Variational Inference Based Spatio-temporal Double Non-stationary Statistical Channel State Information Estimation for Massive MIMO System

Mingfeng Cheng, Wei Peng, *Senior Member, IEEE,* and Tao Jiang, *Fellow, IEEE*

***Abstract*—This paper studies the statistical channel state information (S-CSI) estimation in spatial-temporal double non-stationary massive multiple-input multiple-output (MIMO) systems. First, We propose a structured variational inference model (VSCM), by which incorporates with the hidden Markov model (HMM) is integrated as a latent model. Second, we formulate the state ambiguity problem as a semidefinite programming problem with non-convex rank constraint, and a nonlinear programming based solution is proposed. Then, an iterative EM-like algorithm for model parameters optimization and latent sequence inference is proposed. Last, numerical simulations are carried out, revealing that the S-CSI estimation can be achieved by our model. Besides, the proposed VSCM performs better under low SNR condition compared to the traditional HMM.**

***Index Terms*—Massive MIMO, Spatio-temporal double non-stationary, S-CSI, Variational inference, HMM, Semidefinite programming.**

I. Introduction

Massive multiple-input multiple-output (MIMO) systems employ a large number of base station (BS) antennas to serve several user equipments (UE) simultaneously. With the enormous benefits of diversity gain, massive MIMO has been playing a pivotal role in next-generation communication system [1], [2]. While the great benefits of massive MIMO are based on the assumption that BS has adequate knowledge of channel state information (CSI) [3]. Therefore, the CSI has to be estimated accurately.

In time-division duplex (TDD) systems, the downlink channel can be estimated through uplink training due to the channel reciprocity. While the complexity of channel estimaton is extremely high because of the matrix inversion. Much effort has been devoted to statistical channel state information (S-CSI) estimation. In [4], [5], S-CSI is estimated via averaging instantaneous CSI over time. In [6], [7], channels are modeled as Gaussian-mixture distributions, and the parameters are trained via expectation maximization (EM) algorithm.

Methods in [4]–[7] are based on the assumption that channels are wide-sense stationary. While many researches point out the non-stationary property of channels in massive MIMO systems [8]–[10]. Actually, a number of models have already been proposed to solve the non-stationary problems in high-speed railway communication systems [11], [12]. Besides, a hidden Markov model (HMM) based statistical channel model has proposed an HMM-based model to estimate the S-CSI in temporal non-stationary massive MIMO systems [13]. However,the large aperture antenna arrays and the relatively small propagation distance between BS and UE in a massive MIMO system give rise to more complicated conditions and one of them is the spatial non-stationarity [14], [15]. It is elaborated in [14] that the spatial non-stationarity exists due to the fact that different regions of the BS antenna array receive diverse levels of signal power, which is caused by the different visibility regions of subarrays as well as the various reflecting scatters on the propagation paths.

This paper proposes a subchannel-wise channel model to explore the spatio-temporal double non-stationarity in massive MIMO systems. The contamination-free single-cell scenario is considered for simplicity. Firstly, we propose an extended variational auto-encoder (VAE) model, which integrates the HMM structure to present the probabilistic relations between the observed signal sequence and the statistical channel states by introducing latent variables. Secondly, we analyze the state ambiguity problem that commonly occurs in the multi-user scenario as a semidefinite programming problem and provide a factorization based solution. Then, an interative EM-like algorithm for model parameter optimization ans S-CSI sequence is propsed. Simulation results show that our proposed model and the S-CSI sequence estimation algorithm can get an acceptable accuracy rate and performs better than the traditional HMM, especially under the low SNR condition.

The following notations are used through this paper: $\mathbf{I}_M$ denotes the M- dimensional identity matrix. $\mathbb{R}$ and $\mathbb{C}$ denote real numbers and complex numbers, respectively.$(\cdot)^{\mathrm{T}}$ is the transpose and $(\cdot)^{\mathrm{H}}$ stands for conjugate transpose. $\mathbb{E}[\cdot]$ represents the expectation. $\|\cdot\|_F$ is Frobenius norm. $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote the random vector complying with real Gaussian distribution and complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, respectively. To sum up, Table I present all defined variables in this paper.

II. System model

A. Massive MIMO system

In this paper, the uplink transmission in the single-cell scenario is studied. Considering one base station with a uniform linear array of M antennas and N single-antenna user equipment (UE). At time t, the equivalent baseband received signal is given as:

$$\mathbf{y}_t{=}\mathbf{H}_t\mathbf{x}_t{+}\mathbf{n}_t, \tag{1}$$

where $\mathbf{y}_t{=}\left[y_t^1,y_t^2,\ldots,y_t^M\right]^{\mathrm{T}}{\in}\mathbb{C}^{M\times 1}$ is the received signal vector. $\mathbf{H}_t{=}\left[\mathbf{h}_t^1,\mathbf{h}_t^2,\ldots,\mathbf{h}_t^N\right]{\in}\mathbb{C}^{M\times N}$ is the uplink transmission channel matrix at time t where $\mathbf{h}_t^i{\in}\mathbb{C}^{M\times 1}$ denotes the

TABLE I
Variable definition

Variables	Definitions
$\mathbf{H}_t\in\mathbb{C}^{M\times N}$	Channel matrix
$\mathbf{x}_t \in \mathbb{C}^{M\times 1}$	Transmitted signal vector
$\mathbf{y}_t \in \mathbb{C}^{N\times 1}$	Received signal vector
$\mathbf{n}_t\in\mathbb{C}^{M\times 1}$	Noise vector
$\mathbf{P}\in\mathbb{R}^{N\times 1}$	Power allocate vector
$\mathbf{Y}=[\mathbf{y}_1,\cdots,\mathbf{y}_T]$	The observed sequence from moment 1 to T
$\mathbf{Z}=[\mathbf{z}_1,\cdots,\mathbf{z}_T]$	The hidden sequence from moment 1 to T
$\boldsymbol{\mathcal{V}}=\{\mathcal{V}_1,\mathcal{V}_2,\cdots,\mathcal{V}_D\}$	Statistical channel state set (SCSS)
$\boldsymbol{\mathcal{H}}=\{\mathcal{H}_1,\cdots,\mathcal{H}_K\}$	Maekov hidden state set
$\mathcal{L}(\cdot)$	The variational lower bound
$\mathcal{Z},\mathcal{S},\mathcal{Y}$	Variable space
$\{\theta,\Phi,\delta\}$	Graphical model parameter set
$\mathbf{D}_{ij}$	Distance matrix
$\mathfrak{L}(\cdot)$	Lagrangian function

channel vector for the i-th UE. $\mathbf{x}_t=\left[x_t^1,x_t^2,\dots,x_t^N\right]^{\mathrm{T}}\in\mathbb{C}^{N\times 1}$ represents the transmitted signal vector. $\mathbf{n}_t\in\mathbb{C}^{M\times 1}$ is the vector of complex additive white Gaussian noise. The average transmit power of the i-th UE is

$$P_i=\mathbb{E}\left[\|x_t^i\|_2^2\right], \tag{2}$$

and the total transmit power of N UEs is:

$$P_{\text{total}}=\sum_{i=1}^{N}P_i. \tag{3}$$

B. Spatial-temporal double Non-stationary environment

In practical massive MIMO systems, channels vary not only in the time domain but also in the spatial domain [12], [16], [17]. As illustrated in the Fig.1, the reflecting clusters will change when the UEs move. In addition, if UEs or reflecting clusters locate within the Rayleigh distance, the conventionally used assumption of planar wavefront would no longer be applicable, instead, the spherical wavefront will appear, which is called near-field propagation [18]–[20]. As a result, the channel is non-stationary in the time domain. On the other hand, for large aperture antenna arrays, different regions of array will receive varying levels of signal power [14]. Thus the channel is non-stationary in the spatial domain. To be more practical, in this paper, the spatio-temporal double non-stationarity is considered.

Assume that elements of the channel vector $\mathbf{h}_t^i\in\mathbb{C}^{M\times 1}$ in the uplink channel matrix $\mathbf{H}_t$ are independent and identical random variables following Rayleigh distribution with zero mean and variance σ_i^2. Note that columns of the channel matrix $\mathbf{H}_t$ represent the channel vectors of different UEs, and they obey independent but non-identical distributions. Following [11], suppose the statistical channel state set (SCSS) of all the UEs has a finite cardinality D. Denoting the SCSS by $\mathcal{V}=\{\mathcal{V}_1,\mathcal{V}_2,\cdots,\mathcal{V}_D\}$ where $\sigma_i^2\in\mathcal{V}$. Hereafter it is assumed that the elements of SCSS can be acquired through pilot-training and can be used as a priori knowledge.

