PAPER

Robust Bilinear Form Identification: A Subgradient Method with Geometrically Decaying Stepsize in the Presence of Heavy-Tailed Noise

Guowei YANG[†a)], ***Nonmember***

SUMMARY This paper delves into the utilisation of the subgradient method with geometrically decaying stepsize for Bilinear Form Identification. We introduce the iterative Wiener Filter, an l_2 regression method, and highlight its limitations when confronted with noise, particularly heavy-tailed noise. To address these challenges, the paper suggests employing the l_1 regression method with a subgradient method utilizing a geometrically decaying step size. The effectiveness of this approach is compared to existing methods, including the ALS algorithem. The study demonstrates that the l_1 algorithm, especially when paired with the proposed subgradient method, excels in stability and accuracy under conditions of heavy-tailed noise. Additionally, the paper introduces the standard rounding procedure and the $\mathcal{S}$-outlier bound as relaxations of traditional assumptions. Numerical experiments provide support and validation for the presented results.

key words: *bilinear, subgradient, l_1 regression*

1. Introduction

The investigation into bilinear forms has been a topic of exploration across various studies, particularly due to the versatile applications of bilinear models. These applications span a wide spectrum, such as object recognition [1], compressed sensing [2], digital filter synthesis [3], prediction problems [4], channel equalization [5], and echo cancellation [6]. In [7], the authors synthesized the findings of those studies and introduced a novel method known as the iterative Wiener Filter. The iterative Wiener Filter, categorized as an l_2 regression method, demonstrates commendable performance in the identification of bilinear forms. In [8], this method can also be referred to as the Alternated Least Squares (ALS) algorithm. However, this performance is contingent upon a strict limitation–namely, that the signal system is assumed to be in a noiseless environment or subjected to white Gaussian noise. Given the ubiquity of noise in real-world scenarios and the limited information available about its nature, the applicability of the filter is constrained. In the realm of compressed sensing, as noted in [9], l_2 regression methods excel in signal retrieval when the system operates in a noiseless or Gaussian noise environment. However, when confronted with heavy-tailed noise, l_2 regression struggles to converge effectively.

To address system identification challenges under heavy-tailed noise conditions, we employ the l_1 regression method. The superiority of l_1 regression over l_2 regression is very intuitive in the presence of outlier observations, as l_1 regression is less affected by unusual observations due to its use of the absolute loss function. As far as we know, utilizing subgradient methods is the most practical approach to solve the l_1 regression problem. In [10], the authors discuss the Polyak subgradient method (which we will not consider in this paper since, under noiseless conditions, l_2 regression methods would be more effective) and the subgradient method with a geometrically decreasing step size. The convex version of the second algorithm can be traced back to Goffin [11]. Additionally, [12] analyzed these two methods for sharp weakly convex functions.

In this paper, we introduce the use of the subgradient method with a geometrically decaying stepsize, as introduced by Davis [12], as an effective l_1 algorithm for addressing the identification of bilinear forms under heavy-tailed noise conditions. To the best of our knowledge, the l_1 algorithm exhibits enhanced stability and attains greater accuracy when dealing with scenarios involving heavy-tailed noise. We have further demonstrated that a technique known as the standard rounding procedure [13] and an assumption, specifically the $\mathcal{S}$-outlier bound [14, Page 9], can be employed as a relaxation of conventional assumptions such as the Lipschitz bound and sharpness assumptions. Our numerical experiments have validated our results.

2. Identification of Bilinear Forms and ALS Algorithm

We consider the system with the bilinear forms given by:

$$y_i = \boldsymbol{\alpha}^T X_i \boldsymbol{\beta} + z_i, \quad i = 1, \ldots, p \tag{1}$$

in which $\boldsymbol{\alpha} \in \mathbb{R}^m$ is an unknown m-dimensional vector and $\boldsymbol{\beta} \in \mathbb{R}^n$ is also an unknown n-dimensional vector, $y_i, z_i \in \mathbb{R}$ are scalars, which denote the outcome of the system and the noise respectively. We assume $X_i = [(X_i)_1, (X_i)_2, \ldots, (X_i)_n]$ denotes an $m \times n$ matrix where $(X_i)_j$, $j = 1, \ldots, n$ are the m-dimensional column vectors of X_i. Throughout this paper, we will differentiate between scalars and vectors by denoting vectors as bold letters. For example, x represents a scalar, whereas $\boldsymbol{x}$ represents a vector.

The aim of this paper is to approximate the feasible solutions for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ within their respective feasible sets.

Manuscript received December 18, 2023.
Manuscript revised March 7, 2024.
Manuscript publicized May 6, 2024.

[†] School of Computer Science and Technology, Zhejiang Sci-Tech University, China.
a) E-mail: gwyang12@foxmail.com

DOI: 10.23919/transcom.2023EBP3210

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

Given that solving this problem is generally NP-hard and, therefore, computationally infeasible, our approach involves an approximate solution to the best subset problem.

Let $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ be the estimations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively, $\boldsymbol{y} = [y_1, y_2, \cdots, y_p]^T$ is the p-dimensional vector of outcomes, and the estimation of $\boldsymbol{y}$ is $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_p]^T$.

Then we have

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_p \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}^T X_1 \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}}^T X_2 \hat{\boldsymbol{\beta}} \\ \vdots \\ \hat{\boldsymbol{\alpha}}^T X_p \hat{\boldsymbol{\beta}} \end{bmatrix}.$$

For further information and methods related to the system of bilinear forms, we recommend that readers refer to [15]. To facilitate our analysis, it will be very helpful to use the following relationships [7]:

$$\hat{\boldsymbol{y}} = \mathcal{X}\left(\hat{\boldsymbol{\alpha}} \otimes \hat{\boldsymbol{\beta}}\right) = \mathcal{X}\left(\hat{\boldsymbol{\alpha}} \otimes I_n\right)\hat{\boldsymbol{\beta}} = \mathcal{X}\left(I_m \otimes \hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\alpha}}, \quad (2)$$

where

$$\mathcal{X} = \begin{bmatrix} \mathrm{Vec}\left(X_1\right)^T \\ \mathrm{Vec}\left(X_2\right)^T \\ \vdots \\ \mathrm{Vec}\left(X_p\right)^T \end{bmatrix},$$

and $\otimes$ denotes the Kronecker product, I_n and I_m are the identity matrices of sizes $n \times n$ and $m \times m$, respectively. We use operation $\mathrm{Vec}(X_i)$ to vectorize matrix X_i to a vector with mn entries, which means that stacking $(X_i)_j$ up. By employing well-established identities from the realm of linear algebra, these relationships can be readily derived.

Next we introduce the ALS algorithm as described in [8], and in [7] authors refer to this algorithm as the iterative Wiener filter. To avoid notation ambiguity between iterations and powers, we use $\boldsymbol{x}^{(k)}$ or $x^{(k)}$ to represent iterations (e.g. superscript enclosed in brackets), and x^k to represent x raised to the power of k. We use $\boldsymbol{x}^{(*)}$ to specifically denote that $\boldsymbol{x}$ belongs to the set of optimal solutions. We can define the function $G : (\mathbb{R}^m, \mathbb{R}^n) \longrightarrow \mathbb{R}$ as:

$$\begin{aligned} G\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) &:= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 \\ &= \left\|\boldsymbol{y} - \mathcal{X}\left(\hat{\boldsymbol{\alpha}} \otimes I_n\right)\hat{\boldsymbol{\beta}}\right\|_2^2 \\ &= \left\|\boldsymbol{y} - \mathcal{X}\left(I_m \otimes \hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\alpha}}\right\|_2^2, \end{aligned}$$

the second and third equations follow from the Eq. (2), and then we can minimize G by the following update equations:

$$\hat{\boldsymbol{\alpha}}^{(k+1)} \leftarrow \left(\left(I_m \otimes \hat{\boldsymbol{\beta}}^{(k)}\right)^T \mathcal{X}^T \mathcal{X}\left(I_m \otimes \hat{\boldsymbol{\beta}}^{(k)}\right)\right)^{-1} \cdot \left(I_m \otimes \hat{\boldsymbol{\beta}}^{(k)}\right)\boldsymbol{y},$$

$$\hat{\boldsymbol{\beta}}^{(k+1)} \leftarrow \left(\left(\hat{\boldsymbol{\alpha}}^{(k)} \otimes I_n\right)^T \mathcal{X}^T \mathcal{X}\left(\hat{\boldsymbol{\alpha}}^{(k)} \otimes I_n\right)\right)^{-1} \cdot \left(\hat{\boldsymbol{\alpha}}^{(k)} \otimes I_n\right)\boldsymbol{y}.$$

The update equations above reveal that we iterate $\hat{\boldsymbol{\alpha}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$ alternately. We will demonstrate that our subgradient method follows the same iterative procedure in next section.

