Joint Channel and Location Estimation of Massive MIMO System With Phase Noise

Xuanyu Zheng, *Student Member, IEEE*, An Liu, *Senior Member, IEEE*, and Vincent Lau, *Fellow, IEEE*

***Abstract*—Massive multiple-input multiple-output (MIMO) is a key technique in the fifth-generation (5G) networks for its attractive advantages in wireless communication. Besides, it also provides localization capability with high accuracy, yet the massive MIMO localization problem is not well addressed with the presence of some physical impairments in the circuits, such as phase noise (PHN). In this paper, we consider localization problem with PHN, which is a critical bottleneck in massive MIMO. Traditionally, existing methods exploit channel sparsity by sampling the space into discrete grids to estimate the position of the user in the compressive sensing framework. However, as the PHN affects the received signal in a nonlinear manner, it is difficult to obtain accurate basis vectors for sparse channel recovery. Besides, they often suffer from performance loss due to the location quantization error introduced by the fixed location grid. To deal with the aforementioned problems, we introduced a sparse representation model for the channel with dynamic-grid parameters to eliminate the location quantization error and derived an approximation for the likelihood with the presence of PHN. Based on these, we proposed an efficient algorithm for joint estimation of user location and sparse channel utilizing the majorization-minimization algorithm, which is shown to achieve higher localization accuracy than existing methods with the presence of PHN.**

***Index Terms*—Localization, massive multiple-input multiple-output (MIMO), compressive sensing, majorization-minimization (MM), dynamic-grid refinement.**

I. Introduction

MASSIVE multiple-input multiple-output (MIMO) will be widely adopted in the Fifth-Generation (5G) networks, in order to cope with the increasing demand in data request. There are lots of works studying how massive MIMO can enhance the capacity and link reliability of wireless networks [1], [2]. Exploiting the spatial degrees of freedom, massive MIMO can also enhance the resolution of localization of mobile users (MU).

A challenging problem in massive MIMO for communication is to estimate the channel state information (CSI) under limited pilot symbols. In time-division duplex (TDD) MIMO systems, CSI can be obtained via uplink pilot transmission by exploiting the channel reciprocity [3], [4], and the pilot overhead is high when the number of active MUs is large. In frequency-division duplex (FDD) mode, the pilot overhead increases proportionally with the number of antennas in the BSs, which poses great challenges for massive MIMO. Recently, some researchers exploited different channel sparsity to perform efficient channel estimation algorithms. For example, [5] proposed a CSI acquisition scheme with limited pilots by exploiting the spatially common sparsity of massive MIMO channels. When several MUs share some common scatterers, the channels are jointly sparse in the angular domain and several efficient algorithms was proposed to reduce the pilot overhead [6].

MIMO localization also receives significant attention as an important application of wireless systems. In some early works, a typical method called bearings-only target localization (BOTL), which harnesses the angle of arrival (AoA) estimation as key techniques, is studied in [7]–[9]. In BOTL, AoA are measured at all base stations first, then a maximum likelihood (ML) estimator is designed to estimate the coordinate of the target based on triangulation. However, before obtaining the location information, BOTL primarily requires accurate AoA measurements in all base stations, which is sub-optimal in nature. A direct localization approach was proposed in [10], [11] to estimate the MU position directly from data without any intermediate parameters. Nevertheless, the proposed 2-dimensional exhaustive search in the area is computationally difficult, especially when a high localization resolution is required. Moreover, the received signal is nonlinear with respect to the user location. Hence, the ML cost function or least square optimization is generally non-convex and the estimation problem suffers from many local minimum. A finger-printing based localization algorithm is proposed in [12] to locate multiple users offline with massive MIMO systems. However, such method requires collection of a large amount of training samples at different locations in a specific area, which may pose great challenges in practice.

Similar to the channel estimation, there are some works that considered exploiting sparsity to improve localization performance in wireless networks. By sampling the target area into high resolution grids, [13] approximates the coordinate of the target on a continuous map by a discrete-grid based solution and exploits target sparsity in the grids. Least absolute shrinkage and selection operator (LASSO) based solutions were proposed to find the most likely grid of the target. Nevertheless, there is usually location quantization error between the coordinates of the sample grids and the user location. A denser grid partition in the area can alleviate the location quantization error problem

Manuscript received October 9, 2019; revised March 11, 2020; accepted April 3, 2020. Date of publication April 20, 2020; date of current version May 1, 2020. This work was supported by the Science and Technology Plan of Shenzhen, China, under Grant JCYJ20170818114014753. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gonzalo Mateos. *(Corresponding author: An Liu.)*

Xuanyu Zheng and Vincent Lau are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: xzhengaj@connect.ust.hk; eeknlau@ece.ust.hk).

An Liu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: anliu@zju.edu.cn).

Digital Object Identifier 10.1109/TSP.2020.2986551

at the expense of higher complexity due to the increase in the dimensionality of the unknown vector to be recovered. Besides, a denser grid may cause the $l_{2,1}$ minimization based sparsity recovery methods to fail due to a higher correlation between the basis vectors.

In addition, all theses works have ignored practical issues in localization such as phase noise (PHN). The presence of PHN will have significant impact on localization performance especially when the target resolution is high because the PHN will induce random rotation on the AoA estimates [14]. There are a number of works that considered PHN mitigation in communication problems. For example, [15] studied the effect of PHN on high data rate wireless communication with large constellation size. The PHN rotates the phase of received signal on the constellation map and increase the detection error rate. [16] studied channel estimation with PHN in MIMO systems and gives Cramr-Rao lower bounds for channel gain and PHN estimation. A Weighted Least-Squares (WLS) data aided estimator is proposed to estimate and eliminate the phase noise so as to help reduce bit error rate. However, there is extensive pilot overhead in these existing PHN estimation and mitigation methods, and the overhead is more significant for the case of massive MIMO.

In this paper, we considered a dynamic-grid model for massive MIMO joint localization and channel estimation in the presence of PHN. Through some adjustable parameters, the dynamic-grid adjustment selectively enables finer resolution in the area of likely target locations and sparse resolution in other areas. Hence, it enables improvement of location accuracy without sacrificing the complexity in terms of the dimension of the unknown location vector. Using dynamic grid refinement, we model the massive MIMO channel and localization estimation as a unified sparse representation and recovery problem in the presence of PHN perturbation. By exploiting specific structured sparsity in the massive MIMO channels [13], we proposed a spherically-contoured radial exponential distribution (SRED) to enforce the hidden group sparsity in the channel so that the location and channel estimation as well as the grid adjustment can be modeled as a (maximum a posteriori) MAP estimation problem. The MAP problem is a challenging non-convex problem due to the non-linearity of the PHN and the hidden complex prior. By using approximation and majorization minimization techniques (MM), we propose effective surrogate for the MAP objective function and derive a robust and effective solution to estimate the accurate location of the mobile user and massive MIMO channel in the presence of PHN and limited pilot symbols. The proposed SRED regularized Sparse Bayes Inference (SRED-SBI) algorithm is shown to have low complexity and superior performance compared with the other state-of-the-art baselines.

The rest of the paper is organized as follows. In Section II, we present the system model as well as the sparse channel representation model for the localization problem. In Section III, we provide the MAP problem formulation for the joint localization and channel estimation with PHN, and then, in Section IV, we present the Sparse Bayesian Inference (SBI) based algorithm for solving the MAP problem. Numerical experiments and conclusions are provided in Section V and Section VI, respectively.

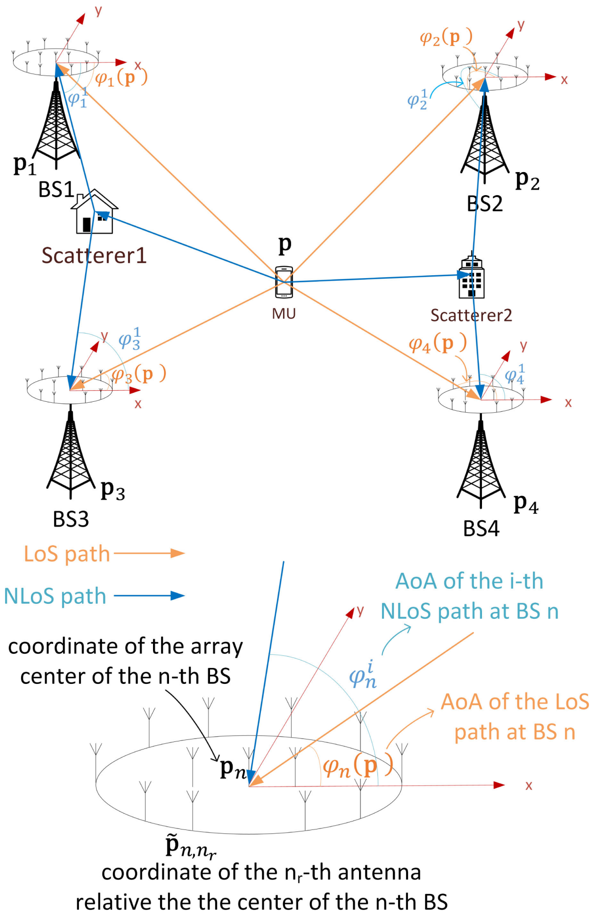


Fig. 1. Illustration of the uplink MIMO localization model with $N = 4$ BSs

Throughout the paper, we use the following notations. Upper case and lower case bold face letters denote matrices and vectors, respectively. $(\cdot)^{-1}$, $(\cdot)^T$ and $(\cdot)^H$ denote the inverse, transpose and conjugate transpose of a vector or matrix, respectively. $\mathbf{A} = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2)$ means $\mathbf{A}$ is block diagonal matrix with $\mathbf{A}_1$ and $\mathbf{A}_2$ on the diagonal. $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2]$ means $\mathbf{X} = [\mathbf{X}_1^T, \mathbf{X}_2^T]^T$. $\mathbf{v}(l)$ and $[\mathbf{v}]_l$ both denotes the l-th element of the vector $\mathbf{v}$. $\mathbf{v}_{a:b}$ means the sub-vector of $\mathbf{v}$ formed by the a-th element to the b-th element of $\mathbf{v}$. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes real and complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, respectively. $a = \mathcal{O}(1)$ means that a is at the order 1. $\mathbf{1}_N$ and $\mathbf{0}_N$ denotes the all one and all zero vector of length N, respectively.

II. System Model

A. System Setup and Signal Model

Consider a two-dimensional scenario where there are one mobile user (MU) with single transmitter and N distributed base stations (BS) equipped with N_r antennas each. The user is positioned at an unknown coordinate $\mathbf{p} = [p_x, p_y]^T$, in a target area $\mathcal{R}$. The coordinate of the array center of the n-th BS is $\mathbf{p}_n = [p_{n,x}, p_{n,y}]$ for $n = 1, \ldots, N$, as illustrated in Fig. 1. The locations of the N BSs are known and they act as anchors for MU position estimation.

Through channel sounding, the MU transmits a known pilot sequence $\mathbf{x} = [x_1, x_2 \ldots, x_T]^T \in \mathbb{C}^{T\times 1}$ with $|x_t| = 1$ for $t = 1, \ldots, T$ to the BSs. These BSs will cooperatively estimate $\mathbf{p}$ based on the received signals. Assume each BS has independent PHN, at time slot t, the received signal at the n-th BS $\mathbf{y}_{t,n} \in \mathbb{C}^{N_r}$ is given by

$$\mathbf{y}_{t,n} = \sqrt{P}\mathbf{\Lambda}_n \mathbf{h}_n x_t + \mathbf{z}_{t,n}, \quad (1)$$

where the diagonal matrix $\mathbf{\Lambda}_n \triangleq \text{diag}(e^{j\theta_{n,1}}, \ldots, e^{j\theta_{n,N_r}})$ captures the phase noise perturbation at the n-th BS. θ_{n,n_r} denotes the PHN on the n_r-th antenna at the n-th BS. $\mathbf{h}_n$ is the channel vector from the MU to the n-th BS. $P = \frac{P_T}{\sigma^2}$ is the transmit SNR, where P_T is the transmit power and σ^2 is the additive noise power, which are known to the receiver. $\mathbf{z}_{t,n} \in \mathbb{C}^{N_r}$ is the zero mean unit variance additive white Gaussian noise (AWGN) vector with $\mathbf{z}_{t,n} \sim \mathcal{CN}(0, \mathbf{I}_{N_r})$.

B. PHN Model

In this paper, we assume that the collective PHN vector on the n-th BS $\mathbf{\Theta}_n \triangleq [\theta_{n,1}, \ldots, \theta_{n,N_r}]^T$ is zero mean jointly Gaussian distributed with $\mathbf{\Theta}_n \sim \mathcal{N}(0, \mathbf{Q}_n)$ with covariance matrix $\mathbf{Q}_n$. We assumed that the PHN covariance matrix is known aprior. However, the PHN covariance matrix can also be easily measured from the phase noise spectrum of the oscillator [17], [18]. The phase noise covariance matrix can be estimated as

$$\mathbf{Q}_\theta = \sigma_\theta^2 \mathbf{I}_{Nr}$$

where $\mathbf{I}_{Nr}$ is an $N_r \times N_r$ identity matrix[1] and σ_θ^2 (in rad^2) is the PHN variance of the oscillator. PHN variance σ_θ^2 can be computed from [17], [18]

$$\sigma_\theta^2 = \int_0^B \frac{2L(f)}{C} df$$

where $L(f)$ is the single-side-band power spectral density (SSPSD) of PHN, B is the signal bandwidth and C is the carrier power.

This Gaussian PHN model is commonly used to model the residual phase error in systems with phase-tracking devices such as phase-locked loops (PLL), and can be found in literature explicitly in [15], [19]. We denote the PHN on all the BSs as $\mathbf{\Theta} \triangleq [\mathbf{\Theta}_1; \ldots; \mathbf{\Theta}_N]$. Since the PHN is a very slow process [20], we assume that the PHN remains constant during the transmission of T consecutive pilot symbols.

