Unsupervised Channel Estimation with Knowledge-Aware Variational Auto-Encoder

Zhiheng Guo*, Yuanzhang Xiao† and Xiang Chen*‡
*School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China
guozhh7@mail2.sysu.edu.cn; chenxiang@mail.sysu.edu.cn
†Hawaii Advanced Wireless Technologies Institute, University of Hawaii, Honolulu, HI
yxiao8@hawaii.edu

***Abstract*—This paper proposes an unsupervised channel estimation method for massive MIMO systems. Our method builds on a novel improvement of the traditional Variational Auto-Encoder (VAE) by incorporating the knowledge of signal propagation models into the decoder. Specifically, instead of having learnable parameters, the decoder in our proposed knowledge-aware VAE has fixed weights that implement the signal propagation model. Such modification forces the encoder of our proposed VAE to learn meaningful parameters (i.e., angle-of-arrivals, path gains and path angles), as opposed to standard VAE whose encoder has no control of the physical meaning of its output. We rigorously analyze the multiplicity of global optima in unsupervised channel estimation problems. Based on our analysis, we adopt the sectorization method to alleviate the issue of multiple global optima. Numerical simulation results corroborate our analysis, and demonstrate the performance improvement of our proposed method over existing representative channel estimation methods.**

I. Introduction

The popularity of fifth generation (5G) communication networks comes with many challenges, such as explosive data traffic growth, high quality of service (QoS) requirement, and increased energy efficiency requirement. Massive multiple-input-multiple-output (MIMO) systems is one of the vital technologies to meet 5G wireless communication system requirements. Theoretically, massive MIMO has been proven to be able to enhance the capacity of a communication system with additional antennas [1]. However, there are still difficulties in putting massive MIMO technology into practical use, resulting from the sophisticated channel modeling [2], the high cost of channel state information (CSI) [3], the huge amount of calculation due to the super-large matrix generated by massive MIMO, etc.. These difficulties will degrade the performance of the massive MIMO system. As a result, we urgently need some effective schemes to overcome these difficulties.

In this paper, we focus on the channel estimation problem in massive MIMO systems. There are three major ways of channel estimation: estimation from the known CSI, estimation based on compressed sensing (CS) [4] and deep learning-based estimation [5].

Classical algorithms such as ESPRIT [6] and MUSIC [7] perform angle-of-arrival (AoA) estimation based on the known CSI (i.e., the covariance matrix of the received signal). The ESPRIT algorithm requires the rotation invariance of the signal subspace of the covariance matrix while the MUSIC algorithm requires the orthogonality between the signal subspace and the noise subspace. However, both techniques will suffer from a huge amount of calculation due to the large size of the covariance matrix under the large amount of antennas in massive MIMO systems.

To avoid huge amount of calculation, CS-based estimation is introduced to channel estimation and AoA estimation. By exploiting the potential sparsity of the massive MIMO covariance matrix in the certain transform domain, the CSI transmission matrix can be effectively compressed so that computational complexity can be greatly reduced. However, most CS-based channel estimation methods have a strong assumption that the channel need to have strong sparsity. Thus it can hardly work well when the channel has correlated path likes Rayleigh channel. Furthermore, CS-based estimation is always iterative which brings additional delay and is not suitable for real-time estimation.

In recent years, with the rapid development of deep learning (DL), there is a growing literature of applying DL to channel estimation. All these DL-based channel estimation methods are model-free, in the sense that they do not have strong assumptions on channel characteristics and have low computational complexity in the deployment phase (i.e., after the model is trained). In [8], by leveraging the spatial structure, they integrated DL technology into the massive MIMO system and first proposed the use of DL to achieve AoA estimation and channel estimation based on deep neural network (DNN). However, to the best of our knowledge, all the existing DL-based channel estimation algorithms are *supervised*, i.e., they need the labels of the channel features (e.g., AoAs, path gains, path angles) to be estimated. We aim to develop a unsupervised DL-based channel estimation that achieves similar performance as the supervised methods.

There is a seemingly unbreakable hurdle in developing unsupervised channel estimation methods. Specifically, without the labels, the output of the neural networks usually has no physical meaning and therefore will not be the channel features we want to estimate. This is because without the labels, the network can only take the difference between the input (i.e., the received signal) and the output (i.e., the reconstructed signal from the estimated features) as its loss function. As a result, the features estimated from the network will only be a lower-dimensional embedding of the received signal, which may not be the actual channel features with physical meanings. A related issue, resulting from the lack of labels, is that there can be multiple

‡ Xiang Chen is the corresponding author.

sets of channel features that result in the same received signal. These channel features are all global optima of the unsupervised estimation problem. Such multiplicity of global optima makes it hard for the supervised method to make accurate estimation.

To overcome this hurdle, we can utilize our knowledge of the wireless channel model and the signal propagation model. In particular, we propose a novel variation of the classic unsupervised DL method, Variational Auto-Encoder (VAE) [9] [10]. The main idea is to use the output of its encoder as the channel features extracted from the network input (i.e., the received signal), reconstruct the signal with the decoder, and compare the output of the decoder with the network input to evaluate the estimation accuracy. Our key innovation is how to enforce the output of the encoder to be the channel features to be estimated. To achieve this purpose, we propose Knowledge-Aware Variational Auto-Encoder (KA-VAE), which instills the knowledge of the signal propagation model into the decoder by hardwiring it with the signal propagation model.

In summary, this paper proposes the first unsupervised DL-based channel estimation in massive MIMO systems. We analyzed the issue of multiple global optima in the unsupervised channel estimation problem, and adopted the sectorization method to alleviate the issue. Numerical simulations were carried out to validate our theoretical analysis. Comparisons with carefully designed benchmark algorithms demonstrated that our proposed method achieves almost identical performance as the fully supervised method, and that the achievement is enabled by our novel design of the decoder.