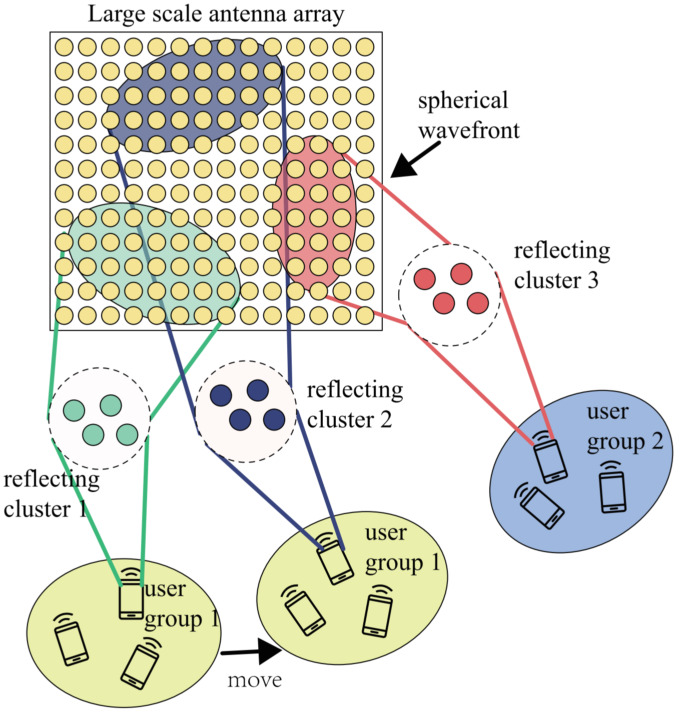


Fig. 1. Spatial-temporal double non-stationary environment

C. HMM structured channel model

The HMM is comprised of a bivariate stochastic process denoted as $\{\mathbf{z}_t,\mathbf{y}_t\}$, where $\{\mathbf{z}_t\}$ is the stochastic process for the hidden state and $\{\mathbf{y}_t\}$ is stochastic process for the observed signal. Specifically, $\{\mathbf{z}_t\}$ is an N-variate, discrete-time, finite-state, and homogeneous Markov chain. A 2-variate example of HMM structured channel model $\{\mathbf{z}_t,\mathbf{y}_t\}$ is shown in Fig.2.

Let $\mathbf{Y}=[\mathbf{y}_1,\cdots,\mathbf{y}_T]$ be the observed sequence from moment 1 to T, $\mathbf{Z}=[\mathbf{z}_1,\cdots,\mathbf{z}_T]$ be the hidden sequence with $\mathbf{z}_i\in\boldsymbol{\mathcal{H}}=\{\mathcal{H}_1,\cdots,\mathcal{H}_K\}$, where $\boldsymbol{\mathcal{H}}$ is the K-element hidden state set and $\mathcal{H}_k$ is the k-th hidden state.

A discrete-time, finite-state HMM is determined by three parameters which are the initial state probability vector $\boldsymbol{\pi}$, the transition probability matrix $\mathbf{A}$ and the emission matrix $\mathbf{B}$. The observed process $\{\mathbf{y}_t\}$ is determined by the hidden process $\{\mathbf{z}_t\}$ in the form of the emission matrix $\mathbf{B}$.

The definitions of HMM parameters are given below:

Definition of $\boldsymbol{\pi}$: $\boldsymbol{\pi}=[\pi_1,\cdots,\pi_K]^{\mathrm{T}}\in\mathbb{R}^{K\times 1}$ represents the initial probability distribution of hidden state $\mathbf{z}_1$ where $\pi_k=p(\mathbf{z}_1=\mathcal{H}_k)$ and $\sum_{i=1}^{K}\pi_i=1$.

Definition of $\mathbf{A}$: $\mathbf{A}\in\mathbb{R}^{K\times K}$ is the transition matrix, its (i,j)-th element means the transition probability from hidden state $\mathbf{z}_{t-1}=\mathcal{H}_i$ to $\mathbf{z}_t=\mathcal{H}_j$, and $\sum_{j=1}^{K}\mathbf{A}_{ij}=1$.

$$\mathbf{A}_{ij}=p(\mathbf{z}_t=\mathcal{H}_j|\mathbf{z}_{t-1}=\mathcal{H}_i). \tag{4}$$

Definition of γ: $\gamma_k(t)$ is the probability of being in state k at moment t given the observed sequence $\mathbf{Y}$,

$$\gamma_k(t)=p(\mathbf{z}_t=\mathcal{H}_k|\mathbf{Y}). \tag{5}$$

Definition of ξ: $\xi_{jk}(t)$ is the probability of being in state j and k at moments t and $t+1$ respectively given the observed sequence $\mathbf{Y}$,

$$\xi_{jk}(t){=}p(\mathbf{z}_t{=}\mathcal{H}_j, \mathbf{z}_{t+1}{=}\mathcal{H}_k|\mathbf{Y}). \tag{6}$$

With regard to the massive MIMO system, the received signal sequence can be modeled as the observed process and the statistical channel states are the hidden process. Due to the spatial non-stationary property, the hidden state vector $\mathbf{z}_t{=}\left[z_t^1, z_t^2, \ldots, z_t^N\right]^T{\in}\mathbb{R}^{N\times 1}$ is consist of statistical channel state (SCS) of N-UEs , where $z_t^i{\in}\mathcal{V}$ is the SCS of the i-th UE. Naturally, the cardinality of the finite state set of the Markov chain $\left\{\mathbf{z}_t\right\}$ is $K{=}D^N$, which equals to the permutation number of N UEs' SCSs.

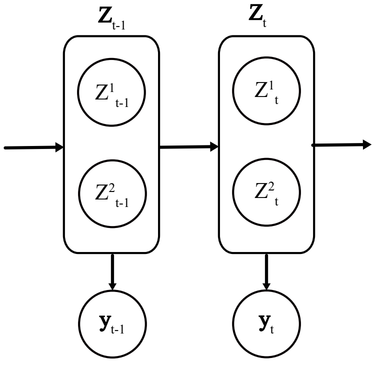


Fig. 2. A 2-variate example for HMM

Note that the S-CSI to be estimated is actually the variance of each element in the channel matrix $\mathbf{H}_t$. The covariance matrix of $\mathbf{y}_t$ is calculated, given as:

$$\mathbb{E}\left[\mathbf{y}_t{\mathbf{y}_t}^{\mathrm{H}}\right]{=}\mathbb{E}\left[(\mathbf{H}_t\mathbf{x}_t)(\mathbf{H}_t\mathbf{x}_t)^{\mathrm{H}}\right]{+}2\mathbb{E}\left[\mathbf{H}_t\mathbf{x}_t\right]\mathbb{E}[\mathbf{n}_t]{+}\mathbb{E}\left[\mathbf{n}_t\mathbf{n}_t^{\mathrm{H}}\right]. \tag{7}$$

While $\mathbb{E}\left[\mathbf{H}_t\mathbf{x}_t\right]\mathbb{E}[\mathbf{n}_t]{=}0$, (7) turns to

$$\mathbb{E}\left[\mathbf{y}_t{\mathbf{y}_t}^{\mathrm{H}}\right] = \left(\langle\mathbf{P}, \mathbf{z}_t\rangle{+}\sigma_n^2\right)\mathbf{I}_M, \tag{8}$$

where $\mathbf{P}{=}[P_1, P_2, \cdots, P_N]^{\mathrm{T}}{\in}\mathbb{R}^{N\times 1}$ is the power allocating vector, and P_i is the average transmit power of the i-th UE, $\mathbf{I}_M$ is an M-dimensional identity matrix, and $\langle\cdot\rangle$ denotes the inner product. It's clear that $\mathbb{E}[\mathbf{y}_t]{=}\mathbf{0}$. Therefore, the probability relationship between the statistical channel state $\mathbf{z}_t$ and the observed signal vector $\mathbf{y}_t$ can be described as a multi-variate Gaussian distribution, given as:

$$P(\mathbf{y}_t|\mathbf{z}_t) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}). \tag{9}$$

The above conclusion can also be found in [21].

III. Variational inference based statistical channel state estimation

A drawback of HMM is that the effect of the noise has not been mathematically formulated. In this section, we propose a variational statistical channel model (VSCM) for the massive MIMO system, by which both the HMM structure and the noise modeling are integrated.

A. Variational auto-encoder

Variational auto-encoder (VAE) is a generative model based on the classic auto-encoder [22]. The graphical VAE model is shown in Fig.3. VAE aims at rebuilding the dataset $\mathcal{Y}{=}\{\mathbf{Y}_n\}$ by introducing a latent variable space $\mathcal{Z}{=}\{\mathbf{Z}_n\}$, so that the unknown distribution $\ln p(\mathcal{Y})$ can be approximate through an encode and decode process. In Fig.3, the hollow circle represents the latent variable, and the shaded circle represents the observable variable. The rounded rectangle represents the iterative loop where C is the loop number. The dashed lines represent the encoding process from $\mathbf{Y}$ to $\mathbf{Z}$ and the solid lines denote the decoding process from $\mathbf{Z}$ to regenerated $\mathbf{Y}$, where parameters Θ and Φ are jointly trained through one neural network.