C. Wireless Channel Model

We consider flat fading channel with finite scatterers near the BSs. Assume that there are K_n non-LoS (NLoS) paths for the n-th BS, the channel vector from the MU to the n-th BS is given by [21]

$$\mathbf{h}_n = \xi_n \mathbf{a}_n(\varphi_n(\mathbf{p})) + \sum_{i=1}^{K_n} \xi_n^i \mathbf{a}_n(\varphi_n^i), \quad (2)$$

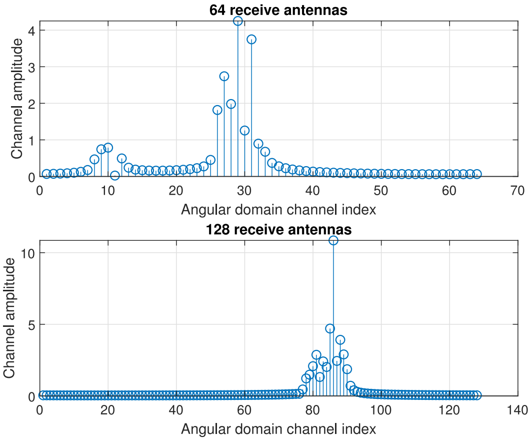


Fig. 2. Simulation of angular domain massive MIMO channel in urban macro scenario based on 3GPP SCM [22].

where ξ_n and ξ_n^i are the channel gain[2] for the LoS and i-th NLoS path, respectively, and are unknown complex scalars. $\varphi_n(\mathbf{p})$ and φ_n^i denote the AoAs of the LoS path and the i-th NLoS path between the MU to the n-th BS, as illustrated in Fig. 1. Note that the LoS AoA is related to the user and BS positions through [13]

$$\varphi_n(\mathbf{p}) = \arctan\left(\frac{p_{n,y} - p_y}{p_{n,x} - p_x}\right) + \pi \cdot \mathbb{1}(p_x < p_{n,x}), \quad (3)$$

where $\mathbb{1}(\boldsymbol{P})$ is one if the logic statement $\boldsymbol{P}$ is true. For an arbitrary array structure, let $\tilde{\mathbf{p}}_{n,n_r} = [\tilde{p}_{x,n,n_r}, \tilde{p}_{y,n,n_r}]^T$ be the coordinate of the n_r-th antenna of n-th BS, with respect to the center of the array. The steering vector $\mathbf{a}_n(\varphi)$ for a given AoA φ is given by the following formula

$$[\mathbf{a}_n(\varphi)]_{n_r} = \exp\left(\frac{2\pi j}{\lambda_c}\tilde{\mathbf{p}}_{n,n_r}^T \begin{bmatrix} \cos(\varphi) \\ \sin(\varphi) \end{bmatrix}\right) \quad (4)$$

where $[\cdot]_{n_r}$ denotes the n_r-th component of the vector, j is the imaginary unit and λ_c is the carrier wavelength. The steering vector implicitly explains the array structure and serves as an identifier of the AoA φ.

D. Sparse Representation for Localization With Dynamic Grid

For MIMO systems equipped with a large number of antennas, the angular resolution is very high, and hence the limited scatterers results in small angular spread in the angular domain [23]. This enables us to have a sparse representation for the channel in the angular domain. Fig. 2 shows two realizations of the massive MIMO channel energy in angular domain using Spatial Channel Model (SCM) [22] developed in 3GPP, which is widely used for evaluating the LTE system performance. We simulated the channel model in urban macro scenario where there are 6 random scatterers and each produces 20 sub-paths concentrated within a small angular spread. As seen from Fig. 2, the MIMO channel is indeed sparse in the angular domain. Such sparsity

[1] The PHN covariance matrix is diagonal because we assume that each antenna at the BS is connected to an independent oscillator, thus the PHNs on each antenna is independent.

[2] Note that we did not consider transmitter phase noise for we can always absorb it into the unknown channel gain ξ_n.

in the massive MIMO channel is also supported by field test measurements in [23], [24].

As seen from (1) and (2), the coordinate of the MU is related to the received measurements $\mathbf{y}_{t,n}$ in a complicated and nonlinear manner. Hence, it is difficult to recover the MU position and channel via traditional least-square (LS) methods with limited pilots. Alternatively, one can exploit sparse recovery method to obtain location information and channel estimates from limited measurements. The main idea is to sample the continuous map into grids and find a sparse representation to the MIMO channel. First, the target area $\mathcal{R}$ is uniformly sampled into L uniform grid locations for MU position, which is denoted by

$$\mathcal{L} = \{\mathbf{q}_1, \ldots, \mathbf{q}_L\} \in \mathcal{R},$$

where $\mathbf{q}_l = [q_{l,x}, q_{l,y}]^T$ is the coordinate of the l-th grid. We further introduce a uniform grid of M_n ($M_n \gg K_n$) discrete angles, denoted by $\mathcal{M}_n$ for AoAs of NLoS path i, i.e.,

$$\mathcal{M}_n = \{\phi_1, \ldots, \phi_{M_n}\} \in [0, 2\pi), \ \forall n = 1, \ldots, N.$$

If the MU precisely locates on the grid in $\mathcal{L}$ and the AoAs of NLoS paths comes exactly on the grid in $\mathcal{M}_n$, we can model $\mathbf{h}_n$ by

$$\mathbf{h}_n = \mathbf{A}_n \mathbf{w}_n + \mathbf{B}_n \mathbf{v}_n \tag{5}$$

where

$$\mathbf{A}_n = [\mathbf{a}_n(\varphi_n(\mathbf{q}_1)), \ldots, \mathbf{a}_n(\varphi_n(\mathbf{q}_L))] \in \mathbb{C}^{N_r \times L},$$

and

$$\mathbf{B}_n = [\mathbf{a}_n(\phi_1), \ldots, \mathbf{a}_n(\phi_{M_n})] \in \mathbb{C}^{N_r \times M_n}$$

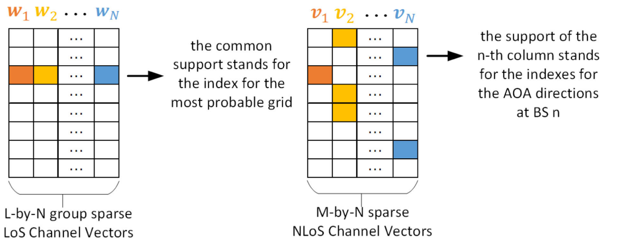

w_1 w_2 ⋯ w_N
the common support stands for the index for the most probable grid
L-by-N group sparse LoS Channel Vectors
v_1 v_2 ⋯ v_N
the support of the n-th column stands for the indexes for the AOA directions at BS n
M-by-N sparse NLoS Channel Vectors

Fig. 3. Illustration of the sparse channel representation, where we assumed the special case that $M_n = M$ for $n = 1, \ldots, N$.

$\varphi_n(\mathbf{q}_l)$ is the AoA from the l-th grid to the n-th BS and $\mathbf{a}_n(\varphi)$ is the steering vector of an arbitrary array structure, which are defined in (3) and (4), respectively. $\mathbf{w}_n = [w_{n,1}, \ldots, w_{n,L}] \in \mathbb{C}^{L \times 1}$ is called the sparse LoS channel coefficients with $w_{n,l}$ corresponding to the channel gain of the LoS path from grid location l to the n-th BS. There is only one nonzero element in $\mathbf{w}_n$ which corresponds to the LoS channel gain ξ_n since the MU can only occupy one grid, and the support of $\mathbf{w}_n$ correspond to the true position of the grid coordinates $\mathcal{L} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_L\}$. For example, if the MU position is at $\mathbf{q}_{l^*}$, then the corresponding l^*-th element of $\mathbf{w}_n$ is nonzero with $[\mathbf{w}_n]_{l^*} = \xi_n$ for all $n = 1, \ldots, N$. Similarly, $\mathbf{v}_n = [v_{n,1}, \ldots, v_{n,M_n}] \in \mathbb{C}^{M_n \times 1}$ is the sparse NLoS channel coefficients where $v_{n,m}$ denotes the channel gain from the MU to BS n with AoA ϕ_m. By definition, there are K_n nonzero elements in $\mathbf{v}_n$, which corresponds to the NLoS channel gains ξ_n^i, $i = 1, \ldots, K_n$ in (2). This is illustrated in Fig. 3. However, as discussed in Section I, the assumption that the true position and AoA of NLoS paths are always located on the grid points is not valid, and the mismatch between the coordinates of the true position\AoA and grid points is inevitable. One can reduce the location quantization errors by increasing the density of the position grids. However, such brute force approach substantially increase the dimension of the basis $\mathbf{A}_n$ and $\mathbf{B}_n$, which will increase the complexity of the algorithm. To address this problem, we introduce the dynamic-grid model. The basic idea is to impose a non-uniform grid based on the likelihood of the user location by adjusting the dynamic-grid parameters that changes the distribution of the grids. A finer grid will be used in the high probability region and a sparse grid will be used in other lower probability region. If the true MU position $\mathbf{p} \notin \mathcal{L}$ and let $\mathbf{q}_{l_r}, l_r \in \{1, \ldots, L\}$, denotes the nearest grid point to $\mathbf{p}$, the MU position can be expressed as

$$\mathbf{p} = \mathbf{q}_{l_r} + \boldsymbol{\beta}_{l_r} \tag{6}$$

where $\boldsymbol{\beta}_{l_r} = [\Delta x_{l_r}, \Delta y_{l_r}]^T$ with Δx_{l_r} and Δy_{l_r} denoting the off-grid gap in the x-axis and y-axis, respectively. Similarly, let ϕ_{m_i} be the nearest grid point to φ_n^i in $\mathcal{M}_n$, we can write

$$\varphi_n^i = \phi_{m_i} + \vartheta_{n,m_i},$$

where ϑ_{n,m_i} is the angular dynamic-grid parameter. In such we can represent the channel $\mathbf{h}_n$ by

$$\mathbf{h}_n = \mathbf{A}_n(\boldsymbol{\beta}) \mathbf{w}_n + \mathbf{B}_n(\boldsymbol{\vartheta}_n) \mathbf{v}_n \tag{7}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T \ldots, \boldsymbol{\beta}_L^T]^T \in \mathbb{C}^{2L \times 1}$, $\mathbf{A}_n(\boldsymbol{\beta}) = [\mathbf{a}_n(\varphi_n(\mathbf{q}_1 + \boldsymbol{\beta}_1)), \ldots, \mathbf{a}_n(\varphi_n(\mathbf{q}_L + \boldsymbol{\beta}_L))] \in \mathbb{C}^{N_r \times L}$, and

$$\boldsymbol{\beta}_l = \begin{cases} \mathbf{p} - \mathbf{q}_{l_r}, & l = l_r \\ [0, 0]^T, & \text{otherwise.} \end{cases}$$

$\boldsymbol{\vartheta}_n = [\vartheta_{n,1}, \ldots, \vartheta_{n,M_n}]^T \in \mathbb{R}^{M_n \times 1}$, $\mathbf{B}_n(\boldsymbol{\vartheta}_n) = [\mathbf{a}_n(\phi_1 + \vartheta_{n,1}), \ldots, \mathbf{a}_n(\phi_{M_n} + \vartheta_{n,M_n})] \in \mathbb{C}^{N_r \times M_n}$ and

$$\vartheta_{n,m_i} = \begin{cases} \varphi_n^i - \phi_{m_i}, & i = 1, \ldots, K_n \\ 0, & \text{otherwise.} \end{cases}$$

The parameterized basis vectors $\mathbf{A}_n(\boldsymbol{\beta})$ and $\mathbf{B}_n(\boldsymbol{\vartheta}_n)$ can fully address the location quantization error problem since there is always such $\boldsymbol{\beta}_{l_r}$ and ϑ_{n,m_i} such that (7) holds exactly. The estimated parameters are the sparse vectors $\mathbf{w}_n$ and $\mathbf{v}_n$, the support index l_s of the sparse vectors $\mathbf{w}_n$ that indicates the most probable grid position, as well as the dynamic grid parameters $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}_n$.

With the above dynamic-grid localization model, the received signal across all the BSs at time slot t is given by

$$\mathbf{y}_t = x_t \cdot \sqrt{P} \boldsymbol{\Lambda} (\boldsymbol{\Omega}(\boldsymbol{\beta}) \mathbf{w} + \boldsymbol{\Phi}(\boldsymbol{\vartheta}) \mathbf{v}) + \mathbf{z}_t \tag{8}$$

where $\mathbf{y}_t = [\mathbf{y}_{t,1}; \ldots; \mathbf{y}_{t,N}] \in \mathbb{C}^{NN_r \times 1}$,

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) = \text{diag}(\mathbf{A}_1(\boldsymbol{\beta}), \ldots, \mathbf{A}_N(\boldsymbol{\beta})) \in \mathbb{C}^{NN_r \times NL},$$

$$\boldsymbol{\Phi}(\boldsymbol{\vartheta}) = \text{diag}(\mathbf{B}_1(\boldsymbol{\vartheta}_1), \ldots, \mathbf{B}_N(\boldsymbol{\vartheta}_N)) \in \mathbb{C}^{NN_r \times M_S},$$

$\mathbf{w} = [\mathbf{w}_1; \mathbf{w}_2, \ldots; \mathbf{w}_N] \in \mathbb{C}^{NL \times 1}$, $\mathbf{v} = [\mathbf{v}_1; \mathbf{v}_2, \ldots; \mathbf{v}_N] \in \mathbb{C}^{M_S \times 1}$, $\mathbf{z}_t = [\mathbf{z}_{t,1}; \ldots; \mathbf{z}_{t,N}]$, $M_S = \sum_{n=1}^N M_n$ and

$\mathbf{\Lambda} = \mathrm{diag}(\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_N) \in \mathbb{C}^{NN_r \times NN_r}$. Stacking the received signal across time t, the signal model can be further rewritten as

$$\mathbf{y} = \sqrt{P}\bar{\mathbf{\Lambda}}\left(\bar{\mathbf{\Omega}}(\boldsymbol{\beta})\mathbf{w} + \bar{\mathbf{\Phi}}(\boldsymbol{\vartheta})\mathbf{v}\right) + \mathbf{z},$$
$$= \sqrt{P}\bar{\mathbf{\Lambda}}\bar{\mathbf{F}}(\boldsymbol{\beta}, \boldsymbol{\vartheta})\mathbf{u} + \mathbf{z}$$

where $\bar{\mathbf{\Lambda}} = \mathrm{diag}(\mathbf{\Lambda}, \ldots, \mathbf{\Lambda})$, $\mathbf{z} = [\mathbf{z}_1; \ldots; \mathbf{z}_T]$, $\bar{\mathbf{\Omega}}(\boldsymbol{\beta}) = [x_1\mathbf{\Omega}(\boldsymbol{\beta}); \ldots; x_t\mathbf{\Omega}(\boldsymbol{\beta})] \in \mathbb{C}^{TNN_r \times NL}$, $\bar{\mathbf{\Phi}}(\boldsymbol{\vartheta}) = [x_1\mathbf{\Phi}(\boldsymbol{\vartheta}); \ldots; x_T\mathbf{\Phi}(\boldsymbol{\vartheta})] \in \mathbb{C}^{TNN_r \times M_{\mathrm{S}}}$, $\bar{\mathbf{F}}(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = [\bar{\mathbf{\Omega}}(\boldsymbol{\beta}), \bar{\mathbf{\Phi}}(\boldsymbol{\vartheta})]$ and $\mathbf{u} = [\mathbf{w}; \mathbf{v}]$. The LoS channel coefficient $\mathbf{w}$ is group sparse since the support of $\mathbf{w}_n$ are the same. Due to the dynamic grid adjustment $\boldsymbol{\beta}, \boldsymbol{\vartheta}$ and the unknown PHN $\mathbf{\Lambda}$, the basis vectors $\bar{\mathbf{\Lambda}}\mathbf{F}(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ are unknown at the BS. Traditional $l_{2,1}$-norm based sparse recovery algorithm [13] cannot be directly applied to recover the MU position and channel from $\mathbf{y}$. Consequently, we proposed a MAP based formulation to jointly recover the location and the channel matrix under limited pilot resource and PHN perturbation.