II. System Model

In this paper, a typical massive MIMO uplink system is considered in which there is one base station (BS) with a uniform linear array (ULA) of N_t antennas and M users. The received signal at the BS can be represented as follows:

$$\boldsymbol{y} = \sum_{i=1}^{M} \alpha_i \boldsymbol{a}(\theta_i) x_i + \boldsymbol{n}. \tag{1}$$

where α_i is the complex Gaussian distributed channel gain from user i to the BS, $\theta_i \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ is the angle of arrival (AoA) of user i's signal, x_i is user i's transmit signal, $\boldsymbol{n} \sim \mathcal{CN}(0, \sigma_n^2 \boldsymbol{I}_{N_t}) \in \mathbb{C}^{N_t \times 1}$ is the complex additive white Gaussian noise, and $\boldsymbol{a}(\theta_i)$ is the steering vector defined as :

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

where λ represents the wavelength of the received signal and $d \geq \frac{\lambda}{2}$ represents the antenna spacing. We represent the complex amplitude α_i in the polar form as $\alpha_i = |\alpha_i| e^{j\phi_i}$, where $|\alpha_i|$ is the path gain, which follows the Rayleigh distribution, and ϕ_i is the path angle, which follows the uniform distribution in $[-\pi, \pi]$. Without loss of generality, we assume the transmit signal x in (1) to have unit magnitude. Then we decompose the received signal into real and imaginary parts $\boldsymbol{y} = \boldsymbol{y}_R + j \cdot \boldsymbol{y}_I$ and write each part analytically as

$$\begin{cases} \boldsymbol{y}_R = \sum_{i=1}^{M} |\alpha_i| \cos\left(\phi_i - \boldsymbol{\Psi}(\theta_i)\right) + \boldsymbol{n}_R \\ \boldsymbol{y}_I = \sum_{i=1}^{M} |\alpha_i| \sin\left(\phi_i - \boldsymbol{\Psi}(\theta_i)\right) + \boldsymbol{n}_I \end{cases}, \tag{3}$$

where $\Psi(\theta_i) = [1, 2\pi\frac{d}{\lambda}\sin(\theta_i), 2\pi\frac{d}{\lambda}2\sin(\theta_i), \ldots, 2\pi\frac{d}{\lambda}(N_t - 1)\sin(\theta_i)]^T$, and $\boldsymbol{n}_R$ and $\boldsymbol{n}_I$ are the real and imaginary parts of the noise, respectively.

Our goal is to estimate the AoAs θ_i as well as the path gains α_i and path angles ϕ_i based on the received signal $\boldsymbol{y}$.

III. Proposed Solution

We propose an unsupervised channel estimation method by modifying the VAE, which is illustrated in Fig. 1. Our proposed neural network consists of an encoder and an decoder. The encoder takes the received signal $\boldsymbol{y}_R$ and $\boldsymbol{y}_I$ as input, and outputs the estimated parameters of the distributions of AoAs θ_i, path gains α_i, and path angles ϕ_i. Then the decoder tries to reconstruct the received signal based on the output of the encoder. The key difference from the traditional VAE is that *our decoder is fixed*, as opposed to have learnable parameters. This is because we know the signal model as in (1), and thus utilize such knowledge by implementing the signal model with the decoder. Next, we give a detailed description of our proposed KA-VAE.

As shown in Fig. 1, the real and imaginary parts of the received signal $\boldsymbol{y}_R$ and $\boldsymbol{y}_I$, called the target signal hereafter, are the $2N_t$-dimensional input of the encoder. The encoder is a multilayer perceptron with the rectified linear unit (ReLU) activation function. From Section II, we know that AoAs θ_i are uniform random variables and α_i are complex Gaussian random variables. Therefore, the real and imaginary parts of α_i, denoted $\alpha_{i,R}$ and $\alpha_{i,I}$, follow Gaussian distribution. Hence, the output of the encoder is the embedding vector $\boldsymbol{Z} = [(\hat{\theta}_{i,min}, \hat{\theta}_{i,max}), (\mu_{\hat{\alpha}_{i,R}}, \sigma^2_{\hat{\alpha}_{i,R}}), (\mu_{\hat{\alpha}_{i,I}}, \sigma^2_{\hat{\alpha}_{i,I}})]$, which is a collection of the parameters of the distributions of estimated AoAs $\hat{\theta}_i$ and estimated channel gains $\hat{\alpha}_{i,R}$ and $\hat{\alpha}_{i,I}$. For reconstruction of the received signal by the decoder, we generate random samples of the estimated AoAs and channel gains using the reparameterization trick, so that we get an unbiased estimate of the gradient with respect to the weights in the encoder) [9]. Specifically, given the embedding vector $\boldsymbol{Z}$, we generate the following samples

$$\begin{cases} \tilde{\theta}_i = (\hat{\theta}_{i,max} - \hat{\theta}_{i,min}) \times \epsilon^{\theta_i} + \hat{\theta}_{i,min} \\ \tilde{\alpha}_{i,R} = \mu_{\hat{\alpha}_{i,R}} + \sqrt{\sigma^2_{\hat{\alpha}_{i,R}}} \times \epsilon^{\alpha_{i,R}} \\ \tilde{\alpha}_{i,I} = \mu_{\hat{\alpha}_{i,I}} + \sqrt{\sigma^2_{\hat{\alpha}_{i,I}}} \times \epsilon^{\alpha_{i,I}} \end{cases}, \tag{4}$$

where $\epsilon^{\theta_i} \sim \mathcal{U}(0,1)$ and $\epsilon^{\alpha_{i,R}}, \epsilon^{\alpha_{i,I}} \sim \mathcal{N}(0,1)$.