With the help of latent variables, the marginal likelihhod of $\mathbf{Y}$ is obtained as

$$\ln p(\mathcal{Y}){=}\sum_{n=1}^{C}\ln p(\mathbf{Y}_n), \tag{10}$$

where

$$\ln p(\mathbf{Y}_n){=}\mathrm{KL}\big(q(\mathbf{Z}_n|\mathbf{Y}_n)||p(\mathbf{Z}_n|\mathbf{Y}_n)\big){+}\mathcal{L}(\mathbf{Y}_n). \tag{11}$$

In (11), $p(\mathbf{Z}_n|\mathbf{Y}_n)$ is the true posterior probability which is intractable to obtain and $q(\mathbf{Z}_n|\mathbf{Y}_n)$ is the approximate posterior probability. $\mathrm{KL}\big(q(\mathbf{Z}_n|\mathbf{Y}_n)||p(\mathbf{Z}_n|\mathbf{Y}_n)\big)$ is the Kullback-Leibler (KL) divergence, representing the divergence of approximate from the true posterior, given by

$$\mathrm{KL}\big(q(\mathbf{Z}_n|\mathbf{Y}_n)||p(\mathbf{Z}_n|\mathbf{Y}_n)\big){=}\int q(\mathbf{Z}_n|\mathbf{Y}_n)\frac{q(\mathbf{Z}_n|\mathbf{Y}_n)}{p(\mathbf{Z}_n|\mathbf{Y}_n)}d\mathbf{Z}_n. \tag{12}$$

$\mathcal{L}(\mathbf{Y}_n)$ is the variational lower bound on the marginal likelihood of $\mathbf{Y}_n$, defined as

$$\begin{aligned}&\mathcal{L}(\mathbf{Y}_n)\\&{=}\mathbb{E}_{q(\mathbf{Z}_n|\mathbf{Y}_n)}\left[\ln p(\mathbf{Y}_n|\mathbf{Z}_n)\right]{-}\mathrm{KL}\big(q(\mathbf{Z}_n|\mathbf{Y}_n)||p(\mathbf{Z}_n|\mathbf{Y}_n)\big).\end{aligned} \tag{13}$$

In the decoding process, the observation data $\mathbf{Y}_n$ is represented as the sum of a non-linear transformation of the latent variable $\mathbf{Z}_n$ and an additive white Gaussian noise. Denoting non-linear function as $f(\cdot;\Theta)$, $\mathbf{Y}_n$ is given as

$$\mathbf{Y}_n{=}f(\mathbf{Z}_n;\Theta){+}\mathbf{v}_n, \tag{14}$$

where $\mathbf{v}_n$ is the noise.

With the objective to approximate the distribution of $\mathbf{Y}_n$, the KL divergence in (11) need to be minimized. Equivalently, the lower bound in (11) will be maximized. However, the posterior probability $p(\mathbf{Z}_n|\mathbf{Y}_n)$ is intractable in the straightforward way. Alternatively, a variational inference will be performed, denoted as

$$q(\mathbf{Z}_n|\mathbf{Y}_n;\Phi){=}\mathcal{N}(\mathbf{Z}_n;\mu,\Sigma), \tag{15}$$

where μ and Σ are obtained through neural network, which use $\mathbf{Y}_n$ as input, denoted as

$$(\mu,\Sigma){=}g(\mathbf{Y}_n;\Phi). \tag{16}$$

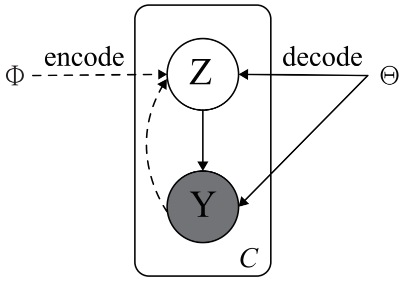


Fig. 3. Graphical model of VAE

In short , VAE can be optimized by alternately performing the expectation step and maximization step. To be specific, the expectation step will infer $q(\mathbf{Z}_n)$ to approximate $p(\mathbf{Z}_n|\mathbf{Y}_n)$, and the maximization step will maximize the lower bound $\mathcal{L}$ with respect to its parameters.

B. Variational statistical channel estimation model

As discussed in the proceeding subsections, in (7), we have derived that the covariance matrix is spherical, all the entries in the main diagonal are the same. Therefore, (9) could be transformed as:

$$\ln\left[p(\mathbf{y}_t|\mathbf{z}_t)\right]=\sum_{j=1}^{M}\ln\left(\frac{1}{\pi\sigma^2}\exp\{-\frac{1}{\sigma^2}(y_t^j)^2\}\right), \qquad (17)$$

where σ^2 is the diagonal entry of covariance matrix. The transformation bring us the benefit of multi-antennas, because (9) refer $\mathbf{y}_t$ as a single sample while (17) regard it as M samples.

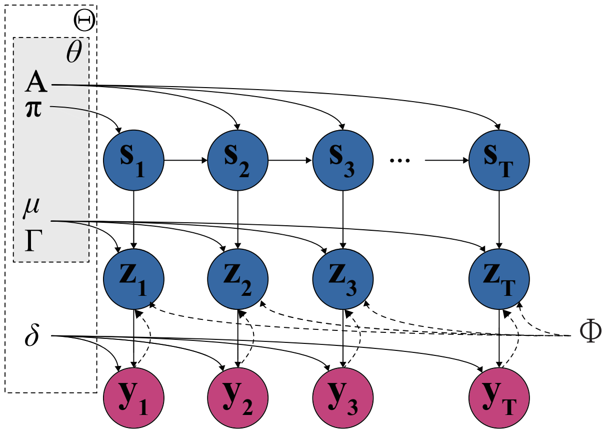


Fig. 4. Graphical model of VSCM where $\{\theta, \Phi, \delta\}$ is the model parameter set.

Considering the noise and nondeterminacy, a extra latent variable space $\mathcal{S}$=$\{\mathbf{S}_n\}$ is introduced into the graphical model as in Fig. 4. Inspired by VAE, the latent variable $\mathbf{z}_t$ could be considered as a sample from a state specific Normal distribution instead of a standard Normal distribution as in Fig. 5, expressed as:

$$p(\mathbf{z}_t|\mathbf{s}_t;\theta)=\mathcal{N}(\mathbf{z}_t;\mu_{\mathbf{s}_t},\Gamma_{\mathbf{s}_t}), \qquad (18)$$

where θ is graphical model parameters. The similar structure has been proposed in [23]. Considering joint distribution of length T with observations $\mathbf{Y}$=$[\mathbf{y}_1,\cdots,\mathbf{y}_T]$ and latent variables $\mathbf{Z}$=$[\mathbf{z}_1,\cdots,\mathbf{z}_T]$ and $\mathbf{S}$=$[\mathbf{s}_1,\cdots,\mathbf{s}_T]$ can get

$$\ln p(\mathbf{Y}|\mathbf{Z};\theta)=\sum_{t=1}^{T}\ln p(\mathbf{y}_t|\mathbf{z}_t;\delta), \qquad (19)$$

$$\ln p(\mathbf{Z}|\mathbf{S};\theta)=\sum_{t=1}^{T}\ln p(\mathbf{z}_t|\mathbf{s}_t;\theta), \qquad (20)$$

$$\ln p(\mathbf{S};\theta)=\sum_{t=1}^{T}\ln p(\mathbf{s}_t|\mathbf{s}_{t-1};\theta), \qquad (21)$$

where δ is the parameters in (17).

