFDM receivers," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2627–2630, 2018.

[47] J. Liao, J. Zhao, F. Gao, and G. Y. Li, "A model-driven deep learning method for massive MIMO detection," *IEEE Communications Letters*, vol. 24, no. 8, pp. 1724–1728, 2020.

[48] A. Barthelme and W. Utschick, "DoA estimation using neural network-based covariance matrix reconstruction," *IEEE Signal Processing Letters*, vol. 28, pp. 783–787, 2021.

[49] A. M. Elbir, "DeepMUSIC: Multiple signal classification via deep learning," *IEEE Sensors Letters*, vol. 4, no. 4, pp. 1–4, 2020.

[50] Y. Yuan, S. Wu, M. Wu, and N. Yuan, "Unsupervised learning strategy for direction-of-arrival estimation network," *IEEE Signal Processing Letters*, vol. 28, pp. 1450–1454, 2021.

[51] Y. Yang, N. Zou, E. Lin, F. Suo, and Z. Chen, "A neural network method for nonconvex optimization and its application on parameter retrieval," *IEEE Transactions on Signal Processing*, 2021.

[52] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ica: A unifying framework," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2207–2217.

[53] S. Y. Lee, "Gibbs sampler and coordinate ascent variational inference: A set-theoretical review," *Communications in Statistics-Theory and Methods*, pp. 1–21, 2021.

[54] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[55] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[56] Q. He, L. Xiao, X. Zhong, and S. Zhou, "Increasing the sum-throughput of cells with a sectorization method for massive MIMO," *IEEE Communications Letters*, vol. 18, no. 10, pp. 1827–1830, 2014.

[57] K. Kreutz-Delgado, "The complex gradient operator and the cr-calculus," *arXiv preprint arXiv:0906.4835*, 2009.