The generated samples go into the decoder, who reconstructs the received signal, called the recovery signal

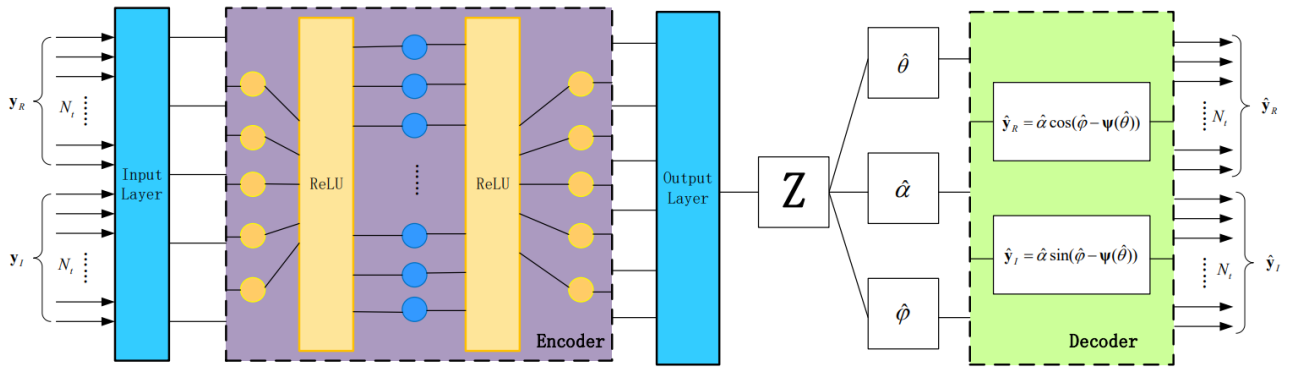


Fig. 1: Illustration of the proposed Knowledge-Aware VAE neural network.

hereafter, based on the known channel model in (3) as follow:

$$\begin{cases} \hat{\boldsymbol{y}}_R = \sum_{i=1}^{M} |\tilde{\alpha}_i| \cos\left(\tilde{\phi}_i - \boldsymbol{\Psi}(\tilde{\theta}_i)\right) \\ \hat{\boldsymbol{y}}_I = \sum_{i=1}^{M} |\tilde{\alpha}_i| \sin\left(\tilde{\phi}_i - \boldsymbol{\Psi}(\tilde{\theta}_i)\right) \end{cases}. \tag{5}$$

Since we focus on unsupervised channel estimation, the network can only take the MSE between the target signal and recovery signal as the loss function without any labels about AoAs, path gains or path angles. The loss function (per training sample of the received signal) is shown below:

$$loss = \frac{1}{N_t}\left(\|\boldsymbol{y}_R - \hat{\boldsymbol{y}}_R\|_2^2 + \|\boldsymbol{y}_I - \hat{\boldsymbol{y}}_I\|_2^2\right). \tag{6}$$

Our goal is to get the accurate estimation of the mean values of the channel features and the minimum and maximum of the estimation AoA from the output of the encoder. After the network is trained, we get the channel estimation from the output of the encoder: $\hat{\theta}_i = (\hat{\theta}_{i,min} + \hat{\theta}_{i,max})/2$, $\hat{\alpha}_{i,R} = \mu_{\hat{\alpha}_{i,R}}$, and $\hat{\alpha}_{i,I} = \mu_{\hat{\alpha}_{i,I}}$. We get the estimation of path angles and path gains by $\hat{\phi}_i = \arctan(\hat{\alpha}_{i,I}/\hat{\alpha}_{i,R})$ and $|\hat{\alpha}_i| = \sqrt{\hat{\alpha}_{i,R}^2 + \hat{\alpha}_{i,I}^2}$.

IV. Multiplicity of Global Optima and Sectorization

A. Multiplicity of Global Optima

A fundamental performance limit of our proposed unsupervised channel estimation method comes from the existence of multiple solutions (i.e., multiple global optima that minimize the loss in (6)). In other words, there are more than one possible AoAs θ_i and channel gains α_i that result in the same noiseless version of the received signal $\boldsymbol{y}$. Since our method is unsupervised (i.e., no labels on AoAs and channel gains), we cannot distinguish between AoAs and channel gains that result in the same received signals.

The following proposition proves the existence of multiple global optima, even when there is only one user.

Proposition 1: Consider the single-user case (i.e., $M = 1$). There exist at least $K = 2 \cdot \lfloor 2\frac{d}{\lambda} \rfloor + 1$ global optima that minimize the loss (6). More specifically, these K global optima have the same path gain and path angle but different AoAs as follows:

$$\theta^k = \arcsin\left(\sin\theta - \left\lceil -2\frac{d}{\lambda} \right\rceil + k\right), \forall k = 0, 1, ..., K-1. \tag{7}$$

Proof: Please see the appendix. ■

Since there are multiple global optima even when there is one user, the number of global optima increase at least exponentially with the number of users. This is because with multiple users, even if the incoming signals from each user are different, the received signal as the supposition of individual users' signals could be the same.

From the characterization of possible AoAs in Proposition 1, we can see that the number of global optima does not depend on the number of antenna elements, and increases with the carrier frequency. Therefore, increasing the number of antenna elements may not solve the problem, and the problem is more severe in millimeter wave systems of higher frequencies.

B. Sectorization Method

To alleviate the problem of multiple global optima, we introduce the sectorization method [11]. We consider a L-order spatial filter in [12] with N_t antennas and the weight vector $\boldsymbol{\omega_1} = [\omega_0, \omega_1, ..., \omega_{L-1}, 0, ..., 0]^H \in \mathbb{C}^{N_t}$, where $\omega_0, \ldots, \omega_{L-1}$ is the coefficient of the filter taps and the superscript and H means conjugate transpose of the vector. Then the basic spatial response of the ULA can be written as:

$$\eta_1(\theta) = \boldsymbol{\omega}_1^H \boldsymbol{a}(\theta) = \sum_{l=0}^{L-1} \omega_l e^{-j\frac{2\pi d}{\lambda} l sin(\theta)}. \tag{8}$$

We assume that the shift step length is regard as m. Then, we can get the $(b+1)$th weight vector $\boldsymbol{\omega_{b+1}}$ by shifting the elements of $\boldsymbol{\omega_1}$ as follows:

$$\boldsymbol{\omega_{b+1}} = [\underbrace{0, 0, ..., 0}_{bm}, \omega_0, \omega_1, ..., \omega_{L-1}, \underbrace{0, 0, ..., 0}_{N_t - bm - L}]^H. \tag{9}$$

As a result, $\boldsymbol{\omega_{b+1}}$ can be written as:

$$\eta_{b+1}(\theta) = \eta_1(\theta) e^{-j\frac{2\pi d}{\lambda} sin(\theta) bm}, 1 \le b \le (N_t - L)/m. \tag{10}$$

Define $\boldsymbol{W} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_B], B = \lfloor (N_t - L)/m \rfloor + 1$. In this way, we can get the spatial responses vector corresponding to all these weight vectors:

$$\boldsymbol{\eta}(\theta) = \boldsymbol{W}^H \boldsymbol{a}(\theta) = \eta_1(\theta) \boldsymbol{a}^m(\theta), \tag{11}$$

where $\boldsymbol{a}^m(\theta)$ can be written as:

$$\boldsymbol{a}^m(\theta) = [1, e^{-j\frac{2\pi d}{\lambda}sin(\theta)m}, ..., e^{-j\frac{2\pi d}{\lambda}sin(\theta)(B-1)m}]^T. \tag{12}$$

From (11) and (12), all the elements in $\boldsymbol{\eta}(\theta)$ have the same magnitude for a given θ which means that after the transformation each spatial response have the same directional characteristics and the phase differences between the elements are the same as those generated by B elements out of N_t antennas with element spacing $m\frac{d}{\lambda}$.