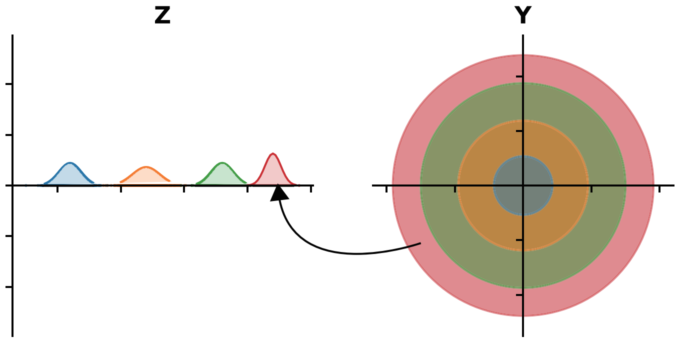


Fig. 5. Generative distributions $p(\mathbf{z}_t|\mathbf{s}_t;\theta)$ and $p(\mathbf{y}_t|\mathbf{z}_t;\delta)$

C. S-CSI estimation

In order to find the optimal path of the hidden state sequence $\mathbf{S}$, Viterbi algorithm is an appropriate method to apply where the posterior distribution $p(\mathbf{S})$ required to be known. While the exact inference of $p(\mathbf{S})$ is approximated by $q(\mathbf{S})$ due to the intractability. General Expectation Maximization (EM) algorithm can be performed to acquire the optimal $q(\mathbf{S})$ which is described as follows in detail:

1) E-step: The objective in E-step is maximizing the marginal loglikelihood in (11). Namely, maximizing the lower bound in (13). Note that (13) can also br written as

$$\mathcal{L}(\mathbf{Y}_n)=\int q(\mathbf{Z}_n|\mathbf{Y}_n)\ln\left\{\frac{p(\mathbf{Z}_n,\mathbf{Y}_n)}{q(\mathbf{Z}_n|\mathbf{Y}_n)}\right\}d\mathbf{Z}_n. \qquad (22)$$

Supposing that the joint distribution of latent variables could be factorized as

$$q(\mathbf{S},\mathbf{Z}|\mathbf{Y};\theta,\Phi)=q(\mathbf{S};\theta)q(\mathbf{Z}|\mathbf{Y};\Phi) \qquad (23)$$

which is known as the mean field approximation. Substituting latent variable $\mathbf{Z}$=$\{\mathbf{S},\mathbf{Z}\}$ in (22) can obtain

$$\mathcal{L}(\mathbf{Y};\Theta,\Phi)=\int q(\mathbf{S},\mathbf{Z}|\mathbf{Y};\theta,\Phi)\ln\left\{\frac{p(\mathbf{S},\mathbf{Z},\mathbf{Y};\Theta,\Phi)}{q(\mathbf{S},\mathbf{Z}|\mathbf{Y};\theta,\Phi)}\right\}d\mathbf{Z}d\mathbf{S}. \qquad (24)$$

Decomposing (24) can get

$$\begin{aligned}\mathcal{L}(\mathbf{Y};\Theta,\Phi)=&\int q(\mathbf{S};\theta)\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\ln\left[p(\mathbf{S},\mathbf{Z},\mathbf{Y};\Theta,\Phi)\right]d\mathbf{S}\\&-\int q(\mathbf{S};\theta)\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\ln\left[q(\mathbf{S};\theta)\right]d\mathbf{S}\\&-\int q(\mathbf{S};\theta)\mathbb{E}\left[\ln q(\mathbf{Z}|\mathbf{Y};\Phi)\right]d\mathbf{S},\end{aligned} \qquad (25)$$

where $q(\mathbf{Z}|\mathbf{Y};\Phi)$ is provided by the probabilistic encoder neural network. Consequently, the last term of (25) is a constant. Observing that (25) is actually the negative Kullback-Leibler divergence between $q(\mathbf{S})$ and $\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y})}\ln\left[p(\mathbf{S},\mathbf{Z},\mathbf{Y})\right]$. Thus maximizing the marginal loglikelihood in (11) is equivalent to minimizing the Kullback-Leibler divergence, and the minimum occurs when

$$q(\mathbf{S};\theta)=\frac{\exp\left(\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\ln\left[p(\mathbf{S},\mathbf{Z},\mathbf{Y};\Theta,\Phi)\right]\right)}{C_1}, \tag{26}$$

where C_1 is a normalizing constant. Hence that (26) can be written as

$$\ln q(\mathbf{S};\theta){=}\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\left[\ln p(\mathbf{S},\mathbf{Z},\mathbf{Y};\Theta,\Phi)\right]{+}C_1, \tag{27}$$

Decomposing $p(\mathbf{S},\mathbf{Z},\mathbf{Y};\Theta,\Phi)$, the right of (27) becomes:

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\left[\ln p(\mathbf{S},\mathbf{Z},\mathbf{Y};\Theta,\Phi)\right]{+}C_1 \qquad (28)\\ {=}&\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\left[\ln p(\mathbf{Z}|\mathbf{S};\theta){+}\ln p(\mathbf{S};\theta)\right]{+}C_2.\end{aligned}$$

Notice that $\ln p(\mathbf{Y}|\mathbf{Z};\delta)$ is a irrelevant constant term integrated into C_2. The marginal of latent variables $\mathbf{S}$ then can be expressed as:

$$\ln q(\mathbf{S};\theta){=}\sum_{t=1}^{L}\left(b(\mathbf{s}_t){+}\ln p(\mathbf{s}_t|s_{t-1};\theta)\right){+}C_2, \tag{29}$$

with $b(\mathbf{s}_t){=}\mathbb{E}_{q(\mathbf{z}_t|\mathbf{y}_t;\Phi)}\left[\ln p(\mathbf{z}_t|\mathbf{s}_t;\theta)\right]$. Knowing the approximate latent marginal distribution, it's easy to find the best path through Viterbi algorithm.

2) M-step: With the marginal of latent variables $\mathbf{S}$ determined, the goal of M-step is maximizing the lower bound $\mathcal{L}(\mathbf{Y}_n)$ with respect to its parameters $\{\Theta,\Phi\}$.

(24) can be written as

$$\mathcal{L}(\mathbf{Y};\Theta,\Phi){=}\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)q(\mathbf{S};\theta)}\left[\ln\frac{p(\mathbf{Y}|\mathbf{Z};\delta)p(\mathbf{Z}|\mathbf{S};\theta)p(\mathbf{S};\theta)}{q(\mathbf{Z}|\mathbf{Y};\Phi)q(\mathbf{S};\theta)}\right], \tag{30}$$

Knowing that

$$\mathcal{L}(\mathbf{Y};\Theta,\Phi){=}\sum_{t=1}^{T}\mathcal{L}(\mathbf{y}_t;\Theta,\Phi), \tag{31}$$

for each time t has

$$\begin{aligned}\mathcal{L}(\mathbf{y}_t;\Theta,\Phi){=}&\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\left[\ln p(\mathbf{y}_t|\mathbf{z}_t;\delta)\right]\\ &-\mathbb{E}_{\hat{q}(\mathbf{s}_t)}\left[\mathrm{KL}\left(q(\mathbf{z}_t|\mathbf{y}_t;\Phi)||p(\mathbf{z}_t|\mathbf{s}_t;\theta)\right)\right]\\ &+\mathbb{E}_{\hat{q}(\mathbf{s}_{t-1},\mathbf{s}_t)}\left[\ln p(\mathbf{s}_t|\mathbf{s}_{t-1};\theta)\right]{+}C_3,\end{aligned} \tag{32}$$

where $C_3=-\mathbb{E}_{q(\mathbf{S};\theta)}\left[\ln q(\mathbf{S};\theta)\right]$. The expectation can approximated by

$$\mathbb{E}_{q(\mathbf{Z}|\mathbf{Y};\Phi)}\left[\ln p(\mathbf{y}_t|\mathbf{z}_t;\delta)\right]\approx\frac{1}{L}\sum_{l=1}^{L}\ln p(\mathbf{y}_t|\hat{\mathbf{z}}_t^{(l)};\delta), \tag{33}$$

where $\hat{\mathbf{z}}_t^{(l)}\sim q(\mathbf{Z}|\mathbf{Y};\Phi)$ is obtained through sampling. The marginal distributions $\hat{q}(\mathbf{s}_t)$ and $\hat{q}(\mathbf{s}_{t-1},\mathbf{s}_t)$ can be given recursively using Forward-Backward algorithm. Note that the calculation of the expectation what is the second term of (31) is intractable when the number of states is large. Hence that the sampling approach can also be performed to approximate the expectation:

$$\begin{aligned}&\mathbb{E}_{\hat{q}(\mathbf{s}_t)}\left[\mathrm{KL}\left(q(\mathbf{s}_t;\Phi)||p(\mathbf{z}_t|\mathbf{s}_t;\theta)\right)\right]\\ \approx&\frac{1}{N}\sum_{n=1}^{N}\mathrm{KL}\left(q(\mathbf{s}_t;\Phi)||p(\mathbf{z}_t|\mathbf{s}_t^{(n)};\theta)\right)\end{aligned} \tag{34}$$

with $\mathbf{s}_t^{(n)}\sim\hat{q}(\mathbf{s}_t)$. Combining with parameters of HMM defined earlier, a more elaborate form of $\mathcal{L}(\mathbf{y}_t;\Theta,\Phi)$ can be written as:

$$\begin{aligned}&\mathcal{L}(\mathbf{y}_t;\Theta,\Phi)\\ \approx&\ln p(\mathbf{y}_t|\hat{\mathbf{z}}_t;\delta){+}H\left(q(\mathbf{z}_t;\Phi)\right)\\ &+\sum_{k=1}^{K}\gamma_k(t)\mathbb{E}_{q(\mathbf{z}_t;\Phi)}\left[\ln p(\mathbf{z}_t|\mathbf{s}_t;\theta)\right]\\ &+\sum_{k=1}^{K}\gamma_k(t)\ln\pi_k{+}\sum_{k=1}^{K}\sum_{j=1}^{K}\ln a_{jk}\sum_{t=2}^{T}\xi_{jk}(t)\\ &+C_3,\end{aligned} \tag{35}$$

where $H(\cdot)$ is the entropy. The loss function $-\mathcal{L}(\mathbf{Y};\Theta,\Phi)$ are trained through neural networks with respect to all parameters $\{\Theta,\Phi\}$. Alternately performing the E-step and M-step unitl the loss function converge.