Algorithm 1 Subgradient Method with A Geometrically Decaying Stepsize

```
Input  The measurement matrix $\mathcal{X}$, observations $\boldsymbol{y}$, iteration times $k$, step
       size coefficient $\lambda_{\boldsymbol{\alpha}}^{(1)}, \lambda_{\boldsymbol{\beta}}^{(1)}$, initialized identifier $\hat{\boldsymbol{\alpha}}^{(1)}, \hat{\boldsymbol{\beta}}^{(1)}$.
Applying the Standard Rounding Procedure
for $i = 1$ to $k$ do
    Choose $\boldsymbol{h}_{\hat{\boldsymbol{\alpha}}}^{(i)} \in \partial F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and $\boldsymbol{h}_{\hat{\boldsymbol{\beta}}}^{(i)} \in \partial F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ .
    Set $\hat{\boldsymbol{\alpha}}^{(i+1)} = \hat{\boldsymbol{\alpha}}^{(i)} - \frac{\lambda_{\boldsymbol{\alpha}}^{(i)}}{\left\|\boldsymbol{h}_{\hat{\boldsymbol{\alpha}}}^{(i)}\right\|_2}\boldsymbol{h}_{\hat{\boldsymbol{\alpha}}}^{(i)}$.
    Set $\hat{\boldsymbol{\beta}}^{(i+1)} = \hat{\boldsymbol{\beta}}^{(i)} - \frac{\lambda_{\boldsymbol{\beta}}^{(i)}}{\left\|\boldsymbol{h}_{\hat{\boldsymbol{\beta}}}^{(i)}\right\|_2}\boldsymbol{h}_{\hat{\boldsymbol{\beta}}}^{(i)}$.
```

3. Subgradient Method with a Geometrically Decaying Stepsize

We present our algorithm in detail as Algorithm 1. In the absence of the Lipschitz bound (3.2) and the μ-sharpness (3.1) assumptions, we employ the Standard Rounding Procedure in the first step. Subsequently, we select the subgradient of function F, with respect to parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. We delineated the function F in (3), and $\partial F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ represents the set of subgradients of F. The determination of the step size is guided by Eqs. (6) and (7), with the method for calculating their coefficients $\lambda_{\boldsymbol{\alpha}}^{(1)}$ and $\lambda_{\boldsymbol{\beta}}^{(1)}$ provided immediately afterward.

In the rest of this section, we show how to use the subgradient method to solve the problem of identification with bilinear forms, defining function $F : (\mathbb{R}^m, \mathbb{R}^n) \longrightarrow \mathbb{R}$ as:

$$F\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) := \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1 . \quad (3)$$

We can easily get the subgradient of $F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ using the following proposition:

Proposition 3.1. *As for the aforementioned function F with fixed $\boldsymbol{\alpha}$,*

$$-\left(\mathcal{X}\left(\boldsymbol{\alpha} \otimes I_n\right)\right)^T \cdot \mathrm{sign}\left(F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right)$$

is a subgradient of F with respect to $\boldsymbol{\beta}$, and for a fixed $\boldsymbol{\beta}$,

$$-\left(\mathcal{X}\left(I_m \otimes \boldsymbol{\beta}\right)\right)^T \cdot \mathrm{sign}\left(F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right)$$

is a subgradient of F with respect to $\boldsymbol{\alpha}$.

Where $\mathrm{sign}\left(F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right)$ denotes the sign of $F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)$, that is a vector with the same dimensions as $F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)$, but with a $+1$ entry when where $F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ has an entry greater than zero, a -1 entry when $F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ has an entry less than zero, and a zero entry where $F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ has an entry equal to zero.

Proof. To simplify our proof and remove ambiguity, for a fixed $\boldsymbol{\alpha}$, we will omit it in the expression $F(\boldsymbol{\alpha},\boldsymbol{\beta})$, and instead denoting it as $F(\boldsymbol{\beta})$.

Then for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in domain of $F(\boldsymbol{\alpha},\boldsymbol{\beta})$, we have

$$\begin{aligned} & F(\boldsymbol{\beta}_1) - F(\boldsymbol{\beta}_2) \\ = & \|Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1)\|_1 - \|Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_2)\|_1 \\ = & (Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1))^T \cdot \operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1)) - \\ & (Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_2))^T \cdot \operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_2)) \\ \leq & \left\{(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1))^T - (Y - \mathcal{X}(\hat{\boldsymbol{\alpha}}\otimes\boldsymbol{\beta}_2))^T\right\} \cdot \\ & \operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1)) \\ = & -(\mathcal{X}(\boldsymbol{\alpha}\otimes I_n)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2))^T \cdot \operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1)) \\ = & -\operatorname{sign}(F(\boldsymbol{\beta}_1))^T \cdot (\mathcal{X}(\boldsymbol{\alpha}\otimes I_n))(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \end{aligned}$$

Then we can use the same way to prove that the subgradient of $F(\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha}$. □

Proposition 3.2. *The Kronecker product is a continuous mapping.*

Proof. Let $\boldsymbol{a}$ be any m-dimensional vector and $\boldsymbol{b}$ be a fixed n-dimensional vector, for any $\epsilon > 0$, there exists a $\delta = \frac{\epsilon}{2\sqrt{n}\|\boldsymbol{b}\|_\infty}$, and let $\boldsymbol{c}$ be a m-dimensional vector which satisfies that

$$\left\{\boldsymbol{c}\,\middle|\,\|\boldsymbol{a}-\boldsymbol{c}\|_\infty \leq \frac{\epsilon}{2\sqrt{mn}\,\|\boldsymbol{b}\|_\infty}\right\}$$

We have

$$\begin{aligned} \operatorname{dist}(\boldsymbol{a},\boldsymbol{c}) = & \|\boldsymbol{a}-\boldsymbol{c}\|_2 \\ = & \sqrt{\sum_i^m (a_i - c_i)^2} \\ \leq & \|\boldsymbol{a}-\boldsymbol{c}\|_\infty \sqrt{m} \leq \delta \end{aligned}$$

and

$$\begin{aligned} \operatorname{dist}(\boldsymbol{a}\otimes\boldsymbol{b},\boldsymbol{c}\otimes\boldsymbol{b}) = & \sqrt{\sum_i^m\sum_j^n \left(a_ib_j - c_ib_j\right)^2} \\ \leq & \sqrt{nm\,(\|\boldsymbol{b}\|_\infty)^2\,\|\boldsymbol{a}-\boldsymbol{c}\|_\infty^2} \\ \leq & \frac{\epsilon}{2} \leq \epsilon. \end{aligned}$$

□

Let $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$ be defined as described in the preceding proof of proposition. Then, there exists a scalar $\theta \in \mathbb{R}$. By applying the definition of a convex function, we obtain:

$$(\theta\boldsymbol{a} + (1-\theta)\,\boldsymbol{c})\otimes\boldsymbol{b} \leq \theta(\boldsymbol{a}\otimes\boldsymbol{b}) + (1-\theta)(\boldsymbol{c}\otimes\boldsymbol{b}).$$

We observe that the Kronecker product constitutes a convex mapping, implying convexity in our objective function F. This assertion stems from the fact that the composition of a convex mapping (Kronecker product) with a convex function (l_1 norm) remains a convex function. The convergence guarantee for the subgradient method applied to convex functions can be found in [11]. Additionally, for weakly convex functions, the convergence assurance is established in [12].

Nevertheless, recent studies on the subgradient method often require assumptions about the objective function, such as the Lipschitz bound and sharpness assumptions.

Assumption 3.1. *(Restricted sharpness [14, Page 7]). A function $F(\cdot)$ is said to be μ-sharp with respect to $\boldsymbol{\xi}$ for some μ if*

$$F(\boldsymbol{\xi}) - F(\boldsymbol{\xi}^{(*)}) \geq \mu\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(*)}\right\|_1 \tag{4}$$

holds for any $\boldsymbol{\xi} \in \mathbb{R}^{mn}$.