III. MAP Formulation of Joint Localization and Channel Estimation With PHN

A. Prior Model for PHNs $\mathbf{\Theta}_n$, Dynamic Grid $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$

The prior model for PHN on the n-th BS is jointly Gaussian distributed with zero mean and covariance matrix $\mathbf{Q}_n$. Since we also assume the PHN on different BSs are independent, the probability density function of $\mathbf{\Theta}$ can be written as

$$p(\mathbf{\Theta}) = \frac{1}{\sqrt{\det(2\pi\mathbf{Q}_\theta)}}\exp\left(-\mathbf{\Theta}^T\frac{\mathbf{Q}_\theta^{-1}}{2}\mathbf{\Theta}^T\right), \quad \mathbf{\Theta} \in \mathbb{R}^{NN_r \times 1} \tag{9}$$

where $\mathbf{Q}_\theta = \mathrm{diag}(\mathbf{Q}_1, \ldots, \mathbf{Q}_N) \in \mathbb{R}^{NN_r \times NN_r}$. For the dynamic grid parameters, we assume non-informative uniform priors on $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$.

B. Prior Model for Massive MIMO Channels

The channel coefficient $\mathbf{u}$ consists of the LoS and NLoS channel coefficients $\mathbf{w}$ and $\mathbf{v}$, i.e., $\mathbf{u} = [\mathbf{w}; \mathbf{v}]$. Assuming $\mathbf{w}$ and $\mathbf{v}_n$ are independent, the prior distribution of $\mathbf{u}$ can be decomposed as

$$p(\mathbf{u}) = p(\mathbf{w})\,p(\mathbf{v}) = p(\mathbf{w})\prod_{n=1}^{N} p(\mathbf{v}_n).$$

By definition, $\mathbf{w}$ is group sparse in the sense that there is only one $l \in \{1, 2, \ldots, L\}$ such that the vector $\bar{\mathbf{w}}_l \triangleq [w_{1,l}, w_{2,l}, \ldots, w_{N,l}]^T \in \mathbb{C}^{N \times 1}$ is elementwisely nonzero. Inspired by the $l_{2,1}$-norm regularized minimization problem, which is known for its capability of enforcing group sparsity, we proposed a d-dimensional spherically-contoured radial exponential distribution (SRED) [25] distribution to enforce group sparsity. For a d-component random vector $\mathbf{s}$, the d-dimensional SRED scale mixture is defined as

$$p(\mathbf{s}|\lambda) = C_d\lambda^d e^{-\lambda\|\mathbf{s}\|_2}, \quad \mathbf{s} \in \mathbb{R}^{d \times 1} \tag{10}$$

where $\lambda > 0$ is called the inverse scale parameter, $C_d = \frac{g(1+\frac{d}{2})}{d \cdot g(d)\pi^{\frac{d}{2}}}$ is a normalization constant that is computed via d-dimensional hyper-sphere volume. $g(\rho)$ is the gamma function and is defined as $g(\rho) = \int_0^\infty \mu^{\rho-1}e^{-\mu}d\mu$. The inverse scale parameter is related to the variance of the elements in $\mathbf{s}$ given by $\mathbb{E}[s_i^2] = \frac{d+1}{\lambda^2}$. We shall denote the distribution in (10) as

$$p(\mathbf{s}|\lambda) \sim \mathrm{SRED}^d(\mathbf{s}; \lambda), \quad \mathbf{s} \in \mathbb{R}^{d \times 1}$$

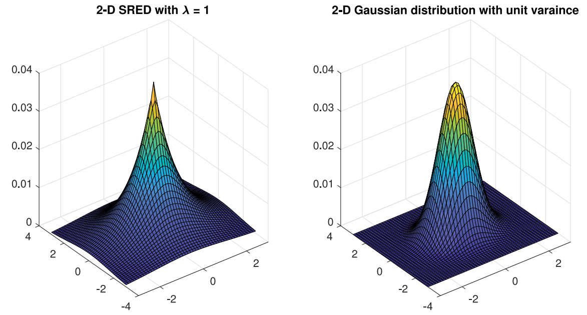


Fig. 4. Comparison between 2-D spherically-contoured radial exponential distribution (SRED) and Gaussian distribution.

The radial profile of the density in (10) is exponential in $\|\mathbf{s}\|$ and when $d = 1$, (10) became the Laplacian Scale Mixture (LSM) [26] distribution for scalar case, which is known for its superior capability of enforcing sparsity over Gaussian Scale Mixture (GSM). Thus (10) can be viewed as Laplacian distribution for vector case. Fig. 4 shows the probability density function of a 2-D SRED and 2-D Gaussian random variable with unit variance. Note that SRED has a sharp point at zero. We then impose the SRED on LoS channels $\bar{\mathbf{w}}_l$ for $l = 1, \ldots, L$. Since the SRED is defined on real random variables, we decompose the channel into real and imaginary parts and denote $\tilde{\mathbf{w}} = [\mathrm{Re}(\mathbf{w}); \mathrm{Im}(\mathbf{w})] \in \mathbb{R}^{2NL \times 1}$ and $\tilde{\mathbf{w}}_l = [\mathrm{Re}(\bar{\mathbf{w}}_l); \mathrm{Im}(\bar{\mathbf{w}}_l)] \in \mathbb{R}^{2N \times 1}$. We assume the LoS channels for each grid l are independent and have the following prior:

$$p(\tilde{\mathbf{w}}_l|\lambda_l) \sim \mathrm{SRED}^{2N}(\tilde{\mathbf{w}}_l; \lambda_l) = \lambda_l^{2N}e^{-\lambda_l\|\tilde{\mathbf{w}}_l\|_2}$$

i.e., $p(\tilde{\mathbf{w}}_l|\lambda_l) = C_{2N}\lambda_l^{2N}e^{-\lambda_l\|\tilde{\mathbf{w}}_l\|_2}$ and λ_l is the corresponding inverse scale parameter and it determines the variance of each element as $\mathbb{E}[\tilde{w}_{n,l}] = \frac{2N+1}{\lambda_l^2}$. Letting $\lambda = [\lambda_1, \ldots, \lambda_L]$, the density of LoS channel $\mathbf{w}$ conditioned on λ can be written as

$$p(\mathbf{w}|\lambda) = p(\tilde{\mathbf{w}}|\lambda) = \prod_{l=1}^{L} p(\tilde{\mathbf{w}}_l|\lambda_l) \tag{11}$$

Following the commonly used method in [26], we further model λ_l as i.i.d. Gamma distributions, i.e.,

$$p(\lambda) = \prod_{l=1}^{L}\Gamma(\lambda_l; a, b), \tag{12}$$

where $\Gamma(\rho; a, b)$ is a Gamma hyper distribution with shape parameter a and rate parameter b. The Gamma distribution is selected here because it is *conjugate* to SRED distribution and therefore the associated Bayesian inference have closed-form expressions [26]. Specifically, the log prior for LoS channel can be computed by

$$p\left(\tilde{\mathbf{w}}_l\right) = \int_{\lambda_l} p\left(\tilde{\mathbf{w}}_l|\lambda_l\right) p\left(\lambda_l\right) d\lambda_l = \frac{C_{2N} a^{2N} b^a}{\left(b + \|\tilde{\mathbf{w}}_l\|_2\right)^{2N+a}} \quad (13)$$

Since the variance of the channel gains are usually unknown, we set $b \to 0$ as in [6] to obtain a broad prior. Such choice of parameter also results in a prior distribution of $\tilde{\mathbf{w}}_l$ that encourages sparsity due to sharp peak at zero [27].

The priors in (11) and (12) can capture the group sparsity property of the LoS channels as explained below. When there is an active LoS path from the grid l to the BSs, the inverse scale parameter λ_l should be $O(1)$, because the variance $\frac{2N+1}{\lambda_l^2}$ of each element $\tilde{w}_{n,l}$ of the vector $\tilde{\mathbf{w}}_l$ is $O(1)$ when it is active. When there is no active LoS path from thr l-th grid to the BSs, λ_l should be large so that the variance $\frac{2N+1}{\lambda_l^2}$ of each element $\tilde{w}_{n,l}$ of the vector $\tilde{\mathbf{w}}_l$ is close to zero when it is inactive. We use different λ_l for different grids to capture the group sparsity in the LoS channel $\mathbf{w}$. We will show how we automatically update λ_l to enforce group sparsity in the next section.

The NLoS channels $\mathbf{v}_n$ is sparse for all n, and we can impose a 2-D SRED on the real and imaginary parts of each element of $\mathbf{v}_n$. Letting $\tilde{\mathbf{v}} = [\mathrm{Re}(\mathbf{v}); \mathrm{Im}(\mathbf{v})]$, $\tilde{\mathbf{v}}_n = [\mathrm{Re}(\mathbf{v}_n); \mathrm{Im}(\mathbf{v}_n)]$ and $\tilde{\mathbf{v}}_{n,m} = [\mathrm{Re}(v_{n,m}), \mathrm{Im}(v_{n,m})]^T$, we have the prior

$$p\left(\tilde{\mathbf{v}}_{n,m}|\gamma_{n,m}\right) \sim \mathrm{SRED}^2\left(\tilde{\mathbf{v}}_{n,m}, \gamma_{n,m}\right) = \gamma_{n,m}^2 e^{-\gamma_{n,m}\|\tilde{\mathbf{v}}_{n,m}\|_2},$$

where $\gamma_{n,m}$ are the corresponding positive inverse scale parameters. Let $\boldsymbol{\gamma}_n = [\gamma_{n,1}, \ldots, \gamma_{n,M_n}]^T$ and $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1; \ldots; \boldsymbol{\gamma}_N]$, the density of NLoS channel $\mathbf{v}$ conditioned on $\boldsymbol{\gamma}$ is

$$p\left(\mathbf{v}|\boldsymbol{\gamma}\right) = \prod_{n=1}^{N} \prod_{m=1}^{M_n} p\left(\tilde{\mathbf{v}}_{n,m}|\gamma_{n,m}\right). \quad (14)$$

Similarly, we model $\gamma_{n,m}$ as i.i.d. Gamma distribution:

$$p\left(\boldsymbol{\gamma}\right) = \prod_{n=1}^{N} \prod_{m=1}^{M_n} \Gamma\left(\gamma_{n,m}; \bar{a}, \bar{b}\right) \quad (15)$$

and the resulting prior for NLoS channel can be computed as

$$p\left(\tilde{\mathbf{v}}_{n,m}\right) = \frac{C_2 \bar{a}^2 \bar{b}^a}{\left(\bar{b} + \|\tilde{\mathbf{v}}_{n,m}\|_2\right)^{2+\bar{a}}}. \quad (16)$$

C. Formulation for SRED-SPI Algorithm via MAP

Given the received signal $\mathbf{y}$, we aim to infer its sparse representation $\mathbf{u}$ in the basis vectors and the dynamic-grid parameters $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$. In this section, we consider the formulation of MAP of the estimated parameters given $\mathbf{y}$. Since we only focus on the accuracy of localization and channel estimation and it is difficult to estimate all the PHN parameters in each antenna with limited pilot symbols, we just integrate over the PHN parameters and consider the average effect of PHN. With Bayes' rule, the log posterior probability can be written as

$$\begin{aligned} &\log p\left(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\vartheta}\right) \\ &\propto \log p\left(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\vartheta}\right) + \log p\left(\mathbf{u}\right) \\ &= \log \int_{\boldsymbol{\Theta}} p\left(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\vartheta}, \boldsymbol{\Theta}\right) p\left(\boldsymbol{\Theta}\right) d\boldsymbol{\Theta} + \log p\left(\mathbf{u}\right) \end{aligned} \quad (17)$$

Thus the estimated parameters are obtained by solving the following MAP Problem

$$\begin{aligned} &\left[\hat{\mathbf{u}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\vartheta}}\right] \\ &= \arg\max_{\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\vartheta}} \log \int_{\boldsymbol{\Theta}} p\left(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\vartheta}, \boldsymbol{\Theta}\right) p\left(\boldsymbol{\Theta}\right) d\boldsymbol{\Theta} + \log p\left(\mathbf{u}\right). \end{aligned} \quad (18)$$

Then, for given estimate of $\mathbf{w} = \hat{\mathbf{w}} = \hat{\mathbf{u}}_{1:NL}$, we obtain the estimate of the index for the most probable grid l_s by

$$\hat{l}_s = \arg\max_{l \in \{1, \ldots, L\}} \left\|\hat{\tilde{\mathbf{w}}}_l\right\|^2. \quad (19)$$

Finally, the coordinate of the MU position can then be recovered by

$$\hat{\mathbf{p}} = \mathbf{q}_{\hat{l}_s} + \hat{\boldsymbol{\beta}}_{\hat{l}_s}, \quad (20)$$

while the estimated uplink channel for the n-th BS is recovered by

$$\hat{\mathbf{h}}_n = \left[\hat{\mathbf{w}}_n\right]_{\hat{l}_s} \mathbf{a}_n\left(\varphi_n\left(\hat{\mathbf{p}}\right)\right) + \boldsymbol{\Phi}\left(\hat{\boldsymbol{\vartheta}}\right)\hat{\mathbf{v}}, \quad (21)$$

where $\hat{\mathbf{v}} = \hat{\mathbf{u}}_{NL+1:NL+M_S}$ is the estimate of $\mathbf{v}$.

However, it is very challenging to calculate the exact likelihood term in (18) because it is difficult to compute the closed-form result of the integration over a nonlinear function of PHN. Besides, it is difficult to jointly optimize the parameters $(\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\vartheta})$ since it is high-dimensional and the objective is non-convex. To overcome these challenges, we use tractable approximation to compute the likelihood term and used MM to obtain effective surrogate for the objective function. Based on these we proposed the SRED-SPI algorithm to find the approximate stationary point of problem (18).

IV. Joint Channel Estimation and Localization Algorithm

In the optimization problem (18), we need to optimize the channel coefficient parameters $\mathbf{u}$ and the dynamic-grid parameters $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ based on the received signals $\mathbf{y}$. As mentioned in Section III-C, it has many challenges. One possible alternative is to jointly optimize all the parameters at the same time but such brute force solution is computational unacceptable for high localization resolution, and it is intractable as it is difficult to numerically compute the likelihood term. To address these challenges, we propose the SRED-SPI algorithm which is mainly based on problem approximation and In-exact Block MM.