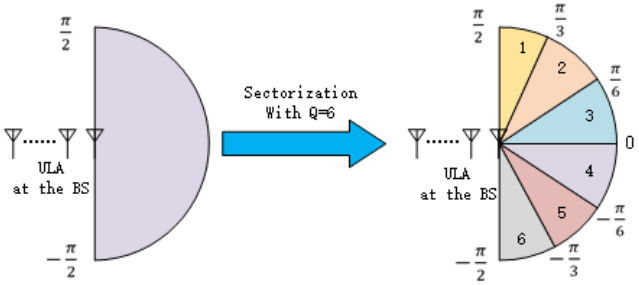


Fig. 2: Sectorization method.

Thus, we can divide the azimuth domain into several sectors by designing the basic vector $\boldsymbol{\omega}_1$ and the shift step length m [13]. In this paper, we assume that the azimuth domain is divided into Q sectors equally. Then the range of AoA to be estimated in the qth sector is $[-\pi/2+(q-1)\pi/Q, \pi/2+q\pi/Q]$. Through sectorization, we divide the estimation range into multiple sectors, which reduce the number of global optima in each sector.

V. Numerical Simulation

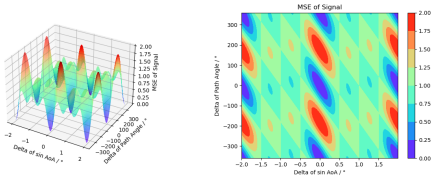
In this section, we first verify our theoretical analysis on the multiplicity of global optima, then demonstrate the effectiveness of sectorization, and finally evaluate the performance improvement of our proposed solution over existing methods.

In all simulations, we fix $d/\lambda = 0.5$ unless otherwise specified. We use a training set of 40000 samples and a testing set of 10000 samples, and set the learning rate to 0.0001, the batch size to 256, and the maximum number of training epochs to 500. There is a batch normalization layer between neurons and activation functions.

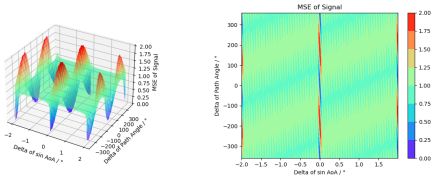
A. Multiplicity of Global Optima

We validate our analysis on the multiplicity of global optima through numerical simulation.

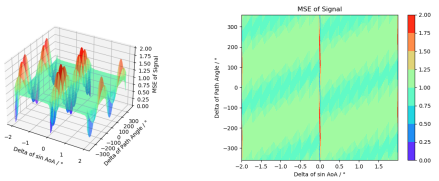
1) Impact of the number of antennas: We consider the single-user case. From Proposition 1, we know that there are at least $K = 3$ global optima with different AoAs. Moreover, the number of antennas should not affect the multiplicity of global optima. In Fig. 3, we show the surfaces of the loss functions under $N_t = 4, 32, 64$ antennas. We can see that there are three distinct AoAs that result in zero loss (i.e., global optima), and that the locations do not change with the number of antennas. The results validate our analysis.



(a) 4 Antennas in ULA (b) 4 Antennas in ULA XOY Projection



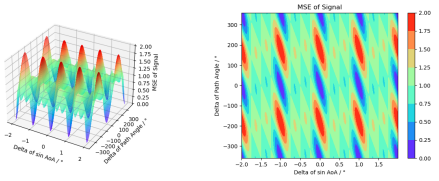
(c) 32 Antennas in ULA (d) 32 Antennas in ULA XOY Projection



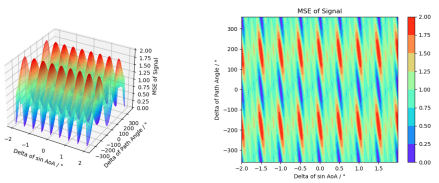
(e) 64 Antennas in ULA (f) 64 Antennas in ULA XOY Projection

Fig. 3: Multiplicity of Global Optima Under Different Numbers of Antennas.

2) Impact of Wavelength and Antennas Spacing: Proposition 1 suggests that the wavelength and antennas spacing, more precisely their ratio d/λ, affect the number of global optima and the AoAs at the global optima. We set the ratio as $d/\lambda = 0.5, 1, 2$. Since the loss surface is plotted in Fig.3(a) and Fig.3(b) for $d/\lambda = 0.5$, we only plot the loss surfaces under $d/\lambda = 1$ and $d/\lambda = 2$ in Fig. 4. We can verify that there are 5 and 9 distinct AoAs at global optima under $d/\lambda = 1$ and $d/\lambda = 2$, respectively, which is consistent with Proposition 1.