Assumption 3.2. *We assume that the function is L-Lipschitz continues, i.e. the function $F(\boldsymbol{\xi})$ satisfies*

$$\|F(\xi_1) - F(\xi_2)\|_2 \leq L\,\|\xi_1 - \xi_2\|_2\,. \tag{5}$$

In [14], not only the two properties mentioned earlier but also the properties of approximate restricted sharpness and mixed-norm restricted isometry property (RIP) are required. The RIP is widely used not only in the compressed sensing field but also in many other fields, as evidenced by studies such as [2], [9], [16], [17]. Here we introduce the rounding procedure [13], using the ellipsoid method to the polytope $P = P(\boldsymbol{x}) := \{\boldsymbol{x}\,|\,\|\mathcal{X}\boldsymbol{x}\|_1 \leq 1\}$. For a given point $\boldsymbol{x} \notin P$ we using the hyperplane

$$\left\{\boldsymbol{y}\,\middle|(\boldsymbol{y}-\boldsymbol{x})^T\,\mathcal{X}^T\operatorname{sign}(\mathcal{X}\boldsymbol{x}) = ((\|\mathcal{X}\boldsymbol{x}\|_1) + 1)/2\right\}$$

to serve as a separation oracle, which separate $\boldsymbol{x}$ and P when $\|\mathcal{X}\boldsymbol{x}\|_1 > 1$.

In this procedure, we make use of the Gram-Schmidt method or an equivalent procedure, to orthogonalize the columns of $\mathcal{X}$ with respect to each other. Additionally, we normalize the columns of $\mathcal{X}$ such that they all have an l_1 norm of 1.

From [13, Theorem 2.1] we have that if $\|\boldsymbol{\xi}\|_2 \leq \sqrt{mn}$, then $\|\boldsymbol{\xi}\|_1 \leq 1$, and $\|\mathcal{X}\boldsymbol{\xi}\|_1 \leq 1$ follows from columns scaling. After applying the rounding procedure, we can observe that a new version of matrices $\mathcal{X}$ possesses an essential nature. This condition states that the matrix $\mathcal{X}$ with the property that for any $\boldsymbol{\xi}$,

$$\|\boldsymbol{\xi}\|_1 \geq \|\mathcal{X}\boldsymbol{\xi}\|_1 \geq \frac{1}{mn\sqrt{mn}}\,\|\boldsymbol{\xi}\|_1\,.$$

A matrix $\mathcal{X}$ with this property will be known as the l_1-conditioned. This property provides insight into the behavior of $\mathcal{X}$ with respect to the ℓ_1 norm of its input vector $\boldsymbol{\xi}$.

With this rounding procedure, we can proof that we do not need the Lipschitz bound and sharpness assumptions.

Theorem 3.1. *In a noisy case, an l_1-conditioned matrix $\mathcal{X}$ ensures that the function $F(\boldsymbol{\xi}) = \|\mathcal{X}\boldsymbol{\xi} - \boldsymbol{y}\|_1$ Lipschitz continuous with a constant of $L = 1$.*

Proof.

$$\begin{aligned}
&|F(\boldsymbol{\xi}_1) - F(\boldsymbol{\xi}_2)| \\
=\ & |\|\boldsymbol{y}_1 - \boldsymbol{y}\|_1 - \|\boldsymbol{y}_2 - \boldsymbol{y}\|_1| \\
=\ & |\|\mathcal{X}\boldsymbol{\xi}_1 - \boldsymbol{y} + z\|_1 - \|\mathcal{X}\boldsymbol{\xi}_2 - \boldsymbol{y} + z\|_1| \\
\leq\ & \|\mathcal{X}\boldsymbol{\xi}_1 - \mathcal{X}\boldsymbol{\xi}_2\|_1 \\
\leq\ & \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_1
\end{aligned}$$

where the second line equality follows from the presence of noise, specifically heavy-tailed noise, the third line inequality follows from the triangle inequality and the fourth inequality follows from the l_1-condition of $\mathcal{X}$. As a result, we have $L = 1$, and (5) follows. □

In prior literature, the RIP plays a crucial role in proving algorithm convergence. However, through the rounding procedure introduced here, we can directly obtain an l_1-conditioned measurement matrix $\mathcal{X}$. This property contributes to the establishment of the fourth inequality in Theorem 3.1.

Assumption 3.3. *($\mathcal{S}$-outlier bound [14, Page 9]) Matrix $X \in \mathbb{R}^{m\times n}$ is said to obey $\mathcal{S}$-outlier bound with respect to a set S with a constant δ if for all vectors $\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^n$, one has*

$$\delta \|\boldsymbol{\alpha} \otimes \boldsymbol{\beta}\|_1 \leq \left\|\boldsymbol{\alpha}^T X_{\mathcal{S}^c}\boldsymbol{\beta}\right\|_1 - \left\|\boldsymbol{\alpha}^T X_{\mathcal{S}}\boldsymbol{\beta}\right\|_1,$$

where $X_{\mathcal{S}^c}$ means $\{X_i\}_{i\in\mathcal{S}^c}$ and $X_{\mathcal{S}}$ means $\{X_i\}_{i\in\mathcal{S}}$.

Theorem 3.2. *In the presence of noise, if assumption 3.3 holds, a function $F(\boldsymbol{\xi}) = \|\mathcal{X}\boldsymbol{\xi} - \boldsymbol{y}\|_1$ is regarded as μ-sharp, with $\mu = c\delta$.*

Proof. If assumption 3.3 holds and noise is present, we have

$$\begin{aligned}
& F(\boldsymbol{\xi}) - F\left(\boldsymbol{\xi}^{(*)}\right) \\
=\ & \left\|\mathcal{X}\boldsymbol{\xi} - \mathcal{X}\boldsymbol{\xi}^{(*)} + z\right\|_1 - \|z\|_1 \\
=\ & \left\|\mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi} - \mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi}^{(*)}\right\|_1 + \\
& \sum_{i\in\mathcal{S}}\left(\left|\boldsymbol{\alpha}^T X_i\boldsymbol{\beta} - \left(\boldsymbol{\alpha}^{(*)}\right)^T X_i\boldsymbol{\beta}^{(*)} + z_i\right| - |z_i|\right) \\
\geq\ & \left\|\mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi} - \mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi}^{(*)}\right\|_1 - \\
& \sum_{i\in\mathcal{S}}\left(\left|\boldsymbol{\alpha}^T X_i\boldsymbol{\beta} - \left(\boldsymbol{\alpha}^{(*)}\right)^T X_i\boldsymbol{\beta}^{(*)}\right|\right) \\
\geq\ & c\delta\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(*)}\right\|_1,
\end{aligned}$$

where the first equality arises from the presence of noise, with the former part simply unfolding the expression of the function F. The latter part arises from the fact that the term $\boldsymbol{\xi}^{(*)}$ is what we subtract within function F, and subtracting two identical terms leaves only a z. The second equality follows from the definition of $\mathcal{S}$, the third inequality follows from the triangle inequality, the last inequality follows from the $\mathcal{S}$-outlier bound, and c is a constant. Therefore, we have $\mu = c\delta$. □

Subsequently, it becomes apparent that we can regard the standard rounding procedure and the assumption 3.3 as a form of relaxation for the Lipschitz bound and sharpness assumptions.

Following this, we can employ a strategy akin to the one presented in [11] to ascertain the algorithm's step size:

$$t_{\boldsymbol{\alpha}}^{(k)} = \frac{\lambda_{\boldsymbol{\alpha}}^{(k)}}{\left\|\boldsymbol{h}_{\boldsymbol{\alpha}}^{(k)}\right\|_2}, \quad t_{\boldsymbol{\beta}}^{(k)} = \frac{\lambda_{\boldsymbol{\beta}}^{(k)}}{\left\|\boldsymbol{h}_{\boldsymbol{\beta}}^{(k)}\right\|_2}, \tag{6}$$

where $\lambda_{\boldsymbol{\alpha}}^{(k)} = \lambda_{\boldsymbol{\alpha}}^{(0)}\rho_{\boldsymbol{\alpha}}^{q}$ and $\lambda_{\boldsymbol{\beta}}^{(k)} = \lambda_{\boldsymbol{\beta}}^{(0)}\rho_{\boldsymbol{\beta}}^{q}$. We initialize $\lambda_{\boldsymbol{\alpha}}^{(0)} = R\mu/(mp)$ and $\lambda_{\boldsymbol{\alpha}}^{(0)} = R\mu/(np)$. Let $\rho_{\boldsymbol{\alpha}}$ and $\rho_{\boldsymbol{\beta}}$ satisfy that

$$\rho_{\boldsymbol{\alpha}} = \begin{cases} \sqrt{1-(\mu/m)^2} & \mu/m \leq \sqrt{2}/2 \\ \mu/(2m) & \mu/m \geq \sqrt{2}/2 \end{cases}, \tag{7}$$

$$\rho_{\boldsymbol{\beta}} = \begin{cases} \sqrt{1-(\mu/n)^2} & \mu/n \leq \sqrt{2}/2 \\ \mu/(2n) & \mu/n \geq \sqrt{2}/2 \end{cases}.$$

Here, R is a constant, and we assume that the iteration algorithm started in close proximity to the feasible solution set. This implies $\left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(*)}\right\|_2 + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(*)}\right\|_2 \leq R$.