IV. State ambiguity and power allocation optimization

As we derived in Eq. (7), the relationship between $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t{}^{\mathrm{H}}]$ and $\mathbf{z}_t{=}\mathcal{H}_i$ could be explicitly computed by the linear mapping $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t{}^{\mathrm{H}}]{=}\left(\langle\mathbf{P},\mathbf{z}_t\rangle{+}\sigma_n^2\right)\mathbf{I}_M$. If the mapping is not bijection, a degenerate phenomenon would be caused, and the phenomenon is referred as state ambiguity. More specifically, considering $\mathcal{H}_1=[0.5,1]^{\mathrm{T}}$ and $\mathcal{H}_2=[1,0.5]^{\mathrm{T}}$, the inner product $\langle\mathbf{P},\mathcal{H}_1\rangle{=}\langle\mathbf{P},\mathcal{H}_2\rangle$ if $\mathbf{P}{=}\mathbf{1}$. In order to avoid such condition, a proper power allocation vector should be chosen.

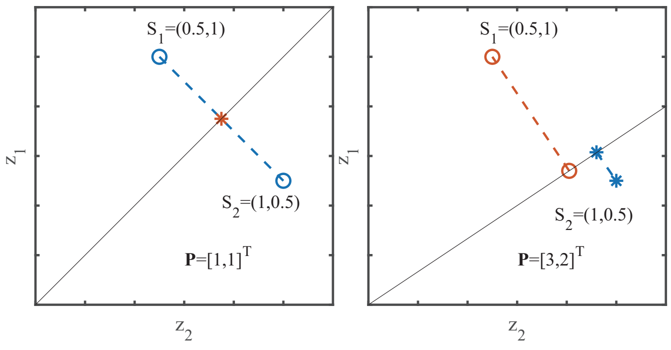


Fig. 6. A projection optimization example to mitigate state ambiguity

The optimization goal is to find a best power allocation vector whose inner product with all the elements in $\mathcal{H}$ are not equal. Inner product could be considered as vector projection, that is to say the length of all points in $\mathcal{H}$ that project to power allocation vector $\mathbf{P}$ should be different, which is illustrated in figure 6. In other words, the projected points' distance should be greater than zero for any two $\mathcal{H}_i$, $\mathcal{H}_j$. The projected distance between $\mathcal{H}_i$ and $\mathcal{H}_j$ is

$$\begin{aligned} &\Delta\big(\mathcal{H}_i,\mathcal{H}_j|\mathbf{P}\big) = \mathrm{Tr}(\mathbf{P}^{\mathrm{T}}\mathbf{D}_{ij}\mathbf{P}) \\ \text{s.t}\quad &\mathbf{P}^{\mathrm{T}}\mathbf{P}=1 \\ &\mathbf{P}\succ\mathbf{0}, \end{aligned} \tag{36}$$

where we refer the matrix $\mathbf{D}_{ij} = \mathcal{H}_i\mathcal{H}_j^{\mathrm{T}}$ as the distance matrix between $\mathcal{H}_i$ and $\mathcal{H}_j$. $\mathbf{P}\succ\mathbf{0}$ represents that all elements of $\mathbf{P}$ are greater than zero. Our problem turns out to maximize the minimum $\Delta\big(\mathcal{H}_i,\mathcal{H}_j|\mathbf{P}\big)$, which is :

$$\begin{aligned} \max\quad \min\quad &\Delta\big(\mathcal{H}_i,\mathcal{H}_j|\mathbf{P}\big),\quad 1\leq i<j\leq K \\ \text{s.t}\quad &\mathbf{P}^{\mathrm{T}}\mathbf{P}=1 \\ &\mathbf{P}\succ\mathbf{0}. \end{aligned} \tag{37}$$

Theorem 1. *$M\succeq 0$ has rank 1 if and only if $M=\beta\beta^{\mathrm{T}}$ for some nonzero vector $\beta\in\mathbb{R}^n$.*

Proof. For sufficient condition, $M\succeq 0$ has rank 1 means M can be decomposed as

$$M=Q\Lambda Q^{\mathrm{T}}=\sum_{i=1}^{n}\lambda_i\beta_i\beta_i^{\mathrm{T}},$$

where Λ is a diagonal matrix and λ_i represents its i-th element of main diagonal, and β_i is the i-th column vector of matrix Q. Let λ_1 be the only entry that greater than 0, and $\beta=\sqrt{\lambda_1}\beta_1$, then $M=\beta\beta^{\mathrm{T}}$.

For necessary condition, $M=\beta\beta^{\mathrm{T}}$ means M is a symmetric matrix obviously. We only need to prove rank(M)=1. Let $\beta=[\beta_1,\cdots,\beta_n]^{\mathrm{T}}$, then $M=[\beta_1\beta^{\mathrm{T}},\cdots,\beta_n\beta^{\mathrm{T}}]$, which means rank($M$)=1. □

The nonsmooth objective function in (37) is intractable, we have to transform it into a convex one by introducing a new variable $\mathbf{X}=\mathbf{P}\mathbf{P}^{\mathrm{T}}$ and an auxiliary variable d, and the trick has been used in [26]. It turns to be :

$$\begin{aligned} \max\quad &d \\ \text{s.t}\quad &\mathrm{Tr}(\mathbf{D}_{ij}\mathbf{X})\geq d \\ &\mathrm{Tr}(\mathbf{X}) = 1 \\ &\mathrm{Tr}(\mathbf{e}_i\mathbf{e}_j\mathbf{X})-|\varepsilon|\geq 0 \\ &\text{rank}(\mathbf{X})=1 \\ &\mathbf{X}\succeq 0 \\ &1\leq i<j\leq K. \end{aligned} \tag{38}$$

In (38), $\mathbf{e}_i$ denotes vector with i-th element 1 and others are 0, ε is infinitesimal number, and $\mathbf{X}\succeq 0$ means $\mathbf{X}$ is a semidefinite positive matrix.

Taking a closer look at (38), a standard semidefinite programming problem with the rank constraint can be recognized, Without the rank constraint, it could be solved through mature method like interior-point method. While the non-convex rank constraint turns it into a NP-hard problem. Readers can refer to [27] for interior-point method and transformation from (38) without rank constraint to a standard form by introducing slack variables, which is :

$$\begin{aligned} \min\quad &\mathrm{Tr}(C\overline{X}) \\ \text{s.t}\quad &\mathrm{Tr}(E_i\overline{X})=b_i \\ &\overline{X}\succeq 0 \end{aligned} \tag{39}$$

Let $\mathfrak{L}(\overline{X},\lambda,S)$ denotes the Lagrangian function of semidefinite program (39), i.e.,

$$\begin{aligned} &\mathfrak{L}(\overline{X},\lambda,S) \\ &=\mathrm{Tr}(C\overline{X})+\sum_i\lambda_i[\mathrm{Tr}(E_i\overline{X})-b_i]-\mathrm{Tr}(S\overline{X}). \end{aligned} \tag{40}$$

Lemma 1. *A symmetric matrix $\overline{X}\in\mathbb{S}^n$ is the optimal feasible solution to (39) if and only if there exist a symmetric matrix $S\in\mathbb{S}^n$ such that the following holds,*

$$\begin{aligned} \mathrm{Tr}(E_i\overline{X})&=b_i \\ \overline{X}&\succeq 0 \\ S&\succeq 0 \\ \mathrm{Tr}(S\overline{X})&=0 \\ C+\sum_i\lambda_iE_i&=S \end{aligned} \tag{41}$$

Proof. These are the first order KKT conditions, which means primal feasible (the first two), dual feasible, complementary slackness and gradient $\nabla_{\overline{X}}\mathfrak{L}=0$ respectively. □