1) **Problem Approximation**: We obtain an approximated MAP problem by replacing the exact likelihood term with a tractable approximation which is accurate for small to moderate PHN. This approximation is easy to compute and is obtained by transforming the nonlinear PHN to equivalent additive perturbation on the received signal.
2) **In-exact Block MM**: We propose to use in-exact block MM algorithm to find a stationary point $(\mathbf{u}^*, \boldsymbol{\beta}^*, \boldsymbol{\vartheta}^*)$ of the approximated MAP problem by alternatively optimizing

each parameter. Because there is no closed-form solution in the subproblem of updating the channel coefficients $\mathbf{u}$, we propose to use MM algorithm to iteratively update the channel coefficients.

A. Approximation to the Likelihood Term

In this section, we propose to approximate the likelihood function in (18) under the assumption that PHN standard deviation on each antenna is not large (smaller than 10°) [15]. With zero mean unit variance AWGN, the log likelihood can be written as

$$\log \mathbb{E}_{\boldsymbol{\Theta}}\left[\exp\left(-\left\|\mathbf{y}-\sqrt{P}\bar{\boldsymbol{\Lambda}}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}\right\|_2^2\right)\right] \tag{22}$$

Let $\bar{\boldsymbol{\Theta}} = [\boldsymbol{\Theta};\ldots;\boldsymbol{\Theta}] \in \mathbb{R}^{TNN_r\times 1}$, the Euclidean norm can be approximated as

$$\begin{aligned}&\left\|\mathbf{y}-\sqrt{P}\bar{\boldsymbol{\Lambda}}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}\right\|_2^2\\ &\quad=\left\|\mathrm{diag}\left(\mathbf{y}\right)\exp\left(-j\bar{\boldsymbol{\Theta}}\right)-\sqrt{P}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}\right\|_2^2\\ &\quad\approx\left\|\mathbf{y}-\sqrt{P}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}-j\mathrm{diag}\left(\mathbf{y}\right)\bar{\boldsymbol{\Theta}}\right\|_2^2\\ &\quad=\left\|\mathbf{y}-\sqrt{P}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}-j\mathbf{D}_y\boldsymbol{\Theta}\right\|_2^2,\end{aligned}$$

where we defined $\mathbf{D}_y = [\mathrm{diag}(\mathbf{y}_1);\ldots;\mathrm{diag}(\mathbf{y}_T)]$ and the approximation is a result of linear expansion $e^{j\theta}\approx 1+j\theta$ for small θ. Decomposing the Euclidean norm to real and imaginary parts, we have the approximation

$$\log\mathbb{E}_{\boldsymbol{\Theta}}\left\{\left\|\mathbf{y}-\sqrt{P}\bar{\boldsymbol{\Lambda}}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}\right\|_2^2\right\}\approx\log\mathbb{E}_{\boldsymbol{\Theta}}\left\{\left\|\tilde{\mathbf{D}}_y\boldsymbol{\Theta}+\mathbf{b}\right\|_2^2\right\} \tag{23}$$

where we define

$$\tilde{\mathbf{D}}_y \triangleq \left[\mathrm{Im}\left(\mathbf{D}_y\right);-\mathrm{Re}\left(\mathbf{D}_y\right)\right]\in\mathbb{R}^{2TNN_r\times NN_r} \tag{24}$$

and

$$\mathbf{b}\triangleq\begin{bmatrix}\mathrm{Re}\left(\mathbf{y}-\sqrt{P}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}\right)\\ \mathrm{Im}\left(\mathbf{y}-\sqrt{P}\bar{\mathbf{F}}\left(\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\mathbf{u}\right)\end{bmatrix}\in\mathbb{R}^{2TNN_r\times 1} \tag{25}$$

Proposition 1: Let $\tilde{\mathbf{D}}_y$ and $\mathbf{b}$ be defined as (24) and (25), respectively. Then the approximated log likelihood in (23) can be calculated in closed-form as

$$\log\mathbb{E}_{\boldsymbol{\Theta}}\left\{\left\|\tilde{\mathbf{D}}_y\boldsymbol{\Theta}+\mathbf{b}\right\|_2^2\right\}=-\mathbf{b}^T\mathbf{W}_y^{-1}\mathbf{b}-\frac{1}{2}\log\det\left(\mathbf{W}_y\right) \tag{26}$$

where

$$\mathbf{W}_y=\mathbf{I}+\tilde{\mathbf{D}}_y\left(2\mathbf{Q}_\theta\right)\tilde{\mathbf{D}}_y^T\in\mathbb{C}^{2TNN_r\times 2TNN_r} \tag{27}$$

Proof: See Appendix VI-A. ■

When $\mathbf{Q}_\theta = 0$, i.e., when there is no PHN, the approximated likelihood term (26) becomes minimizing $\|\mathbf{b}\|^2 = \|\mathbf{y}-\sqrt{P}\bar{\mathbf{F}}(\boldsymbol{\beta},\boldsymbol{\vartheta})\mathbf{u}\|^2$, which correspond to the naive likelihood that ignores the PHN. When $\mathbf{Q}_\theta \neq 0$, the approximated log likelihood function depends on the received signal $\mathbf{y}$ via the matrix $\mathbf{W}_y$. We can think of $\mathbf{W}_y$ as the covariance matrix of the equivalent additive noise and it depends on the received signal $\mathbf{y}$. Besides, for fixed $\mathbf{Q}_\theta$, the impact of PHN is larger when $\|\mathbf{y}\|$ is larger, i.e., when transmit power is higher.

With the approximation to the log-likelihood in (26), we can get the approximated MAP problem (18) as

$$\hat{\mathbf{u}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\vartheta}}=\arg\max_{\mathbf{u},\boldsymbol{\beta},\boldsymbol{\vartheta}}-\mathbf{b}^T\mathbf{W}_y^{-1}\mathbf{b}+\log p\left(\mathbf{u}\right). \tag{28}$$

Let $\mathbf{W}_y^{\frac{1}{2}}\in\mathbb{C}^{2TNN_r\times 2TNN_r}$ be any matrix such that $(\mathbf{W}_y^{-\frac{1}{2}})^H(\mathbf{W}_y^{-\frac{1}{2}})=\mathbf{W}_y^{-1}$. Define $\tilde{\mathbf{y}}=[\mathrm{Re}(\mathbf{y});\mathrm{Im}(\mathbf{y})]\in\mathbb{R}^{2TNN_r\times 1}$, $\tilde{\mathbf{u}}=[\tilde{\mathbf{w}};\tilde{\mathbf{v}}]$,

$$\tilde{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}\right)=\sqrt{P}\begin{bmatrix}\mathrm{Re}\left(\bar{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}\right)\right)&-\mathrm{Im}\left(\bar{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}\right)\right)\\ \mathrm{Im}\left(\bar{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}\right)\right)&\mathrm{Re}\left(\bar{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}\right)\right)\end{bmatrix}\in\mathbb{R}^{2TNN_r\times 2NL}$$

and

$$\tilde{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}\right)=\sqrt{P}\begin{bmatrix}\mathrm{Re}\left(\bar{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}\right)\right)&-\mathrm{Im}\left(\bar{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}\right)\right)\\ \mathrm{Im}\left(\bar{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}\right)\right)&\mathrm{Re}\left(\bar{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}\right)\right)\end{bmatrix}\in\mathbb{R}^{2TNN_r\times 2M_{\mathrm{S}}},$$

the optimization problem in (28) is equivalent to

$$\begin{aligned}\left[\hat{\mathbf{u}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\vartheta}}\right]&=\arg\max_{\mathbf{u},\boldsymbol{\beta},\boldsymbol{\vartheta}}-\left\|\left(\mathbf{W}_y^{-\frac{1}{2}}\right)\mathbf{b}\right\|_2^2+\log p\left(\tilde{\mathbf{u}}\right)\\ &=\arg\max_{\mathbf{u},\boldsymbol{\beta},\boldsymbol{\vartheta}}\Big\{-\left\|\mathbf{W}_y^{-\frac{1}{2}}\left[\tilde{\mathbf{y}}-\tilde{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}\right)\tilde{\mathbf{w}}-\tilde{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}\right)\tilde{\mathbf{v}}\right]\right\|_2^2\\ &\quad+\log p\left(\tilde{\mathbf{w}}\right)+\log p\left(\tilde{\mathbf{v}}\right)\Big\}=\arg\max_{\mathbf{u},\boldsymbol{\beta},\boldsymbol{\vartheta}}f\left(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\end{aligned} \tag{29}$$

where we defined the approximated objective function as $f(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta})$ and the approximated real counterpart of the log-likelihood term is defined as

$$\tilde{h}\left(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\triangleq-\left\|\mathbf{W}_y^{-\frac{1}{2}}\left[\tilde{\mathbf{y}}-\tilde{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}\right)\tilde{\mathbf{w}}-\tilde{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}\right)\tilde{\mathbf{v}}\right]\right\|_2^2$$

B. Block MM for the Approximated MAP

As seen in (29), we have obtained an approximated MAP problem by approximating the log-likelihood term. We can simply substitute the expression of prior terms $p(\tilde{\mathbf{w}}_l)$ and $p(\tilde{\mathbf{v}}_{n,m})$ in (13) and (16) into (29). However, this leads to a non-convex optimization problem in terms of $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{v}}$ since it is maximizing the summation of a concave function (the log likelihood term) and a convex function (the log prior term). In this section, we propose to use block MM algorithm to find a stationary point of problem (29).

The basic idea of MM algorithm is to iteratively construct a continuous surrogate function for the the objective function $f(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta})$ and then maximize the objective function with respect to $(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta})$ iteratively. Specifically, let $u(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta},\boldsymbol{\vartheta})$ be the surrogate function to $f(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta})$ at some fixed point $(\tilde{\mathbf{u}}^{(t)},\boldsymbol{\beta},\boldsymbol{\vartheta})$ in the t-th iteration, and $\forall\boldsymbol{\beta},\boldsymbol{\vartheta}$ the surrogate function satisfy the following properties:

$$u\left(\tilde{\mathbf{u}}^{(t)}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta},\boldsymbol{\vartheta}\right)=f\left(\tilde{\mathbf{u}}^{(t)},\boldsymbol{\beta},\boldsymbol{\vartheta}\right), \tag{30}$$

$$u\left(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta},\boldsymbol{\vartheta}\right)\le f\left(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta}\right), \tag{31}$$

$$\frac{\partial u\left(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta},\boldsymbol{\vartheta}\right)}{\partial\tilde{\mathbf{u}}}|_{\tilde{\mathbf{u}}=\tilde{\mathbf{u}}^{(t)}}=\frac{f\left(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta}\right)}{\partial\tilde{\mathbf{u}}}|_{\tilde{\mathbf{u}}=\tilde{\mathbf{u}}^{(t)}}, \tag{32}$$

Then, we use block coordinate descent method to update the estimated parameters alternatively by

$$\tilde{\mathbf{u}}^{(t+1)} = \arg\max_{\tilde{\mathbf{u}}} u\left(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)}\right) \tag{33}$$

$$\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} u\left(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta},\boldsymbol{\vartheta}^{(t)}\right) \tag{34}$$

$$\boldsymbol{\vartheta}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} u\left(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta}^{(t+1)},\boldsymbol{\vartheta}\right) \tag{35}$$

In order to update the channel coefficient $\tilde{\mathbf{u}}$, we need to find an appropriate surrogate function $u(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta},\boldsymbol{\vartheta})$ that satisfies the three properties as listed in (30)–(32). Since the prior for LoS and NLOS channels $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{v}}$ are independent, we propose to construct two different surrogate functions to $p(\tilde{\mathbf{w}})$ and $p(\tilde{\mathbf{v}})$ respectively, i.e.,

$$u\left(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta},\boldsymbol{\vartheta}\right) = \tilde{h}\left(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta}\right) + u_w\left(\tilde{\mathbf{w}}|\tilde{\mathbf{w}}^{(t)}\right) + u_v\left(\tilde{\mathbf{v}}|\tilde{\mathbf{v}}^{(t)}\right)$$

Inspired by the expectation-maximization (EM) algorithm [28], we construct the surrogate function to $\log p(\tilde{\mathbf{w}})$ as the following lower bound function

$$u_w\left(\tilde{\mathbf{w}}|\tilde{\mathbf{w}}^{(t)}\right) = \int_{\lambda} p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log\left(\frac{p\left(\tilde{\mathbf{w}},\lambda\right)}{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}\right)d\lambda - \tau_w\left\|\tilde{\mathbf{w}}-\tilde{\mathbf{w}}^{(t)}\right\|^2 \tag{36}$$

Similarly, we define the surrogate function to $\log p(\tilde{\mathbf{v}})$ as

$$u_v\left(\tilde{\mathbf{v}}|\tilde{\mathbf{v}}^{(t)}\right) = \int_{\boldsymbol{\gamma}} p\left(\boldsymbol{\gamma}|\tilde{\mathbf{v}}^{(t)}\right)\log\left(\frac{p\left(\tilde{\mathbf{v}},\boldsymbol{\gamma}\right)}{p\left(\boldsymbol{\gamma}|\tilde{\mathbf{v}}^{(t)}\right)}\right)d\boldsymbol{\gamma} - \tau_v\left\|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}^{(t)}\right\|^2, \tag{37}$$

where $\tau_w > 0$ and $\tau_v > 0$ can be any constant. The second term is added to ensure that the surrogate functions (36) and (37) are not only convex but also strongly convex with respect to $\tilde{\mathbf{u}} = [\tilde{\mathbf{w}};\tilde{\mathbf{v}}]$. It also penalizes the objective function such that the solution will not deviate from $\tilde{\mathbf{u}}^{(t)}$ too far.

Lemma 2: With the surrogate functions defined in (36) and (37), $u(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta},\boldsymbol{\vartheta})$ is a surrogate function to $f(\tilde{\mathbf{u}}^{(t)},\boldsymbol{\beta},\boldsymbol{\vartheta})$ with properties (30)–(32) hold true.

Proof: See Appendix VI-B. ■

With the surrogate function defined in (36) and (37), we then discuss the parameter updates for $\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta}$.