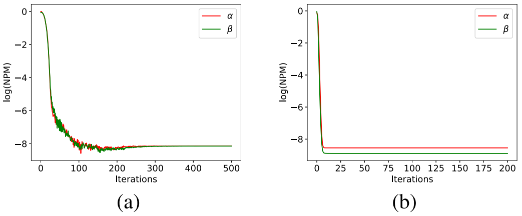
4. Numerical Experiment

This section presents the experimental settings and numerical results of our study. To evaluate the accuracy of the measurements, we use the normalized projection misalignment metric [7, Page 654].

$$\mathrm{NPM}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) = 1 - \left(\frac{\boldsymbol{\alpha}^T\hat{\boldsymbol{\alpha}}}{\|\boldsymbol{\alpha}\|_2\|\hat{\boldsymbol{\alpha}}\|_2}\right)^2$$

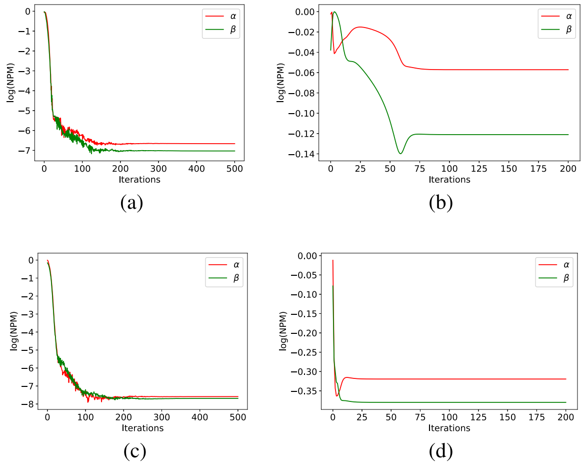
$$\mathrm{NPM}\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}\right) = 1 - \left(\frac{\boldsymbol{\beta}^T\hat{\boldsymbol{\beta}}}{\|\boldsymbol{\beta}\|_2\left\|\hat{\boldsymbol{\beta}}\right\|_2}\right)^2.$$

We generated the entries of vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using the Bernoulli distribution with probability 1/2, while the matrix X_i was generated by independent identically distributed (i.i.d.) $N(0,1)$ random variables. We choose values for $\rho_{\boldsymbol{\alpha}}$ and $\rho_{\boldsymbol{\beta}}$ from the interval $[0.9, 1)$, and set the vector lengths to $m = 30$ and $n = 30$. Let us consider that there are $p = 200$ data samples available for estimating the vectors. In Fig. 1, it is evident that the l_2 regression model exhibits faster convergence than the l_1 regression model within the Gaussian noise structure. In Fig. 2, we examine a system exposed to Cauchy noise and another system subjected to heteroscedastic noise. The heteroscedastic noise structure is characterized by the following distribution: one-third of the entries conform to a Gaussian distribution, another third adhere to a Cauchy distribution, and the remaining entries follow a t-distribution. It is observable that for a system under Cauchy noise or heteroscedastic noise, the l_2 regression method struggles to converge, while the l_1 regression method



(a) (b)

Fig. 1 On the left (a), we have a system under Gaussian noise employing the l_1 regression model with the subgradient method and a geometrically decaying stepsize. On the right (b), we have a system under Gaussian noise using the l_2 regression model with the ALS algorithm.



(a) (b)

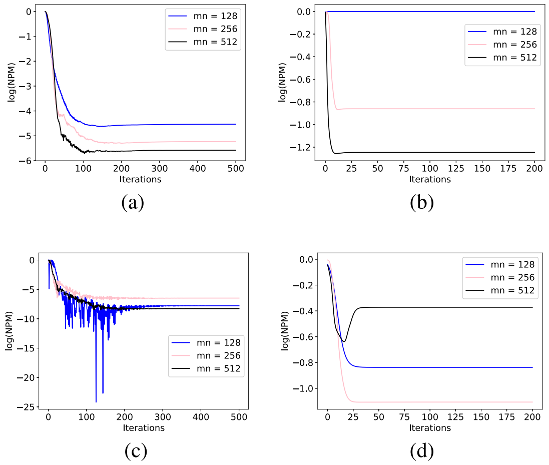
(c) (d)

Fig. 2 Figures (a) and (c) employ the l_1 regression model with the subgradient method and a geometrically decaying stepsize, while figures (b) and (d) utilize the ALS algorithm. Figures (a) and (b) illustrate the system under Cauchy noise, whereas figures (c) and (d) explore the impact of heteroscedastic noise.

continues to perform well.

Next, the algorithm's performance is evaluated from a system identification perspective. We generated the entries of $\boldsymbol{\alpha}$ according to the ITU-T G.168 Recommendation [18], and $\boldsymbol{\beta}$ are generated as $\beta_i = 2^{-(i-1)}$, with $i = 1, 2, \ldots, n$. In this simulation, as depicted in Fig. 3, the length of $\boldsymbol{\beta}$ varies with values of $m = 2, 4,$ and 8; consequently, the length of $\boldsymbol{\alpha}$ is fixed at $n = 64$. From a system identification standpoint, it is evident that under a heavy-tailed noise condition, the l_1 regression consistently outperforms the l_2 regression method. The variation observed in Figure (a) within Fig. 3 is attributed to the relatively short length of the vector $\boldsymbol{\beta}$. For a fixed value of p (e.g., the available data samples), it is evident that increasing the product of mn can contribute to achieving more accurate results.

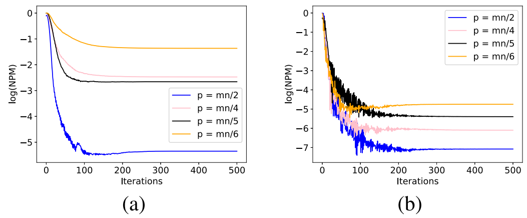
Finally, we will explore the impact of relatively small available data samples, denoted as p, on the algorithm. Notably, not only does the l_1 regression method converge when $p < mn$, but it also performs well when $p < mn/4$. Contrastingly, the l_2 regression method faces challenges in attaining satisfactory results under such conditions, primarily due to the influence of heavy-tailed noise and the limited availability of data samples. See Fig. 4.



(a) (b)

(c) (d)

Fig. 3 In this experiment, we demonstrate the convergence of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In Figure (a) and figure (c), we apply the l_1 regression model using the subgradient method with a geometrically decaying step size. In Figure (b) and Figure (d), the l_2 regression model is employed. Figures (a) and (b) depict the NPM of $\boldsymbol{\alpha}$, while Figures (c) and (d) showcase the NPM of $\boldsymbol{\beta}$. We experiment with various combinations of m and n while maintaining a fixed $p = 200$ under a Cauchy noise condition.



(a) (b)

Fig. 4 Figure (a) illustrates the Normalized Power Mean (NPM) of $\boldsymbol{\alpha}$, while Figure (b) presents the NPM of $\boldsymbol{\beta}$. We conduct tests using the subgradient method with a geometrically decaying step size under a Cauchy noise scenario, varying the parameter p (representing available data samples). The vector lengths of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to $m = 8$ and $n = 64$, respectively.

5. Conclusion

To conclude, the utilization of l_1 regression methods presents the advantageous capability of generating robust solutions, a trait highly beneficial in diverse applications like phase retrieval and compressed sensing. This subgradient method exhibits relatively good performance when dealing with bilinear systems under heavy-tailed noise.

However, the selection between l_1 and l_2 regression methods hinges upon the distinct problem and noise characteristics at hand. In certain instances, the preference may lean towards l_2 regression, particularly when dealing with Gaussian noise and well-conditioned problems. In a broader perspective, the integration of subgradient methods for nonlinear problem-solving has demonstrated promising outcomes, holding significant potential for driving substantial advancements across various application domains.

Acknowledgments

I would like to extend my heartfelt gratitude and sincere appreciation to Dr. Shen for their invaluable guidance, unwavering support, and insightful mentorship throughout my academic journey. I would also like to express my deep gratitude to Zhejiang Sci-Tech University for providing me with an exceptional learning environment and resources that have been crucial to my intellectual and personal growth.