As we noted in theorem 1, rank constraint could be embodied by factorizing the optimization variable $\overline{X}$ as $\overline{X}=RR^{\mathrm{T}}$, and (38) becomes:

$$\begin{aligned} \min\quad &\mathrm{Tr}(CRR^{\mathrm{T}}) \\ \text{s.t}\quad &\mathrm{Tr}(E_iRR^{\mathrm{T}})=b_i. \end{aligned} \tag{42}$$

The Lagrangian function becomes:

$$\begin{aligned} &\mathfrak{L}(R,\lambda) \\ &=\mathrm{Tr}(CRR^{\mathrm{T}})+\sum_i\lambda_i[\mathrm{Tr}(E_iRR^{\mathrm{T}})-b_i]. \end{aligned} \tag{43}$$

Lemma 2. *If there exists a vector $R^*\in\mathbb{R}^n$ and its correspond Lagrangian multiplier vector $\lambda^*\in\mathbb{R}^m$ satisfies the second order optimality sufficient conditions of (43) which are the follows, satisfies the first KKT condition (41) automatically.*

$$\begin{aligned} \nabla_R\mathfrak{L}(R^*,\lambda^*)&=0 \\ \nabla_\lambda\mathfrak{L}(R^*,\lambda^*)&=0 \\ \nabla^2_{RR}\mathfrak{L}(R^*,\lambda^*)&\succeq 0, \end{aligned} \tag{44}$$

where ∇^2 denotes the second order gradient, i.e. its Hessian matrix.

Proof. Lagrangian function (43) can be written as:

$$\mathfrak{L}(R,\lambda)=Tr[(C+\sum_i\lambda_iE_i)RR^{\mathrm{T}}]-\sum_i\lambda_ib_i.$$

Let $S=(C+\sum_i\lambda_iE_i)$ and consider the following easily derived formulas:

$$\begin{aligned} \nabla_R\mathfrak{L}(R,\lambda)&=2SR \\ \nabla_\lambda\mathfrak{L}(R,\lambda)&=\mathrm{Tr}(E_iRR^{\mathrm{T}})-b_i \\ \nabla^2_{RR}\mathfrak{L}(R,\lambda)&=S \end{aligned} \tag{45}$$

Combining these formulas with (44), it's obvious that who satisfies (44) satisfies (41). □

A good candidate method for solving (44) is augmented Lagrangian method, the same method used in [28] for solving semidefinite programs via low-rank factorization. In this paper we choose augmented Lagrangian method while readers can still choose other methods like alternating direction method of multipliers (ADMM).

```
Algorithm 1: Power allocation vector optimization
  Input: Markov finite set $\mathcal{H}$
  Output: approximate optimal vector $R$
  decide the opimization problom in a standard form like (42);
  initialize $\lambda^1$, $\alpha^1$, $v^1$, $\rho$, $\chi$, and $R^1$;
  $k$=1;
  repeat :
      $v = \sum_i [\mathrm{Tr}(E_i(R^k)(R^k)^{\mathrm{T}}) - b_i]^2$;
      if $v < \chi v^k$ :
          $\lambda_i^{k+1} = \lambda_i^k - \alpha^k [\mathrm{Tr}(E_i(R^k)(R^k)^{\mathrm{T}}) - b_i]$ for all $i$ ;
          $\alpha^{k+1} = \alpha^k$ ;
          $v^{k+1} = v$;
      end
      else :
          $\lambda_i^{k+1} = \lambda_i^k$ for all $i$;
          $\alpha^{k+1} = \rho\alpha^k$ ;
          $v^{k+1} = v^k$;
      end
      search $R^{k+1} = \arg\min_R \mathfrak{L}(R^k, \lambda^k, \alpha^k)$;
      $k = k+1$;
  until : converge;
```

The augmented Lagrangian function of (42) is:

$$\mathfrak{L}(R,\lambda,\alpha) = \mathrm{Tr}(CRR^{\mathrm{T}}) + \sum_i \lambda_i [\mathrm{Tr}(E_i RR^{\mathrm{T}}) - b_i] + \frac{\alpha}{2} \sum_i [\mathrm{Tr}(E_i RR^{\mathrm{T}}) - b_i]^2, \tag{46}$$

where α is the penalty parameter. When implementing the augmented Lagrangian method, parameters $\rho > 1$ and $\chi < 1$ need to be given and an auxiliary scalar v^k should be introduced, the choices we make for ρ and χ are 5 and $1/4$. Besides, a detail that the initialization of R can not contain zero at all should be noticed. More detailed dual ascend implementation is given as Algorithm 1. In order to make sure that the approximate optimal vector R satisfies the second order optimality sufficient conditions, Newton's method should be chosen in Algorithm 1 for searching $\arg\min_R \mathfrak{L}(R^k, \lambda^k, \alpha^k)$ where we choose the limited memory BFGS approach employs the strong Wolfe conditions. More specific details could refer to chapter 3 and chapter 7.2 of [29].

V. SIMULATION RESULTS

For the evaluation of the proposed model, several different simulations are performed. Binary phase shift keying modulation is used in the simulation for the reason only second order statistical knowledge is required. The initial distribution vector $\boldsymbol{\pi}$ and the transition matrix $\mathbf{A}$ are randomly initialized. Monte Carlo simulations are performed to reduce the randomness whose flowchart is abbreviated in the Fig.7 .

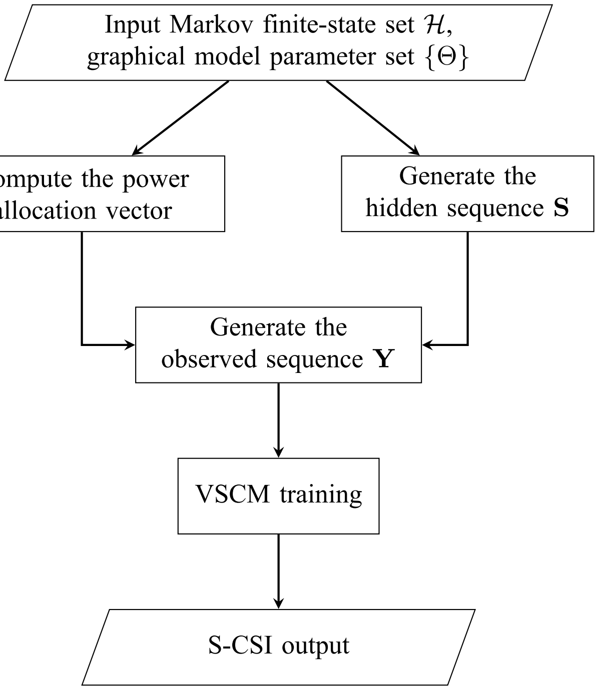

Input Markov finite-state set $\mathcal{H}$,
graphical model parameter set $\{\Theta\}$
Compute the power allocation vector
Generate the hidden sequence $\mathbf{S}$
Generate the observed sequence $\mathbf{Y}$
VSCM training
S-CSI output

Fig. 7. Simulation flowchart

The performance indicator is the accuracy rate P_{acc}, given as:

$$P_{\text{acc}} = \frac{\sum_{n=1}^{\zeta} \sum_{t=1}^{\mathrm{T}} \mathrm{Sign}(\hat{\mathbf{s}}_t - \mathbf{s}_t)}{\zeta}, \tag{47}$$

where ζ represents the number of independent Monte Carlo simulation trials, and we choose ζ=1000 in this paper. Sign function is defined as:

$$\mathrm{Sign}(x) = \begin{cases} 1, x=0, \\ 0, x \neq 0. \end{cases} \tag{48}$$

A. Performance of VSCM

TABLE II
System parameters in simulations

Setup	SCSS	UE number
1	{0.5,3}	2
2	{0.5,3,5}	2
3	{0.5,3,5}	3

This subsection will present the effect of different numbers of antenna, SNR, data sequence length, and state set cardinality on our proposed model. The system parameters are listed in the tableII where the initial state distribution and transition matrix are randomly set.

Firstly, setup 1 is performed which is the simplest case of all 3 to check the feasibility of VSCM where the length of the observation is 100. The result is shown in the Fig. 8. Some easy conclusions can be drawn, which are: 1) increasing

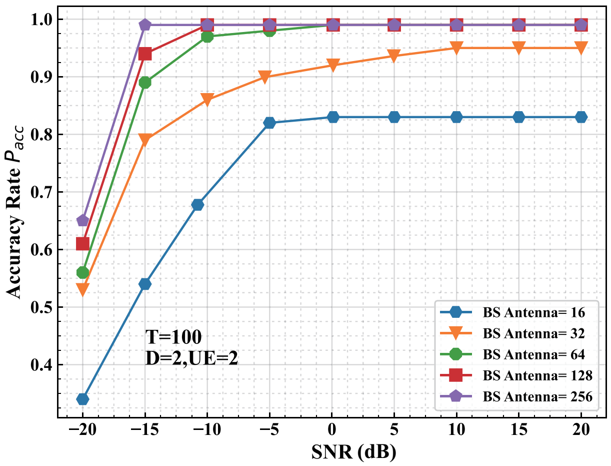


Fig. 8. Effect of SNR on accuracy rate when D=2, UE number=2 and $\mathbf{T}$=100

the number of BS antennas could increase the accuracy rate. This is what we expect because more antennas mean more samples from the same distribution which would reduce the randomness. While the effect is less obvious when increased to 128. 2) The estimation accuracy improves as the SNR increases until meeting the bottleneck where SNR=10 dB.