1) Update for $\tilde{\mathbf{u}}$*:* The update for channel coefficients $\tilde{\mathbf{u}}$ is obtained by solving a convex optimization problem:

Lemma 3: The update for the channel coefficients $\tilde{\mathbf{u}}$ is given by the solution of the following optimization problem

$$\tilde{\mathbf{u}}^{(t+1)} = \arg\min_{\tilde{\mathbf{u}}=[\tilde{\mathbf{w}};\tilde{\mathbf{v}}]}\left\{\left\|\mathbf{W}_y^{-\frac{1}{2}}\left[\tilde{\mathbf{y}}-\tilde{\boldsymbol{\Omega}}\left(\boldsymbol{\beta}^{(t)}\right)\tilde{\mathbf{w}}-\tilde{\boldsymbol{\Phi}}\left(\boldsymbol{\vartheta}^{(t)}\right)\tilde{\mathbf{v}}\right]\right\|_2^2 + \sum_{l=1}^{L}\lambda_l^{(t+1)}\left\|\tilde{\mathbf{w}}_l\right\|_2 + \sum_{n=1}^{N}\sum_{m=1}^{M_n}\gamma_{n,m}^{(t+1)}\left\|\tilde{\mathbf{v}}_{n,m}\right\|_2 + \tau_w\left\|\tilde{\mathbf{w}}-\tilde{\mathbf{w}}^{(t)}\right\|^2 + \tau_v\left\|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}^{(t)}\right\|^2\right\} \tag{38}$$

where $\lambda_l^{(t+1)} = \langle\lambda_l\rangle_{p(\lambda|\tilde{\mathbf{w}}^{(t)})}$, $\gamma_{n,m}^{(t+1)} = \langle\gamma_{n,m}\rangle_{p(\boldsymbol{\gamma}|\tilde{\mathbf{v}}^{(t)})}$ and $\langle\rho_i\rangle_{q(\boldsymbol{\rho})} = \int_{\boldsymbol{\rho}} q(\boldsymbol{\rho})\rho_i d\boldsymbol{\rho}$ denotes the expectation of ρ_i with respect to distribution $q(\boldsymbol{\rho})$. Specifically,

$$\lambda_l^{(t+1)} = \langle\lambda_l\rangle_{p\left(\lambda|\tilde{\mathbf{w}}_l^{(t)}\right)} = \frac{a+2N}{b+\left\|\tilde{\mathbf{w}}_l^{(t)}\right\|_2} \tag{39}$$

and

$$\gamma_{n,m}^{(t+1)} = \langle\gamma_{n,m}\rangle_{p\left(\boldsymbol{\gamma}|\tilde{\mathbf{v}}_{n,m}^{(t)}\right)} = \frac{\bar{a}+2}{\bar{b}+\left\|\tilde{\mathbf{v}}_{n,m}^{(t)}\right\|_2}. \tag{40}$$

Proof: See Appendix VI-C. ■

The optimization problem in (38) is a least-square problem regularized by the weighted sum of the l_2-norm of the channel coefficient vectors. This is a convex quadratic problem and very efficient solver exists.

In each iteration for updating $\tilde{\mathbf{u}}$, we need to compute the weights $\lambda_l^{(t)}$ and $\gamma_{n,m}^{(t)}$. Note that the posteriors $p(\lambda|\tilde{\mathbf{w}})$ and $p(\boldsymbol{\gamma}|\tilde{\mathbf{v}})$ have closed-form expression because we imposed Gamma priors on λ and $\boldsymbol{\gamma}$ as defined in (12) and (15). It is tractable because the SRED and Gamma distribution are *conjugate*, i.e., the posterior probability of λ_l (or $\gamma_{n,m}$) given $\tilde{\mathbf{w}}_l$ (or $\tilde{\mathbf{v}}_{n,m}$) is also Gamma distributed, when the prior on λ_l (or $\gamma_{n,m}$) is Gamma distributed and the conditional probability of $\tilde{\mathbf{w}}_l$ (or $\tilde{\mathbf{v}}_{n,m}$) conditioned on λ_l (or $\gamma_{n,m}$) is SRED distributed. This enables us to compute the prior distribution of the inverse scale parameters conditioned on the channel coefficients.

Remark 4: After the t-th iteration, a small channel coefficient $\tilde{\mathbf{w}}_l^{(t)}$ (or $\tilde{\mathbf{v}}_{n,m}^{(t)}$) that is close to zero but not exactly zero will result in a larger reweighting factor $\lambda_l^{(t+1)}$ (or $\gamma_{n,m}^{(t+1)}$) in the next iteration, which enforces the channel coefficient $\tilde{\mathbf{w}}_l$ (or $\tilde{\mathbf{v}}_{n,m}$) to take smaller values and thus increases the chances of obtaining a (group) sparser solution. On the other hand, when $\tilde{\mathbf{w}}_l^{(t)}$ (or $\tilde{\mathbf{v}}_{n,m}^{(t)}$) are nonzero and takes relatively larger values, it means they are more likely to be the correct location (or NLoS AoA). Therefore, the reweighting factor $\lambda_l^{(t+1)}$ (or $\gamma_{n,m}^{(t+1)}$) should take small values and does not penalize the corresponding channel coefficients.

2) Update for Dynamic-Grid Parameters $\boldsymbol{\beta}$ *and* $\boldsymbol{\vartheta}$*:* As the optimization problems (34) and (35) are non-convex and it is challenging to find a global optimal solution, we propose to use gradient update method to update the dynamic-grid parameters $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$, i.e.,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \Delta_{\beta}^{(t)}\cdot\boldsymbol{\zeta}_{\beta}^{(t)} \tag{41}$$

$$\boldsymbol{\vartheta}^{(t+1)} = \boldsymbol{\vartheta}^{(t)} + \Delta_{\vartheta}^{(t)}\cdot\boldsymbol{\zeta}_{\vartheta}^{(t)} \tag{42}$$

where $\boldsymbol{\zeta}_{\beta}^{(t)}$ and $\boldsymbol{\zeta}_{\vartheta}^{(t)}$ are the derivatives of the objective function $f(\tilde{\mathbf{u}};\boldsymbol{\beta},\boldsymbol{\vartheta})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ at point $(\tilde{\mathbf{u}}^{(t+1)},\boldsymbol{\beta},\boldsymbol{\vartheta}^{(t)})$ and $(\tilde{\mathbf{u}}^{(t+1)},\boldsymbol{\beta}^{(t+1)},\boldsymbol{\vartheta})$, respectively. Δ_{β} and Δ_{ϑ} are the step-sizes that can be determined by the Armijo rule [29]. The detailed derivation of the gradient $\boldsymbol{\zeta}_{\beta}^{(t)}$ and $\boldsymbol{\zeta}_{\vartheta}^{(t)}$ are given in Appendix VI-D.

Algorithm 1: SRED-SPI for MU Localization With PHN.

```
Input: Received signal $\mathbf{y}$, the covariance matrix of PHN
  $\mathbf{Q}_\theta$ and $a$, $b$, $\bar{a}$, $\bar{b}$.
Output: MU position $\hat{\mathbf{p}}$ and uplink channels $\hat{\mathbf{h}}_n$, $\forall n$.
Initialize: $t = 0$, $\mathbf{u}^{(0)} = [\mathbf{w}^{(0)}; \mathbf{v}^{(0)}] = \mathbf{0}$, $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and
  $\boldsymbol{\vartheta}^{(0)} = \mathbf{0}$. Set $\tilde{\mathbf{w}}^{(0)} = [\mathrm{Re}(\mathbf{w}^{(0)}); \mathrm{Im}(\mathbf{w}^{(0)})]$ and
  $\tilde{\mathbf{v}}^{(0)} = [\mathrm{Re}(\mathbf{v}^{(0)}); \mathrm{Im}(\mathbf{v}^{(0)})]$. Compute $\mathbf{W}_y$ using (27)
  and $\mathbf{W}_y^{\frac{1}{2}}$ by Cholesky Factorization.
while not converge do
  Update $\lambda_l^{(t+1)}$ using (39).
  Update $\gamma_{n,m}^{(t+1)}$ using (40).
  Update $\tilde{\mathbf{w}}^{(t+1)}$ and $\tilde{\mathbf{v}}^{(t+1)}$ by solving problem (38).
  Update $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\vartheta}^{(t+1)}$ according to (41–42).
  $t = t + 1$.
end while
Set the estimate of parameters as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$, $\hat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}^{(t)}$,
  $\hat{\mathbf{w}} = \tilde{\mathbf{w}}_{1:NL}^{(t)} + j\tilde{\mathbf{w}}_{NL+1:2NL}^{(t)}$,
  $\hat{\mathbf{v}} = \tilde{\mathbf{v}}_{1:NM}^{(t)} + j\tilde{\mathbf{v}}_{NM+1:2NM}^{(t)}$.
Estimate $\hat{l}_s$ by (19), then the estimated position of the
  MU $\hat{\mathbf{p}}$ is given by (20) while the estimated channels $\hat{\mathbf{h}}_n$
  are given by (21).
```

The above update rules for the parameters ensures that the objective function $f(\tilde{\mathbf{u}}; \boldsymbol{\beta}, \boldsymbol{\vartheta})$ is non-decreasing and the in-exact block MM algorithm will converge to a stationary solution.

Theorem 5: The update rules in (38), (49) and (50) give a non-decreasing sequence $f(\tilde{\mathbf{u}}^{(t)}; \boldsymbol{\beta}^{(t)}, \boldsymbol{\vartheta}^{(t)})$, $t = 1, 2, 3\ldots$ Every limiting point of the iterates generated by the in-exact block MM algorithm will converge to a stationary point of problem (29).

Proof: See Appendix VI-E. ■

C. Implementation Considerations and Complexity Analysis

In the following, we list some practical implementation considerations for the proposed algorithm. Empirical experience shows that the proposed algorithm is robust to choices of parameter initialization and we can just simply set $a = \bar{a} = 1$ and $b = \bar{b} = 0.0001$. We further assign small values to τ_w and τ_v, e.g., $\tau_w = \tau_v = 0.01$. If we have prior information about the path gain between the MU and the BSs, we can initialize $\mathbf{w}_n^{(0)}$ and $\mathbf{v}_n^{(0)}$ such that $\|\mathbf{w}_n^{(0)}\|^2$ and $\|\mathbf{v}_n^{(0)}\|^2$ are equal to the path gain. Otherwise the channel coefficients are initialized near to zero, because the ground truth solution is sparse and most of the elements in the variables $\mathbf{u} = [\mathbf{w}; \mathbf{v}]$, $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ are zero. For example, we can set $\|\mathbf{w}^{(0)}\|_2 \leq Lb$, $\|\mathbf{v}^{(0)}\|_2 \leq M\bar{b}$. The off-grid parameters are initialized to zero, i.e., $\boldsymbol{\beta}^{(0)} = 0$ and $\boldsymbol{\vartheta}^{(0)} = 0$. Simulation results show that the converged performance of the proposed algorithm is not sensitive to the choice of different initial points satisfying the aforementioned conditions. The proposed SRED-SPI algorithm is summarized in Algorithm 1.

Finally, we discuss the computational complexity of the proposed algorithm. The main computational burden is listed as below:

- In the initialization step, the computational complexities in computing $\mathbf{W}_y$ and $\mathbf{W}_y^{-\frac{1}{2}}$ are $\mathcal{O}((TNN_r)^3)$.
- In each iteration, the complexity in updating $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ are $\mathcal{O}(TNN_rL)$ and $\mathcal{O}(TNN_rM_\mathrm{s})$, respectively.
- In each iteration, the complexity of updating $\tilde{\mathbf{u}}$ by solving problem (38) with typical algorithms such as LARS [30] is cubic to variable size, i.e., $\mathcal{O}((NL + M_\mathrm{s})^3)$.

This suggests that the total computational complexity of the proposed method is $\mathcal{O}((NL + M_\mathrm{s})^3)$, because the total number of grids $NL + M_\mathrm{s}$ is usually larger than N_r.

V. Numerical Experiments

In this section, we present the performance of proposed localization algorithm in massive MIMO systems and compare it with the following baselines:

- **Baseline 1**: (DiSoul [13]): The LoS and NLoS channel coefficients are recovered by $l_{2,1}$ norm minimization problem. This algorithm is ignorant about the existence of PHN. The grid is refined by sampling a denser grids around estimated position iteratively.
- **Baseline 2**: (SRED-SPI (ignore PHN)): This is the same SRED-SPI proposed in this paper but it ignores the existence of PHN, i.e., the equivalent covariance matrix is set to $\mathbf{W}_y = \mathbf{I}$ for any PHN covariance matrix.
- **Baseline 3**: (DPD): The Direct Position Determination algorithm proposed in [10] that estimate the location of the MU directly from the received signals. This is essentially a brute force ML estimator that only considers LoS path and does not incorporate the effect of PHN. The location is picked by exhaustively computing the maximum likelihood value in a 2D map and selecting the one with the maximum value.
- **Baseline 4**: (IV): The weighted Instrumental Variables algorithm proposed in [9], utilizes the AoA information which is obtained by applying beamforming [31] on the received signal (1) and selecting the angle associated with the strongest peak. Then the user location is estimated in closed form by triangulation.

In this simulation, we consider the following parameters unless otherwise stated. The MU is located randomly within a area of 100×100 m square map. This square map is sampled into $L = 20 \times 20 = 400$ grid cells and each grid cell is with size 5×5 m. Assume the origin of the coordinate system is at the center of the area, the coordinate of the four BSs are positioned at $[-150\,\mathrm{m}, 150\,\mathrm{m}]$, $[150\,\mathrm{m}, 150\,\mathrm{m}]$, $[-150\,\mathrm{m}, -150\,\mathrm{m}]$ and $[150\,\mathrm{m}, -150\,\mathrm{m}]$ respectively. The carrier frequency is $f_c =$ 30GHz and each BS is equipped with a uniform circular array (UCA) of 20 antennas. The inter antenna space is $\frac{\lambda}{2}$ for the UCA and we set $M_n = M = N_r, \forall n$. Two scatterers uniformly located are considered. A MU transmits $T = 1$ pilot symbol to the BSs for joint localization and channel estimation. The channel gain for the LoS path is modeled as Gaussian random variable, i.e., $\xi_n \sim \mathcal{CN}(0, \rho_n^2)$ where ρ_n is the corresponding path loss coefficient and is computed via [32] $\rho_n = 10^{-\frac{PL(f_c, d_n)}{20}}$ where $PL(f_c, d_n)[\mathrm{dB}] = 20\log_{10}(\frac{4\pi f_c}{c}) + 10k\log_{10}(\frac{d_n}{1m})$ is

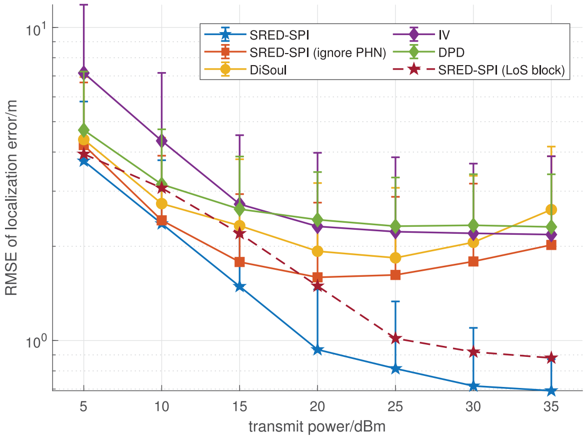


Fig. 5. RMSE and SEE of localization versus P_T at $\sigma_\theta = 4°$ for different algorithms.