References

[1] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, and S. Ma, "A boosting, sparsity-constrained bilinear model for object recognition," IEEE MultiMedia, vol.19, no.2, pp.58–68, 2012.

[2] P. Walk and P. Jung, "Compressed sensing on the image of bilinear maps," 2012 IEEE International Symposium on Information Theory Proceedings, pp.1291–1295, 2012.

[3] U. Forssen, "Adaptive bilinear digital filters," IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process., vol.40, no.11, pp.729–735, 1993.

[4] J. Lee and V.J. Mathews, "Adaptive bilinear predictors," Proc. ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing, vol.iii, pp.III/489–III/492, 1994.

[5] G. Ma, J. Lee, and V.J. Mathews, "A RLS bilinear filter for channel equalization," Proc. ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing, vol.iii, pp.III/257–III/260, 1994.

[6] R. Hu and H. Ahmed, "Echo cancellation in high speed data transmission systems using adaptive layered bilinear filters," IEEE Trans. Commun., vol.42, no.234, pp.655–663, 1994.

[7] J. Benesty, C. Paleologu, and S. Ciochină, "On the identification of bilinear forms with the wiener filter," IEEE Signal Process. Lett., vol.24, no.5, pp.653–657, 2017.

[8] G. Chen, M. Gan, S. Wang, and C.L.P. Chen, "Insights into algorithms for separable nonlinear least squares problems," IEEE Trans. Image Process., vol.30, pp.1207–1218, 2021.

[9] S. Li, D. Liu, and Y. Shen, "Adaptive iterative hard thresholding for least absolute deviation problems with sparsity constraints," J. Fourier Anal. Appl., vol.29, pp.1207–1218, 2022.

[10] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy, "Composite optimization for robust rank one bilinear sensing," Information and Inference: A Journal of the IMA, vol.10, no.2, pp.333–396, Oct. 2020.

[11] J.L. Goffin, "On convergence rates of subgradient optimization methods," Mathematical Programming, vol.13, pp.329–347, 1977.

[12] D. Davis, D. Drusvyatskiy, K.J. MacPhee, and C. Paquette, "Subgradient methods for sharp weakly convex functions," J. Optim. Theory Appl., vol.179, pp.962–982, 2018.

[13] K.L. Clarkson, "Subgradient and sampling algorithms for l_1 regression," Proc. Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '05, USA, pp.257–266, 2005.

[14] T. Tong, C. Ma, and Y. Chi, "Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number," IEEE Trans. Signal Process., vol.69, pp.2396–2409, 2021.

[15] D. Yang, "Solution theory for systems of bilinear equations," Ph.D. thesis, The College of William and Mary, April 2011.

[16] B. Recht, M. Fazel, and P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," SIAM Rev., vol.52, no.3, pp.471–501, 2010.

[17] Y. Chen, Y. Chi, and A.J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," IEEE Trans. Inf. Theory, vol.61, no.7, pp.4034–4059, 2015.

[18] I.T. Union, "Digital network echo cancellers," ITU-T Recommendation G.168, ITU-T, 2002.

Guowei Yang earned his B.S. degree in Software Engineering from Xi'an University of Posts & Telecommunications, Xi'an, China, in 2020. Currently, he is pursuing a master's degree in Computer Technology at Zhejiang Sci-Tech University. His current research interests encompass optimization algorithms, compressed sensing, and phase retrieval.

PAPER

Cascaded Deep Neural Network for Off-Grid Direction-of-Arrival Estimation*

Huafei WANG†a), Xianpeng WANG†b), Xiang LAN†c), *and* Ting SU†d), *Nonmembers*

SUMMARY Using deep learning (DL) to achieve direction-of-arrival (DOA) estimation is an open and meaningful exploration. Existing DL-based methods achieve DOA estimation by spectrum regression or multi-label classification task. While, both of them face the problem of off-grid errors. In this paper, we proposed a cascaded deep neural network (DNN) framework named as off-grid network (OGNet) to provide accurate DOA estimation in the case of off-grid. The OGNet is composed of an autoencoder consisted by fully connected (FC) layers and a deep convolutional neural network (CNN) with 2-dimensional convolutional layers. In the proposed OGNet, the off-grid error is modeled into labels to achieve off-grid DOA estimation based on its sparsity. As compared to the state-of-the-art grid-based methods, the OGNet shows advantages in terms of precision and resolution. The effectiveness and superiority of the OGNet are demonstrated by extensive simulation experiments in different experimental conditions.

key words: *off-grid, direction-of-arrival estimation, deep learning, autoencoder, convolutional neural networks*

1. Introduction

Direction-of-arrival (DOA) estimation has been a hot topic of research for decades, since it plays a crucial role in the field of wireless communication and target sensing [1], [2]. In order to realize accurate DOA estimation, plenty of methods have been proposed over the past few decades. The most well-known methods are the subspace-based methods, which include the multiple signal classification (MUSIC) [3], estimation of signal parameters via rotational invariance techniques (ESPRIT) [4] and their variants [5]–[9]. The basic principle of MUSIC method [3] is that a spatial spectrum is first constructed based on the orthogonal relationship between the signal and noise subspaces, and then the peak search is performed over the spatial spectrum with a specific step size to achieve DOA estimation. While, the ESPRIT method [4], [9] capitalizes the rotational invariance of signal subspace and does not require spectrum search. However, the performance of these subspace-based methods depend on the accuracy of covariance, which depends on the number of snapshots and signal-to-noise ratio (SNR), therefore their performance may suffer significantly degradation at insufficient number of snapshots or low SNRs.

In recent decade, compressed sensing (CS) technique has attracted much attentions [10], [11], which has been successfully applied in DOA estimation [12], [13]. The CS-based methods mainly capture the sparsity of sources in spatial domain, and adopt different strategies of sparse minimization to achieve DOA estimation. Since the sparsity of source signals comes from the discrete grid in spatial domain, the CS-based methods can be classified into three main categories: on-grid, off-grid, and gridless methods [14]. On-grid methods [15], [16] can accurately estimate the angles coincided with the fixed grid points in spatial domain. However, their performance suffers from the off-grid error when angles mismatch with the grid points, especially under a coarse grid condition. Apparently, the problem of off-grid errors faced by the on-grid methods can be alleviated by recursive grid refinement or increasing the degree of spatial discretization. However, it may result in a significant increase in computational complexity. Off-grid methods [17], [18] and gridless methods [19] achieve a balance between estimation accuracy and computational complexity, which can realize high precision DOA estimation under coarse grid conditions with low computational cost. However, since all above CS-based methods are model-driven, they rely on the pre-established mathematical model, and share a common shortcoming that each estimation needs a complete optimization process that need appropriate parameter initialization. The inappropriate initial parameters may cause the performance degradation and even method failure.

Most recently, DOA estimation using deep learning (DL) technique [20], [21] has raised much attentions, which is completely data-driven method. The DL-based methods use the powerful nonlinear mapping capability of neural networks (NN) to learn features in the array data to achieve DOA estimation. As compare to the CS-based approaches, DL-based methods can achieve DOA estimation based on simple multiplications and additions in trained networks with no optimization process, and the training is to be done off-line once and for all. As a current research hotspot, a stream of DL-based methods have emerged for DOA estimation and its application [22]–[24]. Specifically, in [25] and

Manuscript received January 9, 2024.
Manuscript revised March 24, 2024.
Manuscript publicized May 14, 2024.

†School of Information and Communication Engineering and the State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University, 570228 China.

*This work is supported in part by the Hainan Province Science and Technology Special Fund under Grant ZDYF2023GXJS168; in part by the National University of Defense Technology Independent Innovation Fund under Grant 23-ZZCX-JDZ-45; in part by the Important Science and Technology of Haikou City under Grant 2021-004; and in part by the National Natural Science Foundation of China under Grant 62101088.

a) E-mail: wong9525@163.com
b) E-mail: wxpeng1986@126.com (Corresponding author)
c) E-mail: xlan@hainanu.edu.cn
d) E-mail: suting4190@hainanu.edu.cn

DOI: 10.23919/transcom.2024EBP3006

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

Table 1 Glossary of notations throughout the paper.