(a) $d/\lambda = 1$ (b) $d/\lambda = 1$ XOY Projection



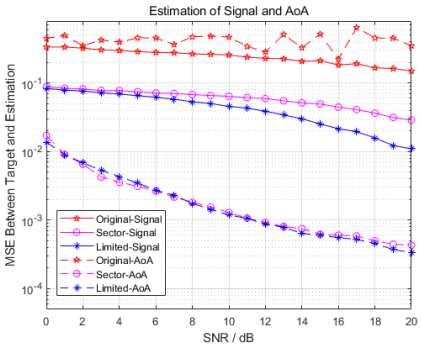
(c) $d/\lambda = 2$ (d) $d/\lambda = 2$ XOY Projection

Fig. 4: Multiplicity of Global Optima under Different d/λ

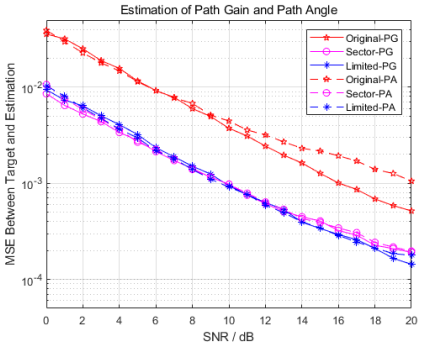
B. Effectiveness of Sectorization

Since there are multiple global optima, we adopt sectorization. Proposition 1 suggests that there are three global optima with distinct AoAs. Therefore, we partition the angle domain $[-\pi/2, \pi/2]$ into three sectors of

$[-\pi/2, -\pi/6]$, $[-\pi/6, \pi/6]$, and $[\pi/6, -\pi/2]$. We compare the performance with the case without sectorization in Fig. 5.



(a) Estimation of Signal and AoA



(b) Estimation of Path Gain and Path Angle

Fig. 5: Effectiveness of Sectorization

From Fig. 5, we can see that sectorization improves the performance across the board, in terms of the signal reconstruction error, and the errors in estimation of AoAs, path angles, and path gains. Such observation is expected because sectorization reduces the number of global optima and decreases the likelihood that the training stops at suboptimal stationary points.

C. Performance Comparison

Finally, we compare the channel estimation accuracy of the KA-VAE network with the following benchmarks:

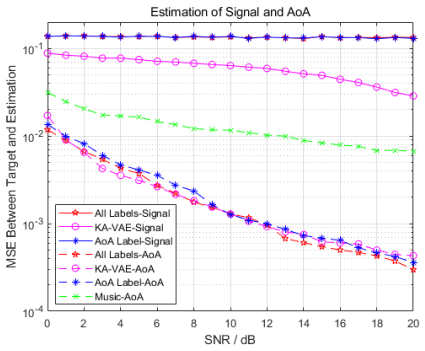
- the MUSIC algorithm, which represents the classic unsupervised channel estimation methods;
- a supervised learning algorithm, for which we have labels of true AoAs, path angles, and path gains (called "All-Label Network") and define the loss function as the total MSE

$$loss_{all} = (\theta - \hat{\theta})^2 + (\alpha - \hat{\alpha})^2 + (\phi - \hat{\phi})^2; \quad (13)$$

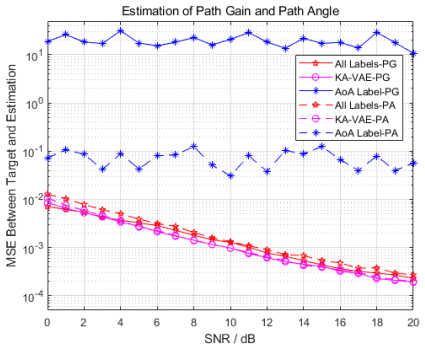
- an variation of standard VAE, for which we have labels of true AoAs and add the MSE of AoAs to the signal reconstruction errors as the loss function:

$$loss_{AoA} = \frac{1}{N_t} \sum_{k=0}^{N_t-1} [(\boldsymbol{y}_R - \hat{\boldsymbol{y}}_R)^2 + (\boldsymbol{y}_I - \hat{\boldsymbol{y}}_I)^2] + (\theta - \hat{\theta})^2. \quad (14)$$

We call the variation of standard VAE "AoA-Label Network". We use the AoA-Label Network to demonstrate that our proposed KA-VAE is able to enforce meaningful output of the encoder by our design of the decoder.



(a) Estimation of Signal and AoA



(b) Estimation of Path Gain and Path Angle

Fig. 6: Performance comparison.

Fig. 6 summarizes the performance evaluation in terms of the signal reconstruction error and the estimation errors of AoAs, path gains, and path angles. Note that for the fully supervised All-Label Network, we do not show the signal reconstruction error because there is no decoder to reconstruct the signal. For the MUSIC algorithm, we do not show the reconstruction error and the estimation errors of path gains and path angles because the algorithm does not produce these estimates.

Our key observation is that the proposed KA-VAE achieves almost identical performance as the fully supervised All-Label Network. This demonstrates the advantage of our 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. In other words, our novel design enables the removal the labels without much performance loss, which would be otherwise impossible.

The performance comparison with the AoA-Label Network illustrates that our proposed decoder enforces meaningful output of the encoder by incorporating the knowledge of the signal propagation model. The AoA-Label Network is the standard VAE with a modified loss function that includes the MSE of the AoA estimation. Note that in the AoA-Label Network, the decoder is the same as the one in the standard VAE, which has learnable parameters. For standard VAEs, we cannot control the physical meanings of the embedding vector (i.e., the output of the encoder). For the AoA-Label Network, we enforce part of the embedding vector to be the AoA through the addition of the MSE of AoA estimation, but have no control over the rest of the embedding vector. As we can expect and

as demonstrated in Fig. 6, the AoA-Label Network has comparable performance with our proposed KA-VAE in terms of AoA estimation, but has much worse performance in signal reconstruction and estimation of path gains and path angles. Such observations further demonstrates that the removal of labels is achieved by our design of the decoder.

VI. Conclusions

This paper proposed a novel variation of VAE for unsupervised channel estimation in massive MIMO systems. The key design idea is to replace the decoder of the standard VAE with a new decoder that "knows the signal propagation model". We instills such knowledge in the proposed decoder by hardwiring it with the signal propagation model. We analyzed the issue of multiple global optima in the unsupervised channel estimation problem, and adopted the sectorization method to alleviate the issue. Numerical simulations were carried out to validate our theoretical analysis. Comparisons with carefully designed benchmark algorithms demonstrated that our proposed method achieves almost identical performance as the fully supervised method, and that the achievement is enabled by our novel design of the decoder.