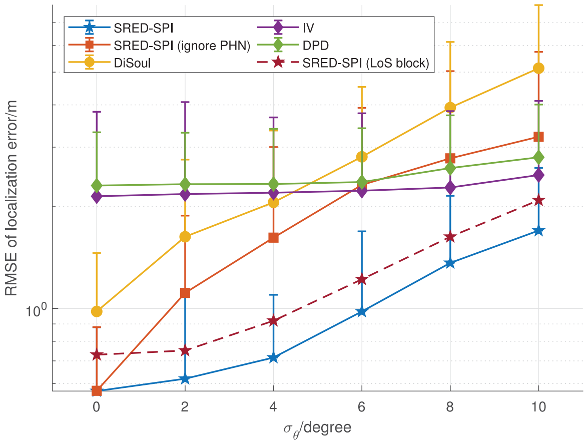


Fig. 6. RMSE and SEE of localization versus σ_θ at $P_T = 30$ dBm for different algorithms.

the pathloss at distance d_n in meters. c is the speed of light and k is the pathloss exponent. The channel gain for NLoS path is distributed as $\xi_n^i \sim \mathcal{CN}(0, \rho_{n,i}^2)$ where $\rho_{n,i} = 10^{-\frac{PL(f_c, d_n)}{20}}$. The pathloss exponent is set to $k = 2$ for LoS paths while $k = 3$ for NLoS paths. We consider PHN standard deviation $\sigma_\theta \leq 10°$. AWGN power σ^2 is set to -108 dBm while the transmit power P_T varies from 5 dBm to 35 dBm.

A. Performance of Localization

We use root-mean-square-error (RMSE) to evaluate the performance of localization accuracy, defined as $\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{j=1}^{K}[e^{(j)}]^2}$ where $e^{(j)} = \|\hat{\mathbf{p}}^j - \mathbf{p}^j\|_2$ is the localization error at j-th simulation run, $\hat{\mathbf{p}}^j$ denotes the estimated location for $\mathbf{p}^j$ at the j-th simulation run. We also compute the standard error estimate (SEE) to evaluate the spread of the estimation error, which is defined as $\mathrm{SEE} = \sqrt{\frac{1}{K}\sum_{j=1}^{K}[e^{(j)} - \bar{e}]^2}$, where $\bar{e}$ is the mean of the localization error. The SSE is represented as the vertical error bar on the RMSE points. Each point in the figure is generated by running $K = 500$ Monte-Carlo trials.

The RMSE performance and SEE for localization of different algorithms versus the transmit power P_T is presented in Fig. 5. PHN standard deviation of $\sigma_\theta = 4°$ is considered in this figure. As seen from the simulation result, we can see that the proposed method outperforms the other baselines for all transmit power considered. At high transmit power, the SRED-SPI algorithm can reach submeter accuracy, which is orders of magnitude smaller than the grid size, implying that the dynamic-grid adjustment helps to improve the localization accuracy.

The RMSE of DiSoul and SRED-SPI (ignore PHN) becomes larger for higher transmit powers. Because as P_T increases, the effective noise power (i.e., the mismatch between the measurements and the signal model) is enlarged. There are two sources of such model mismatch. The first is caused by the PHN. In this paper, we approximate the PHN matrix with first order Taylor expansion and regard it as equivalent additive Gaussian noise with covariance matrix $2\tilde{\mathbf{D}}_y\mathbf{Q}_\theta\tilde{\mathbf{D}}_y^T$. For fixed $\mathbf{Q}_\theta$, the power of effective additive noise caused by PHN increases as transmit power increases. The second effective noise is caused by the position/angle offset. For example in DiSoul baseline, the measurement matrix (for LoS) is $\sqrt{P}\bar{\mathbf{\Omega}}(\mathbf{q})$ while the ideal measurement matrix without model mismatch should be $\sqrt{P}\mathbf{\Omega}(\mathbf{q} + \Delta\boldsymbol{\epsilon})$, where $\mathbf{q}$ is the fixed grid points coordinates and $\Delta\boldsymbol{\epsilon}$ is the offset between the MU and the grid point that is nearest to the MU. Such grid offset also introduces additional effective noise whose power increases as transmit power increases.[3]

The overall effective noise introduced by these model mismatches will cause the baselines fail if they did not consider (all or part of) these effects. For example, in DiSoul, the constraint bound ϵ in [13] should reflect such effective noises, but it only depends on the AWGN power and not adjusted according to the increased transmit power. This makes the baseline algorithms more sensitive to these model mismatches, especially at higher transmit power. However, our proposed method is able to compensate such model mismatch by i) capture the PHN effect through self-interference covariance matrix and ii) introduces dynamic grid parameter that automatically compensate the grid offsets. As a result, our proposed algorithm is more robust to these model mismatches.

The other indirect and non-CS-based IV technique improves very little and has relatively larger SEE because it may select wrong LoS path and result in very large estimation error.

Fig. 6 plots the RMSE for localization of different methods versus the PHN standard deviation σ_θ at $P_T = 30$ dBm. Generally, the RMSE for each method becomes worse as PHN standard deviation increases but the proposed SRED-SPI algorithm is more robust to PHN and outperforms all the others for small to medium PHN considered. Note that for $\sigma_\theta = 0°$, i.e., when there is no PHN, the SRED-SPI also outperforms DiSoul due to the dynamic grid adjustment that compensated the location and angle quantization error. The performance of DPD and IV is bad for all PHN strength simply because they did not consider NLoS paths and may wrongly selecte the true AoA from the MU.

[3] To see that, one can also use first order approximation to derive the equivalent additive noise by expanding $\sqrt{P}\mathbf{\Omega}(\cdot)$ near $\sqrt{P}\mathbf{\Omega}(\mathbf{q})$, like that for PHN approximation.

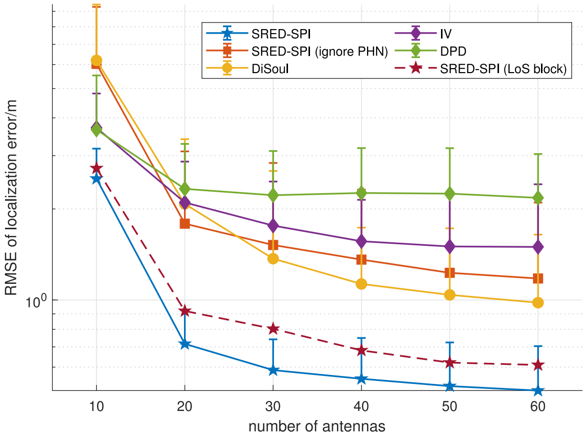


Fig. 7. RMSE and SEE of localization versus antenna number at $\sigma_\theta = 4^\circ$ and $P_T = 30\,\text{dBm}$ for different algorithms.

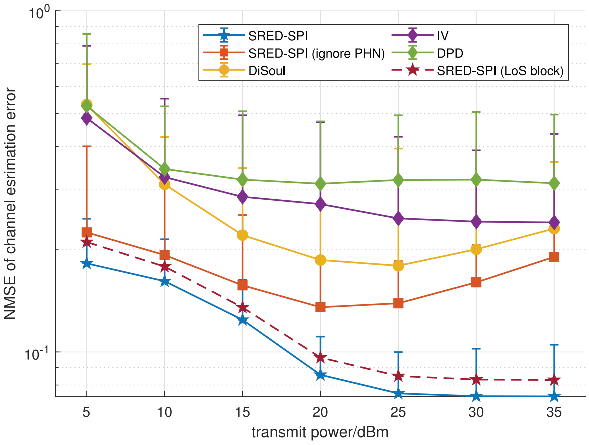


Fig. 8. NMSE and SEE of channel estimation versus P_T at $\sigma_\theta = 6^\circ$ for different algorithms.

The localization error increases dramatically as σ_θ increases for DiSoul because the equivalent additive noise introduced by PHN became the dominate limiting factor when P_T is large and the constraint bound ϵ simply fails to capture such significant noise error.

Fig. 7 evaluates the performance of localization algorithms versus the number of antennas in each base station. It turns out the proposed SRED-SPI algorithm outperforms all the other baselines for all the antenna number considered. The increase in the number of antennas enhanced the angular resolution of each BS, which enables the BSs to get more accurate AoA estimates as well as resolve the LoS and NLoS paths more precisely. As a result, the localization accuracy improves as the number of antennas of each BS increases.

In order to show the robustness of our algorithm to the availability of LoS paths, in the Fig. 5–Fig. 7, we also simulate the scenario when the LoS path between the MU and BS4 (as shown in Fig. 1) is blocked during the localization process. The RMSE of localization error are presented as dashed lines. The simulation results show that even if BS4 can not receive LoS path from the MU, the proposed SRED-SPI algorithm can still work quite well and gives satisfactory localization accuracy. Generally, we do not need every BS to receive LoS path from the MU. We only need a few BS (e.g., 3 BSs with active LoS paths) for the proposed algorithm to perform well. Consequently, our algorithm is quite robust to the blocking of LoS paths.

B. Performance of Channel Estimation

We used normalized mean square error (NMSE) to evaluate the performance of channel estimation accuracy, defined as $\text{NMSE} = \frac{1}{K}\sum_{j=1}^{K}\eta^{(j)}$, where $\eta^{(j)} = \frac{\sum_{n=1}^{N}\|\mathbf{h}_n^{(j)} - \hat{\mathbf{h}}_n^{(j)}\|_2^2}{\sum_{n=1}^{N}\|\mathbf{h}_n^{(j)}\|_2^2}$ is the normalized channel estimation error across all the N BSs at j-th simulation run. $\mathbf{h}_n^{(j)}$ and $\hat{\mathbf{h}}_n^{(j)}$ denote the ground truth uplink channel and the corresponding estimated channel of the n-th BSs at the j-th simulation run, respectively. K is the total number of simulation run. We also compute the standard error estimate for channel estimation and present them as vertical standard error bar with the NMSE in the figures.

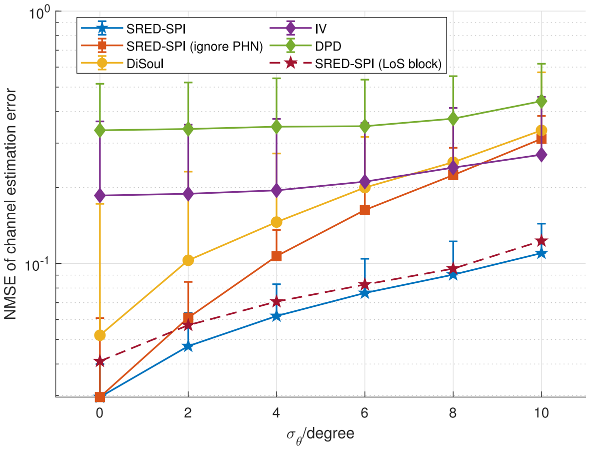


Fig. 9. NMSE and SEE of channel estimation versus σ_θ at $P_T = 30\,\text{dBm}$ for different algorithms.

Fig. 8 illustrates the NMSE performance of channel estimation versus transmit power under different algorithms at $\sigma_\theta = 6^\circ$. It turns out the proposed algorithm achieves the best channel estimation accuracy over the other baselines. The channel estimation accuracy of DiSoul and SRED-SPI (ignore PHN) also get worse as transmit power increases after some point. Such phenomenon is similar to localization performance, and is a result of larger perturbation induced by model mismatch when P_T increases. DiSoul performs bad for small transmit power because $\mathbf{h} = \mathbf{0}$ might be a trivial solution to the $l_{2,1}$ norm minimization problem at low SNR region [13]. IV and DPD generally have large channel estimation error for all the transmit power considered due to the possibility of wrongly selecting the LoS path for certain BSs, thus resulting in a very large channel estimation error. In Fig. 9, the performance of channel estimation as a function of PHN standard deviation is shown for different algorithms at $P_T = 30\,\text{dBm}$. In general, the channel estimation accuracy degrades as σ_θ increases, but the proposed SRED-SPI is more robust to PHN in terms of channel estimation performance as well. Note that at $\sigma_\theta = 0^\circ$, i.e., when there is no presence of PHN, the proposed SRED-SPI also outperforms DiSoul in terms of channel estimation acuracy. This is because SRED-SPI solves a series of $l_{2,1}$ norm regularized minimization problems iteratively and the reweighting factors tends to yield a sparser solution in each iteration than the one-shot $l_{2,1}$ norm minimization problem solved by DiSoul.

We also simulate the NMSE of channel estimation performance when the LoS path between the MU and BS4 is blocked. The results are shown as dashed lines in Fig. 8 and Fig. 9. It turns out the channel estimation module is also robust to LoS blocking.

VI. Conclusion

This paper tackles the localization problem of massive MIMO systems with the presence of PHN and multipath. We proposed a sparse representation model for localization with dynamic grid and transformed the problem to a parameterized sparse recovery problem via a MAP problem. We proposed a SRED-SPI algorithm to solve the localization problem. Specifically, in order to capture the effect of PHN at the receiver, we proposed a tractable approximation to the likelihood term and approximate the PHN as additive perturbation by introducing an effective covariance matrix that is related to the received signals. Then, we proposed to use inexact block MM to find a stationary solution of the approximated problem. Finally, the location of the MU is estimated and uplink channel information is recovered correspondingly. The simulation results shows that our proposed algorithm achieves superior localization and channel estimation performance over the baselines under small to moderate PHN.