Notations	Definitions
Italic (e.g., x or X):	Scalars
Lowercase bold italic (e.g., $\boldsymbol{x}$):	Vectors
Capital bold italic (e.g., $\boldsymbol{X}$):	Matrices
$\boldsymbol{I}_i$:	$(i \times i)$-dimensional unit matrix
$\boldsymbol{0}_i$:	$(i \times i)$-dimensional zero matrix
$\boldsymbol{X}^{-1}$:	Inverse of $\boldsymbol{X}$
$\boldsymbol{X}^T$:	Transpose of $\boldsymbol{X}$
$\boldsymbol{X}^H$:	Hermitian transpose of $\boldsymbol{X}$
$\|\boldsymbol{X}\|_F$:	Frobenius-norm of $\boldsymbol{X}$
$\|\boldsymbol{x}\|_2$:	2-norm of vector $\boldsymbol{x}$
$\|x\|$:	Modulus of scalar x
$\mathbb{E}\{\cdot\}$:	Calculate expectation
$\mathbb{R}^{i\times i}$:	Set of $(i \times i)$-dimensional real value matrix
$\mathbb{C}^{i\times i}$:	Set of $(i \times i)$-dimensional complex value matrix
C_N^n:	Generate all possible combinations of n numbers from N numbers
$utv\{\boldsymbol{X}\}$:	Vectorize the upper triangular part of $\boldsymbol{X}$ by column
$mat\{\boldsymbol{x}\}$:	The reverse operator of $utv\{\boldsymbol{X}\}$

[26], the authors applied the denoising autoencoder to denoise the array covariance of uniform linear array and sparse linear array, respectively. Then implement DOA estimation based on the denoised covariance using MUSIC-based methods, e.g., root MUSIC and spatial smooth MUSIC. Similarly, A. Barthelme et al. used neural network to reconstruct the covariance matrix from sample covariance matrix, then achieve DOA estimation by applying the MUSIC estimator to the reconstructed covariance matrix [27]. On the other hand, a deep convolutional neural network (CNN) was presented in [28] to reconstruct the noiseless covariance matrix by using its Toeplitz structure, then root MUSIC method is applied to realized gridless DOA estimation. However, the methods in [25]–[28] are essentially semi-DL methods with the techniques for final DOA estimation still model-driven, which results in their inability to achieve end-to-end DOA estimation. For purely DL-based methods, the authors in [29], [30] proposed the deep neural networks (DNN) for robust DOA estimation in non-ideal situations such as array imperfections and color noise. However, multiple parallel networks are adopted both in [29] and [30], which leads to large network structures that require large amounts of accurately labeled data for training, and such volume of labeled data for non-ideal situations are very difficult to collect in practice. On the other hand, a deep CNN was developed in [31] with utilizing the sparsity prior. However, since 1-dimensional (1D) convolution is utilized, the method do not exhibit significant performance improvements. Further, G. K. Papageorgiou et al. designed a deep network in [32] for DOA estimation in low SNR, where the 2-dimensional (2D) convolutions are used and the DOA estimation is modeled as a multi-label classification task by inputting 3-channel covariance. Nevertheless, such a network suffers from the similar problem that CS-based methods suffer, i.e., the off-grid errors. Coarse labeling of covariance leads to that the network can only accurately classify the angles been labeled, the off-labeled DOAs (similar to the off-grid DOAs) can only be classified as the nearest angle to it, which cause off-grid errors. While, the dense labeling will undoubtedly require a large amount of labeled data, the collection of such amounts of labeled data is a challenge in practical applications.

Above all, most of the existing DL-based DOA estimation methods either do not provide end-to-end DOA estimation or the estimation precision is restricted by off-grid errors. Thus, this paper try to fill in this gap by proposing a cascaded DNN. The proposed neural network is referred to as off-grid network (OGNet) which is composed of an autoencoder (AE) and a deep CNN (DCNN). The AE behaves like a filter, which takes the upper triangular part of the sampling unitary covariance as input to reduce the divergences between sampling and theoretical unitary covariance. Afterward, the reconstructed unitary covariance by AE is used as the input of the DCNN to predict the off-grid error vector hence to realize DOA estimation based on its sparsity. The major contributions of this paper are summarized as:

- A neural network architecture is proposed for the DOA estimation in off-grid scenarios. The proposed architecture include an AE and a deep CNN.
- In the proposed neural network, the AE behaves like a pre-processor to reduce divergences between the sampling and theoretical unitary covariance. And the deep CNN is designed for off-grid DOA estimation by modeling the off-grid error into labels, which enables it can achieve off-grid DOA estimation based on its sparsity and without a priori information of on-grid angles.
- The proposed neural network architecture achieves more accurate off-grid DOA estimation as compare to the state-of-the-art grid-based methods include traditional methods and DL-based methods.

The remaining part of this paper is structured as follows: A briefly description of the problem formulation of DOA estimation is given in Sect. 2. In Sect. 3, the architecture of the proposed cascaded DNN for off-grid DOA estimation is presented. The network training strategies corresponding to AE and DCNN are introduced in Sect. 4. Simulation experiment results are shown in Sect. 5 to evaluate the effectiveness and superiority of the proposed network. Finally, the conclusions

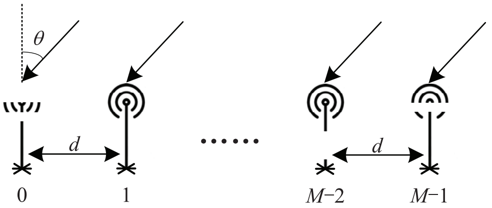


Fig. 1 Uniform linear array with M antennas and inter-antenna distance $d = \lambda/2$.

of this paper are given in Sect. 6. The glossary of notations throughout this paper is given in Table 1 for convince.

2. Problem Formulation

For DOA estimation, there are several array geometries to be applied, such as the linear, circular and planar. In this work, a uniform linear array (ULA) is considered. As shown in Fig. 1, suppose a ULA equipped with M antennas is configured at a inter-antenna distance of $d = \ /2$, where is the wavelength of signals. With P independent far-field narrow-band signals impinging on the ULA from different directions of $\ _p (p = 1; 2; : : : ; P)$, the data received by ULA at l-th sampling snapshot is [3]

$$y(l) = \boldsymbol{A}\boldsymbol{s}(l) + \boldsymbol{n}(l); \ l = 1; 2; \cdots ; L; \tag{1}$$

where $\boldsymbol{s}(l) = [s_l(l); \cdots ; s_P(l)]^T \in C^{P\times 1}$ denotes the signal vector and $\boldsymbol{n}(l)$ denotes the additive Gaussian white noise vector at l-th sampling snapshot. $\boldsymbol{A} = [\boldsymbol{a}(\ _1); \boldsymbol{a}(\ _2); \cdots ; \boldsymbol{a}(\ _P)]$ denotes the $M \times P$ array steering matrix with

$$\boldsymbol{a}(\ _p) = [1; e^{j(2\pi d/\lambda)\sin\theta_p}; \dots ; e^{j(2\pi d/\lambda)(M-1)\sin\theta_p}]^T; \tag{2}$$

where j is the imaginary unit. By collecting L snapshots, the multi-snapshot data is expressed as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{N}; \tag{3}$$

with $\boldsymbol{Y} = [y(1); \cdots ; y(L)]$, $\boldsymbol{S} = [\boldsymbol{s}(1); \cdots ; \boldsymbol{s}(L)]$ and $\boldsymbol{N} = [\boldsymbol{n}(1); \cdots ; \boldsymbol{n}(L)]$.