Appendix A
Proof of Proposition 1

For the reason that the KA-VAE is an unsupervised channel estimation network, which means that it has no label for the estimation of AoA θ, path gain α and path angle ϕ. Thus, the performance of the network may be suffer from the multiplicity of global optima, which may have a low MSE between the target signal and recovery signal while large MSE between target AoA/path gain/path angle and the estimated ones that will bring local convergence points into KA-VAE network. To cover the problem, this section will make a mathematical derivation to find the distribution of local convergence points. Consider (6), it can be divided into two parts where the former part can be regarded as the MSE of the real part between the target signal and recovery signal while the following part is the imaginary part of that, i.e. $loss = loss_R + loss_I$. Substitute (3) and (5) into (6) and ignore the noise:

$$\begin{cases} loss_R = \dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}[(\alpha\cos(\phi-\Psi_k(\theta)) - \hat{\alpha}\cos(\hat{\phi}-\Psi_k(\hat{\theta})))^2] \\ loss_I = \dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}[(\alpha\sin(\phi-\Psi_k(\theta)) - \hat{\alpha}\sin(\hat{\phi}-\Psi_k(\hat{\theta})))^2] \end{cases} \tag{15}$$

Where $\Psi_k(\theta) = 2\pi\frac{d}{\lambda}k\sin(\theta), k\in Z$. For the aim of this paper is to make an accurate estimation on AoA and channel features, in (15), it can be easily found that the estimation of path angle $\hat{\phi}$ has a directly impact on the estimation of AoA $\hat{\theta}$ for they both are the phase of the triangle function cos/sin, while the estimation of path gain $\hat{\alpha}$ has an indirectly impact. As a result, in the following mathematical derivation, we assume that the path gain has been estimated correctly, i.e. $\hat{\alpha} = \alpha$. Then take the loss of the real part into consider first:

$$\begin{aligned} loss_R^{\phi,\theta} =& \frac{1}{N_t}\sum_{k=0}^{N_t-1}\alpha^2[\cos^2(\phi-\Psi_k(\theta)) + \cos^2(\hat{\phi}-\Psi_k(\hat{\theta})) \\ & - 2\cos(\phi-\Psi_k(\theta))\cos(\hat{\phi}-\Psi_k(\hat{\theta}))]. \end{aligned} \tag{16}$$

Similarly, we can get the loss of the image part with the same format. And let $loss^{\phi,\theta} = loss_R^{\phi,\theta} + loss_I^{\phi,\theta}$:

$$\begin{aligned} loss^{\phi,\theta} =& \frac{1}{N_t}\sum_{k=0}^{N_t-1}\alpha^2[2 - 2\cos(\phi-\Psi_k(\theta))\cos(\hat{\phi}-\Psi_k(\hat{\theta})) \\ & - 2\sin(\phi-\Psi_k(\theta))\sin(\hat{\phi}-\Psi_k(\hat{\theta}))] \\ =& 2\alpha^2 - \frac{1}{N_t}\sum_{k=0}^{N_t-1}2\alpha^2[\cos[(\phi-\hat{\phi}) - (\Psi_k(\theta)-\Psi_k(\hat{\theta}))]. \end{aligned} \tag{17}$$

For the reason that $\cos[(\phi-\hat{\phi}) - (\Psi_k(\theta)-\Psi_k(\hat{\theta}))] \leq 1$, we have $loss^{\phi,\theta} \geq 0$. Furthermore, the KA-VAE network will converge in the direction where the loss function drops the fastest. Thus, without the affect of noise, $loss^{\phi,\theta}$ = 0 is the local minimum value as well as global minimum value. Let $f(\hat{\phi},\hat{\theta}) = (\phi-\hat{\phi}) - (\Psi_k(\theta)-\Psi_k(\hat{\theta}))$.

$$\begin{cases} \dfrac{\partial loss^{\phi,\theta}}{\partial\hat{\phi}} = -\dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}2\alpha^2\sin[f(\hat{\phi},\hat{\theta})] \\ \dfrac{\partial loss^{\phi,\theta}}{\partial\hat{\theta}} = \dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}2\alpha^2\sin[f(\hat{\phi},\hat{\theta})]\dfrac{\partial\Psi_k(\hat{\theta})}{\partial\hat{\theta}} \end{cases} \tag{18}$$

Where $\frac{\partial\Psi_k(\hat{\theta})}{\partial\hat{\theta}} = \frac{2\pi dk}{\lambda}\cos(\hat{\theta})$. Let $f_{\hat{\phi}}(\hat{\phi},\hat{\theta}) = \frac{\partial loss^{\phi,\theta}}{\partial\hat{\phi}}$, $f_{\hat{\theta}}(\hat{\phi},\hat{\theta}) = \frac{\partial loss^{\phi,\theta}}{\partial\hat{\theta}}$. Now assume that (ϕ_0,θ_0) is the local minimum of $loss^{\phi,\theta}$. Then it must have $f_{\hat{\phi}}(\hat{\phi}_0,\hat{\theta}_0) = 0$ and $f_{\hat{\theta}}(\hat{\phi}_0,\hat{\theta}_0) = 0$. From (18), we have $f(\hat{\phi}_0,\hat{\theta}_0) = (\phi-\hat{\phi}_0) - (\Psi_k(\theta)-\Psi_k(\hat{\theta}_0)) = n\pi, n\in Z$. Drive the second derivative to find the extreme point:

$$\begin{cases} \dfrac{\partial(loss^{\phi,\theta})^2}{\partial^2\hat{\phi}} = \dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}2\alpha^2\cos[f(\hat{\phi},\hat{\theta})] \\ \dfrac{\partial(loss^{\phi,\theta})^2}{\partial\hat{\phi}\partial\hat{\theta}} = -\dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}2\alpha^2\cos[f(\hat{\phi},\hat{\theta})]\dfrac{\partial\Psi_k(\hat{\theta})}{\partial\hat{\theta}} \\ \dfrac{\partial(loss^{\phi,\theta})^2}{\partial^2\hat{\theta}} = \dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}2\alpha^2\sin[f(\hat{\phi},\hat{\theta})]\dfrac{\partial(\Psi_k(\hat{\theta}))^2}{\partial^2\hat{\theta}} \\ \qquad\qquad + \dfrac{1}{N_t}\displaystyle\sum_{k=0}^{N_t-1}2\alpha^2\cos[f(\hat{\phi},\hat{\theta})](\dfrac{\partial\Psi_k(\hat{\theta})}{\partial\hat{\theta}})^2 \end{cases} \tag{19}$$