Appendix

A. *Proof of Proposition 1.*

With the prior on PHN defined in (9), we can write the approximated likelihood term as

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\exp\left(-\left\|\tilde{\mathbf{D}}_y\boldsymbol{\Theta}+\mathbf{b}\right\|^2\right)\right]$$
$$=\frac{1}{\sqrt{\det\left(2\pi\mathbf{Q}_\theta\right)}}\cdot\int_{\boldsymbol{\Theta}}d\boldsymbol{\Theta}\exp\left(-\left\|\tilde{\mathbf{D}}_y\boldsymbol{\Theta}+\mathbf{b}\right\|^2\right)\exp\left(-\boldsymbol{\Theta}^T\frac{\mathbf{Q}_\theta^{-1}}{2}\boldsymbol{\Theta}\right)$$
$$\frac{1}{\sqrt{\det\left(2\pi\mathbf{Q}_\theta\right)}}\cdot\int_{\boldsymbol{\Theta}}d\boldsymbol{\Theta}\exp\left(-\boldsymbol{\Theta}^T\frac{1}{2}\mathbf{Q}^{-\mathbf{1}}\boldsymbol{\Theta}-2\mathbf{b}^T\tilde{\mathbf{D}}_y\boldsymbol{\Theta}-\|\mathbf{b}\|^2\right) \tag{43}$$

$$=\sqrt{\frac{\det\left(2\pi\mathbf{Q}\right)}{\det\left(2\pi\mathbf{Q}_\theta\right)}}\exp\left(-\|\mathbf{b}\|^2+\mathbf{b}^T\tilde{\mathbf{D}}_y\left(2\mathbf{Q}\right)\tilde{\mathbf{D}}_y^T\mathbf{b}\right) \tag{44}$$
$$=\sqrt{\frac{\det\left(2\pi\mathbf{Q}\right)}{\det\left(2\pi\mathbf{Q}_\theta\right)}}\exp\left(-\mathbf{b}^T\left(\mathbf{I}-\tilde{\mathbf{D}}_y\left[\frac{1}{2}\mathbf{Q}_\theta^{-1}+\tilde{\mathbf{D}}_y^T\tilde{\mathbf{D}}_y\right]^{-1}\tilde{\mathbf{D}}_y^T\right)\mathbf{b}\right)$$
$$=\sqrt{\frac{1}{\det\left(\mathbf{I}+2\tilde{\mathbf{D}}_y\mathbf{Q}_\theta\tilde{\mathbf{D}}_y^T\right)}}\exp\left(-\mathbf{b}^T\left[\mathbf{I}+\tilde{\mathbf{D}}_y\left(2\mathbf{Q}_\theta\right)\tilde{\mathbf{D}}_y^T\right]^{-1}\mathbf{b}\right) \tag{45}$$

where we define $\mathbf{Q}^{-1}=2\tilde{\mathbf{D}}_y^T\tilde{\mathbf{D}}_y+\mathbf{Q}_\theta^{-1}$. (44) holds since (43) is integrating over a probability density function over $\boldsymbol{\Theta}$, while (45) holds with Woodbury identity.

B. *Proof of Lemma 2.*

Let $p(\lambda|\tilde{\mathbf{w}}^{(t)})$ and $p(\boldsymbol{\gamma}|\tilde{\mathbf{v}}^{(t)})$ be any arbitrary conditional distribution of λ and $\boldsymbol{\gamma}$, we have

$$u_w\left(\tilde{\mathbf{w}}^{(t)}|\tilde{\mathbf{w}}^{(t)}\right)=\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log\left(\frac{p\left(\tilde{\mathbf{w}}^{(t)},\lambda\right)}{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}\right)d\lambda$$
$$=\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log p\left(\tilde{\mathbf{w}}^{(t)}\right)d\lambda$$
$$=\log p\left(\tilde{\mathbf{w}}^{(t)}\right)$$

where we used the fact that $p(\lambda|\tilde{\mathbf{w}}^{(t)})$ is a probability distribution s.t. $\int_\lambda p(\lambda|\tilde{\mathbf{w}}^{(t)})d\lambda=1$. Similarly, we have $u_v(\tilde{\mathbf{v}}^{(t)}|\tilde{\mathbf{v}}^{(t)})=\log p(\tilde{\mathbf{v}}^{(t)})$. Then it is clear (30) holds true.

To prove (31), we rewrite the $\log p(\tilde{\mathbf{w}})$ as

$$\log p\left(\tilde{\mathbf{w}}\right)=\log\int_\lambda p\left(\tilde{\mathbf{w}},\lambda\right)d\lambda$$
$$=\log\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\frac{p\left(\tilde{\mathbf{w}},\lambda\right)}{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}d\lambda$$
$$\geq\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log\frac{p\left(\tilde{\mathbf{w}},\lambda\right)}{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}d\lambda-\tau_w\left\|\tilde{\mathbf{w}}-\tilde{\mathbf{w}}^{(t)}\right\|^2 \tag{46}$$

where we used Jensen's inequality in obtaining (46). Similarly, we have $u_v(\tilde{\mathbf{v}}|\tilde{\mathbf{v}}^{(t)})\leq\log p(\tilde{\mathbf{v}})$. Then one can easily check (31) holds true. Finally, to prove (32), we have

$$\frac{\partial u_w\left(\tilde{\mathbf{w}}|\tilde{\mathbf{w}}^{(t)}\right)}{\partial\tilde{\mathbf{w}}}|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}$$
$$=\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\frac{\partial}{\partial\tilde{\mathbf{w}}}\log p\left(\tilde{\mathbf{w}},\lambda\right)d\lambda-2\tau_w\left(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}^{(t)}\right)|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}$$
$$=\int_\lambda\frac{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}{p\left(\tilde{\mathbf{w}}^{(t)},\lambda\right)}\frac{\partial}{\partial\tilde{\mathbf{w}}}p\left(\tilde{\mathbf{w}},\lambda\right)d\lambda|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}$$
$$=\frac{1}{p\left(\tilde{\mathbf{w}}^{(t)}\right)}\cdot\frac{\partial}{\partial\tilde{\mathbf{w}}}\int_\lambda p\left(\tilde{\mathbf{w}},\lambda\right)d\lambda|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}$$
$$=\frac{1}{p\left(\tilde{\mathbf{w}}^{(t)}\right)}\cdot\frac{\partial}{\partial\tilde{\mathbf{w}}}p\left(\tilde{\mathbf{w}}\right)|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}, \tag{47}$$

while on the other hand, we have $\frac{\log p(\tilde{\mathbf{w}})}{\partial\tilde{\mathbf{w}}}|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}=\frac{1}{p(\tilde{\mathbf{w}}^{(t)})}\frac{\partial}{\partial\tilde{\mathbf{w}}}p(\tilde{\mathbf{w}})$. Then we have

$$\frac{\log p\left(\tilde{\mathbf{w}}\right)}{\partial\tilde{\mathbf{w}}}|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}=\frac{\partial u_w\left(\tilde{\mathbf{w}}|\tilde{\mathbf{w}}^{(t)}\right)}{\partial\tilde{\mathbf{w}}}|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}.$$

Similarly, we have

$$\frac{\log p\left(\tilde{\mathbf{v}}\right)}{\partial\tilde{\mathbf{v}}}|_{\tilde{\mathbf{v}}=\tilde{\mathbf{v}}^{(t)}}=\frac{\partial u_w\left(\tilde{\mathbf{v}}|\tilde{\mathbf{v}}^{(t)}\right)}{\partial\tilde{\mathbf{v}}}|_{\tilde{\mathbf{v}}=\tilde{\mathbf{v}}^{(t)}},$$

from which we can conclude that (32) holds true.

C. Proof of Lemma 3.

We first focus on the surrogate function for the LoS prior. At the point $\tilde{\mathbf{w}}^{(t)}$, the majorized problem for $\tilde{\mathbf{w}}$ is given by

$$\begin{aligned}&\arg\max_{\tilde{\mathbf{w}}} u_w\left(\tilde{\mathbf{w}}|\tilde{\mathbf{w}}^{(t)}\right)\\&=\arg\max_{\tilde{\mathbf{w}}}\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log\left(\frac{p\left(\tilde{\mathbf{w}}|\lambda\right)p\left(\lambda\right)}{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}\right)d\lambda\\&-\tau_w\left\|\tilde{\mathbf{w}}-\tilde{\mathbf{w}}^{(t)}\right\|^2\\&=\arg\max_{\tilde{\mathbf{w}}}\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log\left(p\left(\tilde{\mathbf{w}}|\lambda\right)\right)d\lambda-\tau_w\left\|\tilde{\mathbf{w}}-\tilde{\mathbf{w}}^{(t)}\right\|^2\end{aligned}\tag{48}$$

For given factorial probability model $p(\lambda|\tilde{\mathbf{w}}^{(t)})$, since we have closed form expression for $\log(p(\tilde{\mathbf{w}}|\lambda))$, the first term in (48) can be computed by

$$\begin{aligned}&\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log\left(p\left(\tilde{\mathbf{w}}|\lambda\right)\right)d\lambda\\&=\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\log\left\{\prod_{l=1}^{L}e^{-\lambda_l\|\tilde{\mathbf{w}}_l\|_2}\right\}d\lambda+C\\&=-\sum_{l=1}^{L}\left[\int_\lambda p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)\lambda_l d\lambda\right]\|\tilde{\mathbf{w}}_l\|_2+C\\&=-\sum_{l=1}^{L}\left\langle\lambda_l\right\rangle_{p\left(\lambda_l|\tilde{\mathbf{w}}^{(t)}\right)}\|\tilde{\mathbf{w}}_l\|_2+C\end{aligned}$$

where C represent the constant term. As a result we have

$$\begin{aligned}u_w\left(\tilde{\mathbf{w}}|\tilde{\mathbf{w}}^{(t)}\right)=&\sum_{l=1}^{L}\left\langle\lambda_l\right\rangle_{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}\|\tilde{\mathbf{w}}_l\|_2\\&+\tau_w\left\|\tilde{\mathbf{w}}-\tilde{\mathbf{w}}^{(t)}\right\|^2+C\end{aligned}\tag{49}$$

Similarly, we have

$$\begin{aligned}u_v\left(\tilde{\mathbf{v}}|\tilde{\mathbf{v}}^{(t)}\right)=&\sum_{n=1}^{N}\sum_{m=1}^{M_n}\left\langle\gamma_{n,m}\right\rangle_{p\left(\boldsymbol{\gamma}|\tilde{\mathbf{v}}^{(t)}\right)}\|\tilde{\mathbf{v}}_{n,m}\|_2\\&+\tau_v\left\|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}^{(t)}\right\|^2+C.\end{aligned}\tag{50}$$

Substituting (49) and (50) into (33), we obtain the desired optimization problem.

Now we consider the derivation of $\left\langle\lambda_l\right\rangle_{p\left(\lambda|\tilde{\mathbf{w}}^{(t)}\right)}$ and $\left\langle\gamma_{n,m}\right\rangle_{p\left(\boldsymbol{\gamma}|\tilde{\mathbf{v}}^{(t)}\right)}$. With the choice of factorial Gamma distribution on λ_l, i.e., $p(\lambda_l)=b^a\lambda_l^{a-1}e^{-b\lambda_l}/\Gamma(a)$, we can compute the probability density on LoS channel coefficients $\tilde{\mathbf{w}}_l$ as

$$\begin{aligned}p\left(\tilde{\mathbf{w}}_l\right)&=\int_0^\infty p\left(\lambda_l\right)p\left(\tilde{\mathbf{w}}_l|\lambda_l\right)d\lambda_l\\&=\frac{C_{2N}a^{2N}b^a}{\left(b+\|\tilde{\mathbf{w}}_l\|_2\right)^{2N+a}}\end{aligned}$$

Then, we compute the probability of λ_l conditioned on $\tilde{\mathbf{w}}_l$:

$$\begin{aligned}p\left(\lambda_l|\tilde{\mathbf{w}}_l\right)&=\frac{p\left(\tilde{\mathbf{w}}_l|\lambda_l\right)p\left(\lambda_l\right)}{p\left(\tilde{\mathbf{w}}_l\right)}\\&=\frac{C_{2N}b^a}{\Gamma(a)}\lambda_l^{a-1}e^{-b\lambda_l}\lambda_l^{2N}e^{-\lambda_l\|\tilde{\mathbf{w}}_l\|_2}\frac{\left(b+\|\tilde{\mathbf{w}}_l\|_2\right)^{2N+a}}{C_{2N}a^{2N}b^a}\\&=\frac{\left(b+\|\tilde{\mathbf{w}}_l\|_2\right)^{2N+a}}{\Gamma\left(a+2N\right)}\lambda_l^{2N+a-1}e^{-\left(b+\|\tilde{\mathbf{w}}_l\|_2\right)\lambda_l}\\&=\Gamma\left(\lambda_l;a+2N,b+\|\tilde{\mathbf{w}}_l\|_2\right),\end{aligned}$$

i.e., $p(\lambda_l|\tilde{\mathbf{w}}_l)$ is Gamma distributed so the mean is given by $(a+2N)/(b+\|\tilde{\mathbf{w}}_l\|_2)$. Similarly, for the NLoS channel coefficients, we can compute the probability of $\gamma_{n,m}$ conditioned on $\tilde{\mathbf{v}}_{n,m}$, which is given by

$$p\left(\gamma_{n,m}|\tilde{\mathbf{v}}_{n,m}\right)=\Gamma\left(\gamma_{n,m};\bar{a}+2,\bar{b}+\|\tilde{\mathbf{v}}_{n,m}\|_2\right).$$

D. The Expression for the Gradient of Dynamic-Grid Updates

For simplicity, we first consider $T=1$. The derivative $\boldsymbol{\zeta}_\beta$ can be computed as $\boldsymbol{\zeta}_\beta=[\boldsymbol{\zeta}(\boldsymbol{\beta}_1);\ldots;\boldsymbol{\zeta}(\boldsymbol{\beta}_L)]$, where

$$\boldsymbol{\zeta}\left(\boldsymbol{\beta}_l\right)=\begin{bmatrix}\left(-\mathbf{c}_1^T+\mathbf{c}_2\left(\boldsymbol{\beta}\right)^T\mathbf{W}_y^{-1}\right)\tilde{\mathbf{c}}_x\left(\boldsymbol{\beta}_l\right)\\\left(-\mathbf{c}_1^T+\mathbf{c}_2\left(\boldsymbol{\beta}\right)^T\mathbf{W}_y^{-1}\right)\tilde{\mathbf{c}}_y\left(\boldsymbol{\beta}_l\right)\end{bmatrix},$$

with $\mathbf{c}_1=\mathbf{W}_y^{-\frac{T}{2}}\tilde{\mathbf{y}}_{W/v}$, $\tilde{\mathbf{y}}_{W/v}=\mathbf{W}_y^{-\frac{1}{2}}(\tilde{\mathbf{y}}-\tilde{\boldsymbol{\Phi}}(\boldsymbol{\vartheta})\tilde{\mathbf{v}})$, $\mathbf{c}_2(\boldsymbol{\beta})=\tilde{\boldsymbol{\Omega}}(\boldsymbol{\beta})\tilde{\mathbf{w}}$, $\tilde{\mathbf{c}}_x(\boldsymbol{\beta}_l)=[\mathrm{Re}(\mathbf{c}_x(\boldsymbol{\beta}_l));\mathrm{Im}(\mathbf{c}_x(\boldsymbol{\beta}_l))]$ and $\tilde{\mathbf{c}}_y(\boldsymbol{\beta}_l)=[\mathrm{Re}(\mathbf{c}_y(\boldsymbol{\beta}_l));\mathrm{Im}(\mathbf{c}_y(\boldsymbol{\beta}_l))]$. $\mathbf{c}_x(\boldsymbol{\beta}_l)=2x_1[w_{1,q}\mathbf{a}_1'(\boldsymbol{\beta}_l)c_{3,1};\ldots;w_{N,q}\mathbf{a}_N'(\boldsymbol{\beta}_l)c_{3,N}]$ and $\mathbf{c}_y(\boldsymbol{\beta}_l)=2x_1[w_{1,q}\mathbf{a}_1'(\boldsymbol{\beta}_l)c_{4,1};\ldots;w_{N,q}\mathbf{a}_N'(\boldsymbol{\beta}_l)c_{4,N}]$, where $\mathbf{a}_n'(\boldsymbol{\beta}_l)=d\mathbf{a}_n(\mathbf{q}_l+\boldsymbol{\beta}_l)/d\varphi_n(\mathbf{q}_l+\boldsymbol{\beta}_l)$, $c_{3,n}=-(q_{l,y}+\Delta y_l-p_{n,y})/\|\mathbf{q}_l+\boldsymbol{\beta}_l-\mathbf{p}_n\|^2$ and $c_{4,n}=(q_{q,x}+\Delta x_q-p_{n,x})/\|\mathbf{q}_l+\boldsymbol{\beta}_l-\mathbf{p}_n\|^2$.