Based on Eq. (1), with collecting infinite snapshots, the theoretical covariance of the receiving data can be expressed as

$$\boldsymbol{R} = E\{y(l)y(l)^H\} = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \boldsymbol{R}_n; \tag{4}$$

where $\boldsymbol{R}_s$ and $\boldsymbol{R}_n$ denote the covariance of incident signals and noise, respectively. However, the theoretical covariance in Eq. (4) is hard to obtain and unknown in practice, hence it is usually replaced by the sampling covariance, which is

$$\bar{\boldsymbol{R}} = \frac{1}{L}\sum_{l=1} y(l)y(l)^H : \tag{5}$$

Since the theoretical covariance $\boldsymbol{R}$ is a centro-Hermitian matrix, it can be transformed into a real-valued matrix [33], [34], which is called theoretical unitary covariance (TUC), by [8]

$$\boldsymbol{R}_u = (\boldsymbol{U}_M)^H \boldsymbol{R}\boldsymbol{U}_M \tag{6}$$

with

$$\begin{cases} \boldsymbol{U}_{M=even} = \frac{\sqrt{2}}{2}\begin{bmatrix} \boldsymbol{I}_{\frac{M}{2}} & j\boldsymbol{I}_{\frac{M}{2}} \\ \ _{\frac{M}{2}} & -j\ _{\frac{M}{2}} \end{bmatrix}; \\ \boldsymbol{U}_{M=odd} = \frac{\sqrt{2}}{2}\begin{bmatrix} \boldsymbol{I}_{\frac{M-1}{2}} & \mathbf{0}_{\frac{M-1}{2}\times 1} & j\boldsymbol{I}_{\frac{M-1}{2}} \\ \mathbf{0}_{1\times\frac{M-1}{2}} & \sqrt{2} & \mathbf{0}_{1\times\frac{M-1}{2}} \\ \ _{\frac{M-1}{2}} & \mathbf{0}_{\frac{M-1}{2}\times 1} & -j\ _{\frac{M-1}{2}} \end{bmatrix}; \end{cases} \tag{7}$$

where $\ _i$ denotes $(i \times i)$-dimensional matrix with the anti-diagonal elements being 1 others all 0. While, although the sampling covariance $\bar{\boldsymbol{R}}$ in Eq. (5) is a Hermitian matrix, it's not centro-Hermitian, hence it cannot directly be transformed into real-valued by Eq. (6). Fortunately, the forward-backward technique can be applied first to turn $\bar{\boldsymbol{R}}$ into a centro-Hermitian matrix, which is

$$\bar{\boldsymbol{R}}_{fb} = \frac{1}{2}(\bar{\boldsymbol{R}} + \ _M\bar{\boldsymbol{R}}^* \ _M): \tag{8}$$

Then, the sampling unitary covariance (SUC) based on $\bar{\boldsymbol{R}}$ is

$$\bar{\boldsymbol{R}}_u = (\boldsymbol{U}_M)^H \bar{\boldsymbol{R}}_{fb}\boldsymbol{U}_M : \tag{9}$$

Based on Eq. (9), we are interested in estimating unknown DOAs from $\bar{\boldsymbol{R}}_u$. Hence, as considering the powerful nonlinear mapping capability of NN, a cascaded DNN is designed in the following section to predict the off-grid error vector by using $\bar{\boldsymbol{R}}_u$ as input, then achieve off-grid DOA estimation using the sparsity of the predicted off-grid error vector.

3. Proposed Cascaded DNN for Off-Grid DOA Estimation

The overall architecture of the proposed cascaded DNN, named as OGNet, is shown in Fig. 2, which is composed of two components. The first component is a neural network called AE, which is consisted by fully connected (FC) layers. And the second component is the DCNN mainly composed of 2-dimensional (2D) convolutional (Conv.) layers and FC layers. The former is to reduce the divergences between $\bar{\boldsymbol{R}}_u$ and $\boldsymbol{R}_u$, while the latter is to predict the off-grid error vector by using the unitary covariance predicted by AE as input. Finally, the DOA estimation is realized based on the sparsity of the predicted off-grid error vector. The detailed architecture of the components within OGNet is introduced as following.

3.1 The Architecture of AE

From Sect. 2, it is known that the $\boldsymbol{R}_u$ is obtained based on

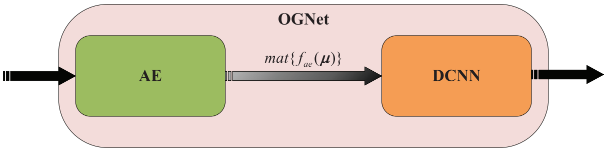


Fig. 2 The overall architecture of the proposed cascaded deep neural network, i.e., the OGNet.

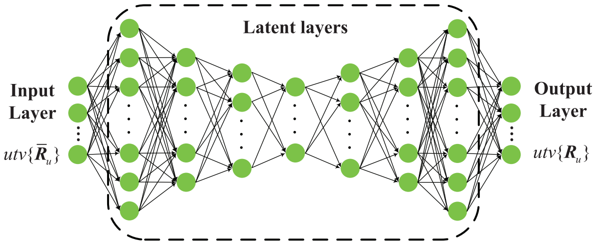


Fig. 3 The architecture of autoencoder within OGNet.

infinite snapshots, which is practically unrealistic, while the $\bar{\boldsymbol{R}}_\text{u}$ is calculated with L snapshots. Therefore, there must exist certain difference between $\bar{\boldsymbol{R}}_\text{u}$ and $\boldsymbol{R}_\text{u}$, i.e.

$$\boldsymbol{R}_\text{u} = \bar{\boldsymbol{R}}_\text{u} + \quad \boldsymbol{R}_\text{u}; \tag{10}$$

where $\quad \boldsymbol{R}_\text{u}$ represents the divergence matrix between $\bar{\boldsymbol{R}}_\text{u}$ and $\boldsymbol{R}_\text{u}$. The AE within the proposed OGNet is designed to reduce $\quad \boldsymbol{R}_\text{u}$. The architecture of AE within OGNet is displayed in Fig. 3, which is consisted of 9 FC layers include 1 input layer, 1 output layer and 7 latent layers (i.e., FC layers). Each latent layer is followed by a ReLU activation layer except for the input and output layers to avoid the gradient disappearing. The specific configurations of each layer in AE is given in Table 2.

Note that $\bar{\boldsymbol{R}}_\text{u}$ and $\boldsymbol{R}_\text{u}$ are all M × M-dimensional Hermitian matrices, therefore the training data pair for AE is respectively denoted as vector consisted of the elements of upper triangular parts of $\bar{\boldsymbol{R}}_\text{u}$ and $\boldsymbol{R}_\text{u}$ by columns, i.e.,

$$\begin{aligned} &= \text{utv}\{\bar{\boldsymbol{R}}_\text{u}\}; \\ \boldsymbol{u} &= \text{utv}\{\boldsymbol{R}_\text{u}\}: \end{aligned} \tag{11}$$

where $\quad \in \text{R}^{(\frac{M(M-1)}{2}+M)\times 1}$ represents the input of AE, and $\boldsymbol{u} \in \text{R}^{(\frac{M(M-1)}{2}+M)\times 1}$ represents the output label of AE. Then, the nonlinear mapping procedure of AE can be parameterized as

$$\text{f}_\text{ae}(\quad) = \text{f}_\text{aout}(\text{f}_\text{a7}(:::(\text{f}_\text{a1}(\text{f}_\text{ain}(\quad))))) = !\ ; \tag{12}$$

where $!\quad \in \text{R}^{(\frac{M(M-1)}{2}+M)\times 1}$ is the predicted output of AE during training; $\text{f}_\text{ae}\{\cdot\}$ represents the nonlinear mapping function of the whole AE, $\text{f}_\text{ain}\{\cdot\}$ and $\text{f}_\text{aout}\{\cdot\}$ respectively denote the mapping function of input layer and output layer, and $\text{f}_\text{ai}\{\cdot\}$ with $\text{i} = 1; 2; \cdots; 7$ denotes the mapping function of i-th latent layer.

Since the AE is modeled for a regression task and completely composed of FC layers, it is potential to over-fitting in the case of small training data. In order to prevent overfitting, the mean-square-error (MSE) with L_2 regularization is

Table 2 Specific configurations of autoencoder.

Layers	Number of neurons	Activation function
Input	$M(M-1)/2+M$	-
FC #1	300	ReLU
FC #2	200	ReLU
FC #3	100	ReLU
FC #4	50	ReLU
FC #5	100	ReLU
FC #6	200	ReLU
FC #7	300	ReLU
Output	$M(M-1)/2+M$	Linear

chosen as the loss function of AE to optimize the trainable weights and biases set $\quad_\text{a}$ during training phase, that is

$$^{\star}_\text{a} = \arg\min_{\text{a}} \frac{1}{\text{D}_\text{a}} \tilde{\text{Õ}}^{\text{D}_\text{a}}_{\text{d}=1} \mathcal{L}(!\ ^{(\text{d})}; \boldsymbol{u}^{(\text{d})}) + \frac{}{2} \tilde{\text{Õ}}^{\text{N}_\text{a}}_{\text{n}=1} \|\mathcal{W}_\text{a}^{(\text{n})}\|_\text{F}^2 \;; \tag{13}$$

where $\mathcal{L}(!\ ^{(\text{d})}; \boldsymbol{u}^{(\text{d})}) = \{\frac{1}{\text{Q}}\ {}''\ ^{\text{Q}}_{\text{i}=1} |!\ ^{(\text{d})}_\text{i} - \boldsymbol{u}_\text{i}^{(\text{d})}|^2\}$ is the MSE loss with $\text{Q} = \frac{\text{M}(\text{M}-1)}{2} + \text{M}$; D_a represents the total number of training data for AE; $!\ ^{(\text{d})}$ and $\boldsymbol{u}^{(\text{d})}$ respectively denote the predicted output and the output label of AE when inputting d-th data; $!\ ^{(\text{d})}_\text{i}$ and $\boldsymbol{u}_\text{i}^{(\text{d})}$ represent the i-th entries of $!\ ^{(\text{d})}$ and $\boldsymbol{u}^{(\text{d})}$, respectively; $\quad_\text{a} = \{\mathcal{W}_\text{a}; \boldsymbol{b}_\text{a}\}$ is the set of trainable weights and biases in AE; is the regularization parameter for the weights in AE, which is $= 10^{-4}$ in this paper; N_a is the total number of hidden layers in AE and $\mathcal{W}_\text{a}^{(\text{n})}$ represents the weights of n-th hidden layer of AE.