Where $\frac{\partial(\Psi_k(\hat{\theta}))^2}{\partial^2\hat{\theta}} = -\frac{2\pi dk}{\lambda}\sin(\hat{\theta})$. Let $A = \frac{\partial(loss^{\phi,\theta})^2}{\partial^2\hat{\phi}}, B = \frac{\partial(loss^{\phi,\theta})^2}{\partial\hat{\phi}\partial\hat{\theta}}, C = \frac{\partial(loss^{\phi,\theta})^2}{\partial^2\hat{\theta}}$. When $f(\hat{\phi}_0,\hat{\theta}_0) = 2n\pi, n\in Z$, e.g. $\cos[f(\hat{\phi}_0,\hat{\theta}_0)] = 1$, we have

$$\begin{aligned} AC - B^2 =& (2\alpha^2\frac{2\pi d}{\lambda}\cos\hat{\theta})^2[\frac{1}{N_t}\sum_{k=0}^{N_t-1}k^2 - (\frac{1}{N_t}\sum_{k=0}^{N_t-1}k)^2] \\ =& \frac{4}{3}(\alpha^2\frac{2\pi d}{\lambda}\cos\hat{\theta})^2[N_t^2 + 7N_t + 5(N_t-1)]. \end{aligned} \tag{20}$$

$$[\boldsymbol{H}|\boldsymbol{b}] = \left[\begin{array}{cc:c} 1 & 0 & 2n_0\pi \\ 1 & \frac{2\pi d}{\lambda} & 2n_1\pi \\ \vdots & \vdots & \vdots \\ 1 & \frac{2\pi d}{\lambda}(N_t-1) & 2n_{N_t-1}\pi \end{array}\right] \overset{transform}{\longrightarrow} \left[\begin{array}{cc:c} 1 & 0 & 2n_0\pi \\ 0 & \frac{2\pi d}{\lambda} & 2(n_1-n_0)\pi \\ \vdots & \vdots & \vdots \\ 0 & 0 & 2(n_{N_t-1}+n_{N_t-3}-2n_{N_t-2})\pi \end{array}\right] = [\boldsymbol{H}'|\boldsymbol{b}']. \tag{23}$$

$$\begin{cases} \Delta\hat{\phi} = 2n_0\pi \\ -\frac{2\pi d}{\lambda}\Delta\hat{\theta} = 2(n_1-n_0)\pi \\ 0 = 2(n_2+n_0-2n_1)\pi \\ \vdots \\ 0 = 2(n_{N_t-1}+n_{N_t-3}-2n_{N_t-2})\pi \end{cases} \Rightarrow \begin{cases} \Delta\hat{\phi} = 2n_0\pi \\ \Delta\hat{\theta} = \frac{\lambda}{d}(n_0-n_1) \\ n_2 = 2n_1-n_0 \\ \vdots \\ n_{N_t-1} = 2n_{N_t-2}-n_{N_t-3} \end{cases}. \tag{24}$$

When $f(\hat{\phi}_0,\hat{\theta}_0) = (2n+1)\pi, n\in Z$, e.g. $\cos[f(\hat{\phi}_0,\hat{\theta}_0)] = -1$, we will get the same result. From the result of (20), it can be easily seen that when $N_t > 1$ and $f(\hat{\phi}_0,\hat{\theta}_0) = 2n\pi$, for $n\in Z$ and $n\in[0, N_t-1]$, we have $AC - B^2 > 0$ and $A = 2\alpha^2 > 0$ which means $(\hat{\phi}_0,\hat{\theta}_0)$ is the local minimum of the $loss^{\phi,\theta}$, e.g. $loss^{\phi,\theta} = 0$, with $\cos[f(\hat{\phi}_0,\hat{\theta}_0)] = 1$. It means that for all $n = 0, 1, ..., N_t-1$, the loss function will reach its local optimal solution where estimated AoA $\hat{\theta}$, estimated path angle $\hat{\phi}$, estimated path gain $\hat{\alpha}$ will be the output of the network, if $\hat{\phi}$ and $\hat{\theta}$ satisfy:

$$\phi - \hat{\phi} - \frac{2\pi dk}{\lambda}(\sin\theta - \sin\hat{\theta}) = 2n\pi, \quad n\in Z. \tag{21}$$

In this problem, the difference between target AoA and the estimated AoA and the difference target path angle and estimated path angle should be placed most attention. Let $\Delta\hat{\phi} = \phi - \hat{\phi}$ and $\Delta\hat{\theta} = \sin\theta - \sin\hat{\theta}$. Then $f(\hat{\phi},\hat{\theta}) = f(\Delta\hat{\phi},\Delta\hat{\theta}) = \Delta\hat{\phi} - \frac{2\pi dk}{\lambda}\Delta\hat{\theta}$, and from what has been discussed above, let $f(\Delta\hat{\phi},\Delta\hat{\theta}) = 2n_i\pi$, for $i = 0, 1, ..., N_t-1$ and $n_i\in Z$. Turn it into matrix format $\boldsymbol{Hx} = \boldsymbol{b}$:

$$\left[\begin{array}{cc} 1 & 0 \\ 1 & \frac{2\pi d}{\lambda} \\ \vdots & \vdots \\ 1 & \frac{2\pi d}{\lambda}(N_t-1) \end{array}\right] \left[\begin{array}{c} \Delta\hat{\phi} \\ -\Delta\hat{\theta} \end{array}\right] = \left[\begin{array}{c} 2n_0\pi \\ 2n_1\pi \\ \vdots \\ 2n_{N_t-1}\pi \end{array}\right]. \tag{22}$$

Perform row transformation on the augmented matrix $[\boldsymbol{H}|\boldsymbol{b}]$. Then, we get $[\boldsymbol{H}'|\boldsymbol{b}']$ as (23) shows.