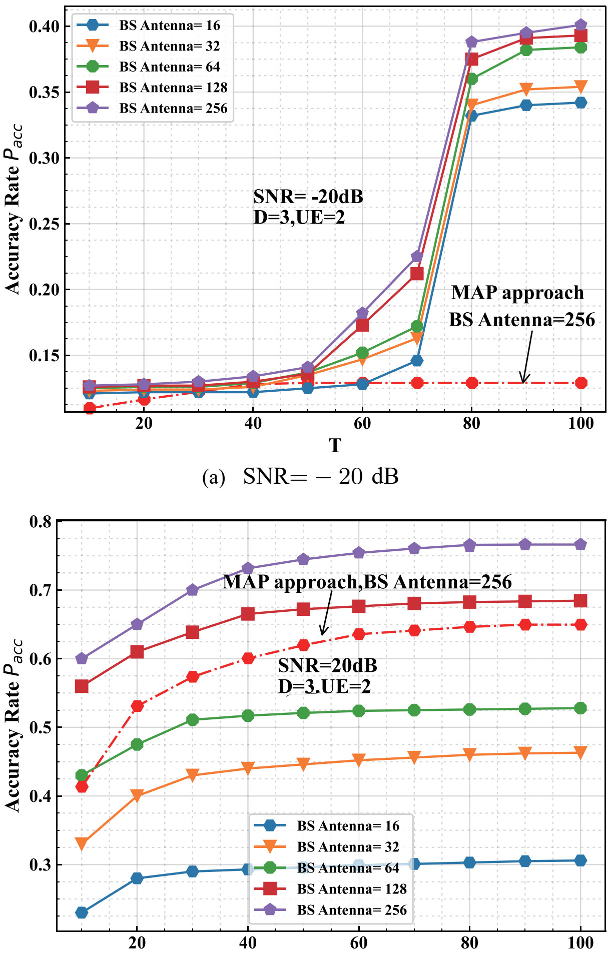
Secondly, Fig. 9 presents the impact of the sequence length under setup 2. As can be seen, longer sequence would provide more information to estimate the hidden sequence. In Fig. 9(b), the accuracy rate increase rapidly when T$\leq$ 40 under great SNR condition. While Fig. 9(a) shows an entirely different picture, the accuracy grows until T$\geq$ 70. Note that the transition matrix contains 81 parameters need to be estimated when D=3 and UE=2. If the provided data sequence is not long enough, severe model degradation will occur. The reason why we still get an acceptable result under high SNR condition is the prior knowledge about the elements in SCSS is obtaied, which or else need long enough data sequence to estimate its elements and gaurantee the degenerate phenomenon would not happen. Meanwhile, Fig. 9 concludes the MAP approach results, where estimating the S-CSI only by the prior knowledge of the SCS set, revealing the MAP approach malfunctions under low SNR condition and worse than our approach under high one.

Next, given enough data sequence, we deliberate the effect of sequence length on accuracy rate when D=3, UE=3 which is presented in the Fig. 10. The performance bottleneck occurs when sequence length T≈1400 which is nearly twice the number of parameters to be estimated. Besides, the performance improvement is not significant when the number of antennas increased from 128 to 256.

On balance, it is better to provide twice the sequence length of the number of parameters to be estimated to avoid degeneration, especially under low SNR condition. Furthermore, setting the number of antennas as 128 is an acceptable choice for balance between accuracy and computational complexity.

B. Comparison between VSCM and HMM

The comparison between HMM and VSCM is presented in this subsection. The simulations are carried out using Setup 2



(a) SNR= − 20 dB

(b) SNR=20 dB

Fig. 9. Effect of sequence length on accuracy rate when D=3, UE=2

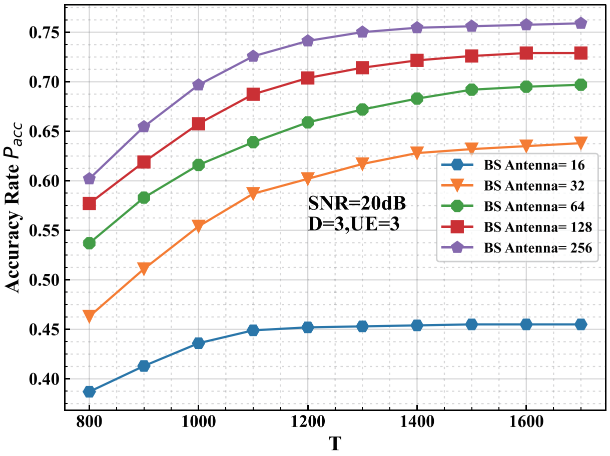


Fig. 10. Effect of sequence length on accuracy rate when D=3, UE=3

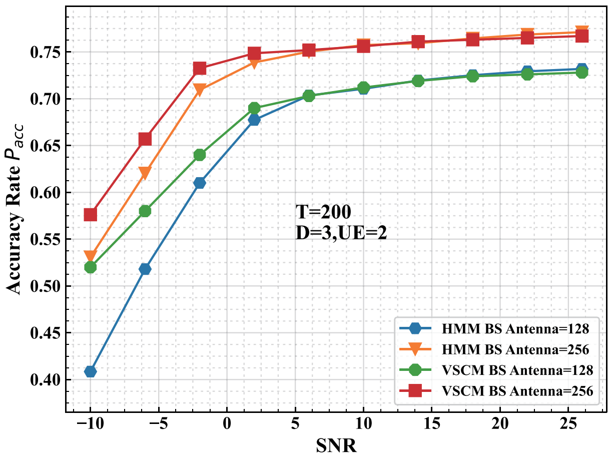


Fig. 11. Comparison between HMM and VSCM when D=3, UE=2, and T=200

and the results are shown in Fig. 11. As we can see, VSCM performs better under low SNR condition while worse under high situation. This phenomenon occurs is because VAE brings Gaussian randomness into the model which is reflected in the fact that the loss function contains two adversarial terms. Under high SNR ocndition, VSCM is still stochastic. In other words, if we let Γ=0 then VSCM degenerates to normal HMM.

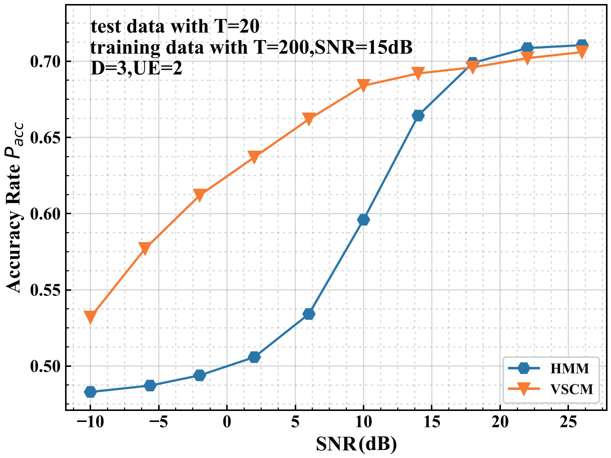


Fig. 12. Estimating S-CSI using well trained model under different SNR condition

In addition, we use the well trained model to estimate S-CSI from new received signals under different SNR condition. The model is set with Setup 2 and training data sequence length T=200 under the circumstance that SNR=15 dB to ensure the model can get steady performance in the training period. The results are present in Fig. 12. Obviously, both HMM and VSCM can not get the same performance as that in the training period. HMM performs bad under low SNR condition and grows rapidly from SNR=5dB to SNR=15dB, while VSCM is more robust and performs better under low SNR condition by contrast, which is a foregone conclusion due to the fact that VSCM is a generative model contains noise part and is more generalized.