The derivative of $\boldsymbol{\zeta}_\vartheta$ can be calculated as $\boldsymbol{\zeta}_\vartheta=[\boldsymbol{\zeta}_1(\boldsymbol{\vartheta});\ldots;\boldsymbol{\zeta}_N(\boldsymbol{\vartheta})]$, where $\boldsymbol{\zeta}_n(\boldsymbol{\vartheta})=[\zeta(\vartheta_{n,1}),\ldots,\zeta(\vartheta_{n,M_n})]^T$. Each element can be computed by

$$\zeta\left(\vartheta_{n,m}\right)=\left(-\mathbf{c}_5^T+\mathbf{c}_6\left(\boldsymbol{\beta}\right)^T\mathbf{W}_y^{-1}\right)\tilde{\mathbf{c}}_\vartheta\left(\vartheta_{n,m}\right)$$

$\mathbf{c}_5=\mathbf{W}_y^{-\frac{T}{2}}\tilde{\mathbf{y}}_{W/w}$, $\tilde{\mathbf{y}}_{W/w}=\mathbf{W}_y^{-\frac{1}{2}}(\tilde{\mathbf{y}}-\tilde{\boldsymbol{\Omega}}(\boldsymbol{\beta})\tilde{\mathbf{w}})$, $\mathbf{c}_6(\boldsymbol{\beta})=\tilde{\boldsymbol{\Phi}}(\boldsymbol{\vartheta})\tilde{\mathbf{v}}$, $\tilde{\mathbf{c}}_\vartheta(\vartheta_{n,m})=[\mathrm{Re}(\mathbf{c}_\vartheta(\vartheta_{n,m}));\ \mathrm{Im}(\mathbf{c}_\vartheta(\vartheta_{n,m}))]$, $\mathbf{c}_\vartheta(\vartheta_{n,m})=2x_1[\mathbf{0}_{(n-1)N_r};\ v_{n,m}\mathbf{a}_n'(\vartheta_{n,m});\ \mathbf{0}_{(N-n)N_r}]$, $\mathbf{a}_n'(\vartheta_{n,m})=d\mathbf{a}_n(\phi_m+\vartheta_{n,m})/d\vartheta_{n,m}$. The result can be extended to general T easily.

E. Proof of Lemma 5

With the property of the surrogate function and gradient update, the non-decreasing property can be obtained by

$$\begin{aligned}f\left(\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)}\right)&=u\left(\tilde{\mathbf{u}}^{(t)}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)}\right)\\&\le u\left(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)}\right)\qquad(51)\\&\le f\left(\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)}\right)\\&=u\left(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)}\right)\qquad(52)\end{aligned}$$

$$\le u\left(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta}^{(t+1)},\boldsymbol{\vartheta}^{(t)}\right) \tag{53}$$

$$\le u\left(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta}^{(t+1)},\boldsymbol{\vartheta}^{(t+1)}\right)=f\left(\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)}\right), \tag{54}$$

where the equality in (53) and (54) holds only when the gradient with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ is zero, i.e., the gradient update in (41) and (42) strictly increases the surrogate function value as well as the original objective function value whenever $\frac{\partial u(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta},\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\beta}}\neq 0$ and $\frac{\partial u(\tilde{\mathbf{u}}^{(t+1)}|\tilde{\mathbf{u}}^{(t+1)};\boldsymbol{\beta}^{(t+1)},\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}\neq 0$. The update step in (38) obtained by maximizing the surrogate function always increases the value of surrogate function and the original objective function whenever $\frac{\partial u(\tilde{\mathbf{u}}|\tilde{\mathbf{u}}^{(t)};\boldsymbol{\beta}^{(t)},\boldsymbol{\vartheta}^{(t)})}{\partial\tilde{\mathbf{u}}}\neq 0$. Besides, the objective function is upper bounded by 0. Therefore, the non-decreasing sequence converges to a limit point denoted by $\bar{p}$. Denote all the updated dynamic-grid parameters $(\tilde{\mathbf{u}},\boldsymbol{\beta},\boldsymbol{\vartheta})$ as a single block $\mathbf{r}$, where $\mathbf{r}$ includes three blocks of parameters where $\mathbf{r}_1=\tilde{\mathbf{u}}$, $\mathbf{r}_2=\boldsymbol{\beta}$ and $\mathbf{r}_3=\boldsymbol{\vartheta}$. The objective value will keep increasing and we must have

$$\lim_{t\to\infty}\frac{\partial u\left(\mathbf{r}_1|\mathbf{r}_1^{(t)};\mathbf{r}_2^{(t)},\mathbf{r}_3^{(t)}\right)}{\partial\mathbf{r}_1}=0, \tag{55}$$

$$\lim_{t\to\infty}\frac{\partial u\left(\mathbf{r}_1^{(t+1)}|\mathbf{r}_1^{(t+1)};\mathbf{r}_2,\mathbf{r}_3^{(t)}\right)}{\partial\mathbf{r}_2}=0, \tag{56}$$

$$\lim_{t\to\infty}\frac{\partial u\left(\mathbf{r}_1^{(t+1)}|\mathbf{r}_1^{(t+1)};\mathbf{r}_2^{(t+1)},\mathbf{r}_3\right)}{\partial\mathbf{r}_3}=0, \tag{57}$$

(otherwise, the objective will keep increasing to infinity, which contradicts with the fact that the objective function is bounded above). Then according to (55–57), the property of gradient update and the strong convexity of $u(\mathbf{r}_1|\mathbf{r}_1^{(t)};\mathbf{r}_2^{(t)},\mathbf{r}_3^{(t)})$ with respect to $\mathbf{r}_1$, we have

$$\lim_{t\to\infty}\left\|\mathbf{r}_i^{(t+1)}-\mathbf{r}_i^{(t)}\right\|=0,\ \forall i. \tag{58}$$

Denote $\mathbf{r}_{-i}^{(t)}=\{\mathbf{r}_1^{(t+1)},\ldots,\mathbf{r}_{i-1}^{(t+1)},\mathbf{r}_{i+1}^{(t)},\ldots,\mathbf{r}_3^{(t)}\}$, the property in (58) insures that all the sequences $\{\mathbf{r}_i^{(t)},\mathbf{r}_{-i}^{(t)}\}$, $i=1,2,3$ converges to the same set of limiting points, denoted by $\bar{\mathbf{r}}$, i.e., $\lim_{t\to\infty}\{\mathbf{r}_i^{(t)},\mathbf{r}_{-i}^{(t)}\}=\bar{\mathbf{r}},\ i=1,2,3$. Let $\{\mathbf{r}_i^{(t_j)},\mathbf{r}_{-i}^{(t_j)},j=1,2,\ldots\}$ denotes a subsequence that converges to $\bar{\mathbf{r}}$. If $\bar{\mathbf{r}}$ is not a stationary point of the original objective function, then $\frac{\partial f(\mathbf{r})}{\partial\mathbf{r}}|_{\mathbf{r}=\bar{\mathbf{r}}}\neq 0$. From (58) we can conclude that at least one of (55–57) does not hold with the subsequence, which results in a contradiction. Therefore, every limiting point $\bar{\mathbf{r}}$ must be a stationary point of the original objective function.

References

[1] F. Rusek *et al.*, “Scaling up MIMO: Opportunities and challenges with very large arrays,” *IEEE Signal Process. Mag.,*vol. 30, no. 1, pp. 40–60, 2012.

[2] T. L. Marzetta *et al.*, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, 2014.

[4] J. Hoydis, S. Ten Brink, and M. Debbah, “Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?” vol. 31, no. 2, pp. 160–171, Feb. 2013.

[5] Z. Gao, L. Dai, Z. Wang, and S. Chen, “Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO,” *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[6] J. Dai, A. Liu, and V. K. Lau, “FDD massive MIMO channel estimation with arbitrary 2D-array geometry,” *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2584–2599, May 2018.

[7] M. Gavish and A. J. Weiss, “Performance analysis of bearing-only target location algorithms,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, no. 3, pp. 817–828, Jul. 1992.

[8] L. M. Kaplan, Q. Le, and N. Molnar, “Maximum likelihood methods for bearings-only target localization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 5, pp. 3001–3004.

[9] K. Doğançay, “Passive emitter localization using weighted instrumental variables,” *Signal Process.*, vol. 84, no. 3, pp. 487–497, 2004.

[10] A. J. Weiss, “Direct position determination of narrowband radio frequency transmitters,” *IEEE Signal Process. Lett.*, vol. 11, no. 5, pp. 513–516, May 2004.

[11] A. J. Weiss and A. Amar, “Direct position determination of multiple radio signals,” *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 37–49, 2005.

[12] V. Savic and E. G. Larsson, “Fingerprinting-based positioning in distributed massive MIMO systems,” in *Proc. IEEE 82nd Veh. Technol. Conf. (Fall)*, 2015, pp. 1–5.

[13] N. Garcia, H. Wymeersch, E. G. Larsson, A. M. Haimovich, and M. Coulon, “Direct localization for massive MIMO,” *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2475–2487, May 2017.

[14] H. Godrich and A. M. Haimovich, “Localization performance of coherent MIMO radar systems subject to phase synchronization errors,” in *Proc. 4th Int. Symp. Commun., Control Signal Process.*, 2010, pp. 1–5.

[15] R. Combes and S. Yang, “An approximate ML detector for MIMO channels corrupted by phase noise,” *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1176–1189, Mar. 2018.

[16] H. Mehrpouyan, A. A. Nasir, S. D. Blostein, T. Eriksson, G. K. Karagiannidis, and T. Svensson, “Joint estimation of channel and oscillator phase noise in MIMO systems,” *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4790–4807, Sep. 2012.

[17] W. Robins, *Phase noise in Signal Sources: Theory and Applications*, vol. 9. London, U.K.: IET, 1984.

[18] A. G. Armada, “Understanding the effects of phase noise in orthogonal frequency division multiplexing (OFDM),” *IEEE Trans. Broadcast.*, vol. 47, no. 2, pp. 153–159, Jun. 2001.

[19] M. A. Richards, “Coherent integration loss due to white Gaussian phase noise,” *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 208–210, Jul. 2003.

[20] Y.-F. Kuo, C.-L. Hsiao, and H.-C. Wei, “Phase noise analysis of 28 GHz phase-locked oscillator for next generation 5G system,” in *Proc. IEEE 6th Global Conf. Consum. Electron.*, 2017, pp. 1–2.

[21] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[22] J. Salo *et al.*, “MATLAB implementation of the 3GPP spatial channel model,” 3GPP, Sophia Antipolis Cedex, France, TR 25.996, Jan. 2005.

[23] K. Haneda *et al.*, “5G 3GPP-like channel models for outdoor urban microcellular and macrocellular environments,” in *Proc. IEEE 83rd Veh. Technol. Conf. (Spring)*, 2016, pp. 1–7.

[24] Q. Spencer, M. Rice, B. Jeffs, and M. Jensen, “Indoor wideband time/angle of arrival multipath propagation results,” in *Proc. IEEE 47th Veh. Technol. Conf.. Technol. Motion*, vol. 3, 1997, pp. 1410–1414.

[25] I. W. Selesnick, “The estimation of laplace random vector in AWGN and the generalized incomplete gamma function,” *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3482–3496, Aug. 2008.

[26] P. Garrigues and B. A. Olshausen, “Group sparse coding with a laplacian scale mixture prior,” in *Proc. Advances Neural Inf. Process. Syst.*, 2010, pp. 676–684.

[27] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[28] B. D. Rao and K. Kreutz-Delgado, “An affine scaling methodology for best basis selection,” *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

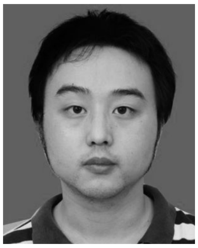
[29] M. Razaviyayn, “Successive convex approximation: Analysis and applications,” Ph.D. Dissertation, Faculty Grad. School, Univ. Minnesota, Minneapolis, MN, USA, 2014.

[30] B. Efron *et al.*, “Least angle regression,” *Annals Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[31] L. C. Godara, “Application of antenna arrays to mobile communications. ii. Beam-forming and direction-of-arrival considerations,” *Proc. IEEE*, vol. 85, no. 8, pp. 1195–1245, 1997.

[32] S. Sun, T. A. Thomas, T. S. Rappaport, H. Nguyen, I. Z. Kovacs, and I. Rodriguez, “Path loss, shadow fading, and line-of-sight probability models for 5G urban macro-cellular scenarios,” in *Proc. Globecom Workshops*, 2015, pp. 1–7.

Xuanyu Zheng (Student Member, IEEE) received the B.Eng. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2017. He is currently working toward the Ph.D. degree at the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology (HKUST), Hong Kong. His research interests include wireless communication and compressive sensing.



An Liu (Senior Member, IEEE) received the Ph.D. and B.S. degrees in electrical engineering from Peking University, China, in 2011 and 2004, respectively. From 2008 to 2010, he was a Visiting Scholar with the Department of ECEE, University of Colorado at Boulder. He has been a Postdoctoral Research Fellow in 2011–2013, Visiting Assistant Professor in 2014, and Research Assistant Professor in 2015–2017, with the Department of ECE, HKUST. He is currently a Distinguished Research Fellow with the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include wireless communications, stochastic optimization and compressive sensing.

Vincent K. N. Lau (Fellow, IEEE) received the B.Eng. degree from the University of Hong Kong (1989–1992) and the Ph.D. degree from the Cambridge University (1995–1997). He was with Bell Labs from 1997–2004 and the Department of ECE, Hong Kong University of Science and Technology (HKUST) in 2004. He is currently a Chair Professor and the Founding Director of Huawei-HKUST Joint Innovation Lab at HKUST. His current research focus includes wireless communications for 5G systems, content-centric wireless networking, wireless networking for mission-critical control, and cloud-assisted autonomous systems.