As a result, $\boldsymbol{H}'\boldsymbol{x} = \boldsymbol{b}'$ which can be expressed as (24).Then it can be find out that the multiplicity of global optima is suffered from the difference between target path angle and estimated path angle $\Delta\hat{\phi}$ and the difference between target AoA and estimated AoA $\Delta\hat{\theta}$, while has nothing to do with the number of antennas of the ULA. For the reason that $\Delta\hat{\phi} = \phi - \hat{\phi} = 2n_0\pi$, where $\phi,\hat{\phi}\in[-\pi,\pi], n_0\in Z$, we can easily get $\Delta\hat{\phi}\in[-2\pi, 2\pi]$ and $n_0 = 0$ or ±1. Because of $\Delta\hat{\theta} = \sin\theta - \sin\hat{\theta}\in[-2, 2]$ and $n_1 \in Z$, there will be a convergence point for the network if $\Delta\hat{\theta} = \frac{\lambda}{d}(n_0 - n_1)$ where $n_0 = 0$ or ±1. Furthermore, n_i can be derived by iteration, where $i = 2, 3, ..., N_t-1$.

We know that if $\Delta\hat{\phi} = 0$ and $\Delta\hat{\theta} = 0$, there will be a correction estimation on the channel feature. However, if $\Delta\hat{\phi} = \pm2\pi$ and $\Delta\hat{\theta} = 0$, there will also be a correction estimation, for $\Delta\hat{\phi} = \pm2\pi$ means that $\phi = \pm\pi$ and $\hat{\phi} = \mp\pi$, respectively, which means $\hat{\phi} = \phi$ is still hold, i.g.$\Delta\hat{\phi} = 0$ is equivalent to $\Delta\hat{\phi} = \pm2\pi$. Thus, (24) can be simplified as:

$$\begin{cases} \Delta\hat{\phi} = 2n_0\pi \\ \Delta\hat{\theta} = \frac{\lambda}{d}(n_0-n_1) \\ n_2 = 2n_1-n_0 \\ \vdots \\ n_{N_t-1} = 2n_{N_t-2}-n_{N_t-3} \end{cases} \Rightarrow \begin{cases} \Delta\hat{\phi} = 0 \\ \Delta\hat{\theta} = -\frac{\lambda}{d}n_1 \\ n_2 = 2n_1 \\ \vdots \\ n_{N_t-1} = (N_t-1)n_1 \end{cases}. \tag{25}$$

Where $n_1 = -\frac{d}{\lambda}\Delta\hat{\theta}\in[-2\frac{d}{\lambda}, 2\frac{d}{\lambda}], n_1\in Z$. For the reason that $d/\lambda \geq 1/2$, which means $[-1, 1]\subset[-2\frac{d}{\lambda}, 2\frac{d}{\lambda}]$, then $n_1 = 0$ and ±1 must be the solution of (25). So, there must be muti-convergence point in the unsupervised KA-VAE network. With the increase of the d/λ, the number of the solution for (25) can be expressed as $N_{loc} = 2\times\lfloor 2\frac{d}{\lambda}\rfloor + 1$, where $\lfloor x\rfloor$ denotes the maximum integer less than or equal to x.

For $n_1 = -\frac{d}{\lambda}\Delta\hat{\theta}\in[-2\frac{d}{\lambda}, 2\frac{d}{\lambda}], n_1\in Z$ and there are N_{loc} solution for (25), then let $n_1^0 = n_1^{min} = \lceil -2\frac{d}{\lambda}\rceil$ and $n_1^{N_{loc}-1} = n_1^{max} = \lfloor 2\frac{d}{\lambda}\rfloor$, where $\lceil x\rceil$ denotes the minimum integer greater than or equal to x. Furthermore, for $n_1\in Z$, then the set of all n_1 satisfying (25) is as follows:

$$\begin{aligned} \boldsymbol{n_1} &= [n_1^0, n_1^1, ..., n_1^{N_{loc}-1}] \\ &= [\underbrace{\left\lceil -2\frac{d}{\lambda}\right\rceil, \left\lceil -2\frac{d}{\lambda}\right\rceil + 1, ..., -1}_{(N_{loc}-1)/2}, 0, \underbrace{1, ..., \left\lfloor 2\frac{d}{\lambda}\right\rfloor - 1, \left\lfloor 2\frac{d}{\lambda}\right\rfloor}_{(N_{loc}-1)/2}]. \end{aligned} \tag{26}$$

Consider that when θ, $\hat{\theta}\in[-\pi/2, \pi/2]$ and the target AoA θ is set, $\Delta\hat{\theta} = \sin(\theta) - \sin(\hat{\theta}) = C - \sin(\hat{\theta})$, where C is a constant which means that $\Delta\hat{\theta}$ is a monotonic function of $\hat{\theta}$. According to (26), when the target AoA is θ, the estimation AoA $\hat{\theta}_i$ can be:

$$\hat{\theta}_k = \arcsin(\sin(\theta) - n_1^k) = \arcsin(\sin(\theta) - \left\lceil -2\frac{d}{\lambda}\right\rceil + k). \tag{27}$$

Where $k = 0, 1, ..., N_{loc} - 1$. And only when $\hat{\theta}_k \in [-\pi/2, \pi/2]$, it can be retained as a solution.

REFERENCES

[1] T. L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 9(11):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*, 19(6):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)*, pages 1–3, 2019.

[4] 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)*, pages 75–78, March 2019.

[5] M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh. Deep learning-based channel estimation. *IEEE Communications Letters*, 23(4):652–655, 2019.

[6] A. Paulraj, R. Roy, and T. Kailath. Estimation of signal parameters via rotational invariance techniques- esprit. In *Nineteeth Asilomar Conference on Circuits, Systems and Computers, 1985.*, pages 83–89, 1985.

[7] 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,*, pages 2605–2608 vol.4, 1989.

[8] 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*, 67(9):8549–8560, 2018.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014.

[10] Q. Zhang, J. Zhu, N. Zhang, and Z. Xu. Multidimensional variational line spectra estimation. *IEEE Signal Processing Letters*, 27:945–949, 2020.

[11] 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*, 18(10):1827–1830, 2014.

[12] B. D. Van Veen and K. M. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

[13] J. Li, Q. He, L. Xiao, X. Xu, and S. Zhou. Uplink sum-throughput evaluation of sectorized multi-cell massive mimo system. In *2015 IEEE International Conference on Communication Workshop (ICCW)*, pages 1143–1148, 2015.