C. Further discussion: Expand the solution space

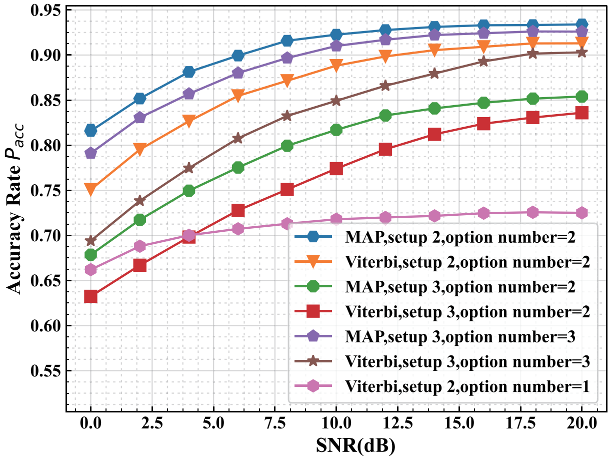


Fig. 13. Results after expanding the solution space.

Our approach can not be applied to the practical system directly due to the impassable estimation results that the accuracy rate of more than 90% cannot be achieved. The number of Markov states, which increase exponentially with the number of states and the number of users, limits the performance of our algorithm. Even though the power allocation optimization procedure is performed preceding the estimation, the minimum distance in the projected space is still small with the case that the cardinality K of MSS $\{\mathcal{H}\}$ is large.

Expanding the solution space by introducing the second best estimation sequence or even the third best one to provide a reliable solution set can give credible information to subsequent system design, for concreteness, transmission scheme. Hence the simulation results of expanding the solution set are listed in Fig.13 under the condition that SNR=20dB , M=128, T=200 and 1800 for setup 2 and 3 respectively. Two decoding algorithms, maximum a posteriori (MAP) and Viterbi, are applied. Choosing the first few hidden states who have the highest MAP probability at each time t, i.e. $\gamma_k(t)$, for MAP and the first few optimal paths for Viterbi to expand the solution space. The option number in Fig.13 denotes the number of states or paths. ACC is given as :

$$P_{\text{acc}}{=}\frac{\sum_{n=1}^{\zeta}\sum_{t=1}^{\text{T}}\overline{\text{Sign}}(\hat{\text{s}}_t)}{\zeta}, \tag{49}$$

with

$$\overline{\text{Sign}}(x){=}\begin{cases}1, x{\in}\text{solution set},\\ 0, x{\notin}\text{solution set}.\end{cases} \tag{50}$$

Results demonstrate that our approach can provide a reliable solution set with accuracy rate more than 90% after expanding the solution space, which is useful for downstream system design. Besides, MAP decode algorithm will achieve better performance due to its lager solution set.

VI. Conclusion

In this paper, we consider a Spatio-temporal double non-stationary massive MIMO channel system and try to estimate the S-CSI of the non-stationary channel. First, the channel system is modeled as an extended VAE to capture correlations of received signals in our channel system by incorporating the structure of HMM called VSCM. We have derived the covariance matrix of received signals is spherical, and the relationships between the diagonal elements and the S-CSI is projection under the power allocation vector. Second, in order to ensure that the projection mapping is bijection which avoids the degeneration of HMM, we abstract the problem as a power allocation optimization problem. Third, we provide a heuristic solution to the optimization problem which is semidefinite programming with non-convex rank constraint. Then we have derived an EM-like algorithm to optimize the loss function and inference the latent variable sequence. Last, several simulations have been performed to illustrate the effectiveness of VSCM. we find that VSCM performs better under low SNR condition which means more robust and it has more power of generalization compared to traditional HMM.

References

[1] Thomas L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 9(11):3590–3600, 2010.

[2] Erik G. Larsson, Ove Edfors, Fredrik Tufvesson, and Thomas L. Marzetta. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, 2014.

[3] Hien Quoc Ngo, Erik G. Larsson, and Thomas L. Marzetta. Energy and spectral efficiency of very large multiuser MIMO systems. *IEEE Transactions on Communications*, 61(4):1436–1449, 2013.

[4] Zhitan Zheng, Chengzhi Zhu, Bin Jiang, Wen Zhong, and Xiqi Gao. Statistical channel state information acquisition for massive MIMO communications. *2015 International Conference on Wireless Communications and Signal Processing, WCSP 2015*, pages 1–5, 2015.

[5] Chen Sun, Xiqi Gao, Shi Jin, Michail Matthaiou, Zhi Ding, and Chengshan Xiao. Beam division multiple access for massive MIMO downlink transmission. *IEEE International Conference on Communications*, 2015-Septe(6):1970–1975, 2015.

[6] Chao Kai Wen, Shi Jin, Kai Kit Wong, Jung Chieh Chen, and Pangan Ting. Channel Estimation for Massive MIMO Using Gaussian-Mixture Bayesian Learning. *IEEE Transactions on Wireless Communications*, 14(3):1356–1368, 2015.

[7] Jeremy P. Vila and Philip Schniter. Expectation-Maximization gaussian-mixture approximate message passing. *IEEE Transactions on Signal Processing*, 61(19), 2013.

[8] Xiang Gao, Fredrik Tufvesson, Ove Edfors, and Fredrik Rusek. Measured propagation characteristics for very-large MIMO at 2.6 GHz. In *Conference Record - Asilomar Conference on Signals, Systems and Computers*, pages 295–299, 2012.

[9] Yu Liu, Cheng Xiang Wang, Jie Huang, Jian Sun, and Wensheng Zhang. Novel 3-D Nonstationary MmWave Massive MIMO Channel Models for 5G High-Speed Train Wireless Communications. *IEEE Transactions on Vehicular Technology*, 68(3):2077–2086, 2019.

[10] Shangbin Wu, Cheng Xiang Wang, El Hadi M. Aggoune, Mohammed M. Alwakeel, and Xiaohu You. A General 3-D Non-Stationary 5G Wireless Channel Model. *IEEE Transactions on Communications*, 66(7):3065–3078, 2018.

[11] Siyu Lin, Zhangdui Zhong, Lin Cai, and Yuanqian Luo. Finite state Markov modelling for high speed railway wireless communication channel. *GLOBECOM - IEEE Global Telecommunications Conference*, pages 5421–5426, 2012.

[12] Ammar Ghazal, Cheng Xiang Wang, Bo Ai, Dongfeng Yuan, and Harald Haas. A Nonstationary Wideband MIMO Channel Model for High-Mobility Intelligent Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, 2015.

[13] Guoliang Wang, Wei Peng, Dong Li, Tao Jiang, and Fumiyuki Adachi. Statistical CSI Acquisition in the Nonstationary Massive MIMO Environment. *IEEE Transactions on Vehicular Technology*, 67(8):7181–7190, 2018.

[14] Elisabeth De Carvalho, Anum Ali, Abolfazl Amiri, Marko Angjelichinoski, and Robert W. Heath. Non-Stationarities in Extra-Large-Scale Massive MIMO. *IEEE Wireless Communications*, 2020.

[15] Anum Ali, Elisabeth De Carvalho, and Robert W. Heath. Linear Receivers in Non-Stationary Massive MIMO Channels with Visibility Regions. *IEEE Wireless Communications Letters*, 8(3), 2019.

[16] Thomas Zwick, Christian Fischer, and Werner Wiesbeck. A stochastic multipath channel model including path directions for indoor environments. *IEEE Journal on Selected Areas in Communications*, 2002.

[17] Fulvio Babich. A markov model for the mobile propagation channel. *IEEE Transactions on Vehicular Technology*, 2000.

[18] Luc Le Magoarou, Antoine Le Calvez, and Stephane Paquelet. Massive MIMO Channel Estimation taking into account spherical waves. In *IEEE Workshop on Signal Processing Advances in Wireless Communications, SPAWC*, 2019.

[19] Zhou Zhou, Xiang Gao, Jun Fang, and Zhi Chen. Spherical wave channel and analysis for large linear array in los conditions. In *2015 IEEE Globecom Workshops, GC Wkshps 2015 - Proceedings*, 2015.

[20] Xuefeng Yin, Stephen Wang, Nan Zhang, and Bo Ai. Scatterer Localization Using Large-Scale Antenna Arrays Based on a Spherical Wave-Front Parametric Model. *IEEE Transactions on Wireless Communications*, 2017.

[21] Amos Lapidoth. *A foundation in digital communication*. 2017.

[22] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[23] Janek Ebbers, Jahn Heymann, Lukas Drude, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. Hidden Markov Model variational autoencoder for acoustic unit discovery. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017-August:488–492, 2017.

[24] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians, 2017.

[25] Ke M. Tran, Yonatan Bisk, Ashish Vaswani, Daniel Marcu, and Kevin Knight. Unsupervised Neural Hidden Markov Models. pages 63–71, 2016.

[26] Wei Bian and Dacheng Tao. Max-min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):1037–1050, 2011.

[27] Bernd Gärtner and Jiří Matoušek. *Approximation algorithms and semidefinite programming*. 2012.

[28] Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B*, 2003.

[29] Jorge Nocedal and Stephen J. Wright. *Numerical optimization 2nd edition*. 2000.