3.2 The Architecture of DCNN

After obtaining a prediction of $\boldsymbol{u}$ from AE, the predicted unitary covariance can be obtained based on its Hermitian property, i.e.,

$$\hat{\boldsymbol{R}}_\text{u} = \text{mat}\{!\ \}; \tag{14}$$

which is taken as the input of DCNN to predict the sparse off-grid error vector. During training phase, $\boldsymbol{R}_\text{u}$ is chosen as the training input. The architecture of the DCNN is displayed in Fig. 4. The DCNN contains 1 input layer, 1 output layer, 10 2D Conv. layers, 1 flatten layer and 4 FC layers. Each Conv. layer has 64 channels and is followed by a batch normalization (BN) layer [35]. The kernel size of all Conv. layers is $\ \times\ $ with $\ = 3$ and stride $\ = 1$ and same padding. Similarly, to prevent the gradient disappearing, the activation function used in each Conv. and FC layer of DCNN is ReLU. The specific configurations of each layer of DCNN is given in Table 3.

The output label of DCNN is designed as an G-dimensional P-sparse vector , where G depends on the number of discrete grid points in spatial domain. For instance, if the spatial domain from $-\ $ to $\ $ is discretized by the grid interval of , then $\text{G} = 2\ /\ + 1$, and the spatial discrete grid can be obtained as $\ = \{\ _1;\ _2; \cdots;\ _\text{G}\}$ with $\ = \ _{g+1} - \ _g$ $(\text{g} = 1; 2; \cdots; \text{G} - 1)$. Suppose that the true

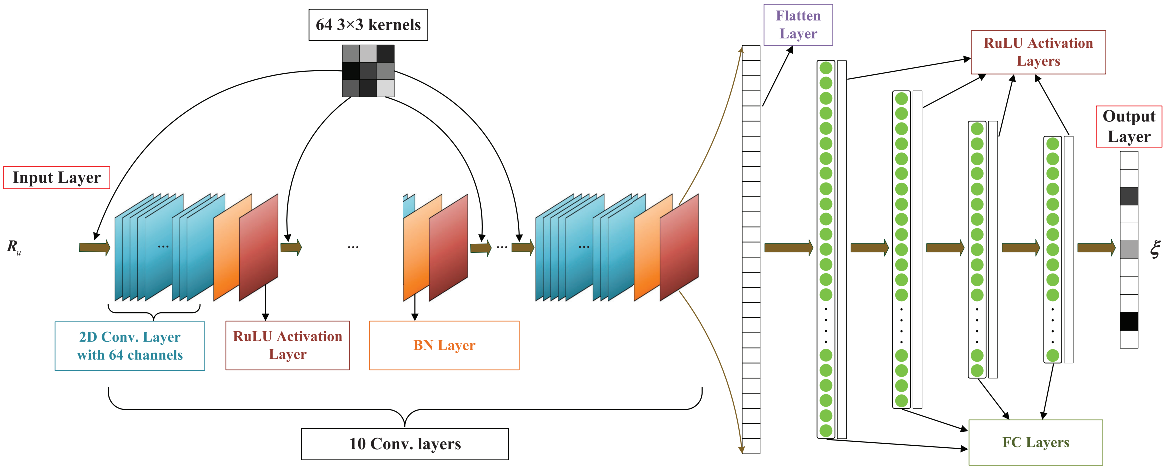


Fig. 4 The architecture of DCNN within OGNet.

Table 3 Specific configurations of DCNN.

Layers	Size/neurons	Activation function
Input	$M \times M \times 1$	-
Conv. #1 with BN	$64 \times M \times M \times 1$	ReLU
Conv. #2 with BN	$64 \times M \times M \times 1$	ReLU
⋮	⋮	⋮
Conv. #9 with BN	$64 \times M \times M \times 1$	ReLU
Conv. #10 with BN	$64 \times M \times M \times 1$	ReLU
Flatten	9216	-
FC #1	2048	ReLU
FC #2	1024	ReLU
FC #3	512	ReLU
FC #4	256	ReLU
Output	G	Linear

DOAs of P sources are $= \{ {}_1; {}_2; \cdots ; {}_P\}$, then the sparse off-grid error vector is expressed as

$$\boldsymbol{e} = [e_1; e_2; \cdots ; e_G]^T \in R^{G\times1}; \qquad (15)$$

where $e_g \subset (- /2; /2]$ with $g = 1; 2; \cdots ; G$, and the entries of $\boldsymbol{e}$ are all 0 except for the g_p-th entry being $e_{g_p} = {}_p - {}_{g_p}$ with $p = 1; 2; \cdots ; P$ and $g_p = 1; 2; \cdots ; G$. ${}_{g_p}$ denotes the angle in nearest to ${}_p$. Note that e_{g_p} could be negative or positive and could be very small when the targets are very close to the discrete grids. In order to enhance the sparsity of $\boldsymbol{e}$ and convert it into a positive vector for better learning by neural networks, a linear transformation is introduced as

$${}_{g_p} = e_{g_p} \times c + \frac{}{2} \times c; \qquad (16)$$

where $g_p = 1; 2; \cdots ; G$, c is a constant which is set to $c = 10$ in this paper. Then the output label of DCNN is expressed as

$$= [{}_1; {}_2; \cdots ; {}_G]^T \in R^{G\times1}; \qquad (17)$$

which shares the same sparsity with $\boldsymbol{e}$.

Remark 1. *We set $e_g \subset (- /2; /2]$ intentionally, because if $- /2$ and $/2$ both are included, there will exist conflict in the network labels, leading to problems during training. Let us take a simple example: suppose the true DOA is −59° and = 2°, then we can labeled the outputs as that $e_1 = /2 = 1°$ or $e_2 = - /2 = -1°$, which are actually pointing at the same DOA. Hence, if $- /2$ and $/2$ are both included, the labels are conflict for training. On the other hand, we set $e_g \subset (- /2; /2]$ instead of $e_g \subset [- /2; /2)$. This is because a linear transformation is made in Eq. (16) to maintain the sparsity of the labels, if we set $e_g \subset [- /2; /2)$, the sparsity will be vanished when $e_g = - /2$, which will also lead to problems during training.*

Similarly, the nonlinear mapping procedure of DCNN can be parameterized as

$$f_{cnn}(\boldsymbol{R}_u) = f_{cout}(f_{c14}(f_{c13}(:::(f_{c2}(f_{c1}(f_{cin}(\boldsymbol{R}_u))))))) = ; \qquad (18)$$

where $\in R^{G\times1}$ represents the predicted output of DCNN during training; $f_{cnn}\{\cdot\}$ denotes the nonlinear mapping function of the entire DCNN, $f_{cin}\{\cdot\}$ and $f_{cout}\{\cdot\}$ respectively stand for the mapping function of input layer and output layer of DCNN; $f_{ci}\{\cdot\}$ with $i = 1; 2; :::; 10$ is the mapping function of i-th Conv. layer, and $f_{ci}\{\cdot\}$ with $i = 11; 12; :::; 14$ is the mapping function of $(i - 10)$-th FC layer.

Likewise, the DCNN is modeled to complete a regression task. Therefore, MSE is chosen as the loss function of the DCNN for the optimization of the trainable weights and biases set ${}_c$ in DCNN, i.e.,

$${}^{\star}_{c} = \arg\min_{c} \left(\frac{1}{D_c} \sum_{d=1}^{D_c} \frac{1}{G} \sum_{i=1}^{G} |{}^{(d)}_i - {}^{(d)}_i|^2 \right); \qquad (19)$$

where D_c represents the total number of training data for DCNN; ${}^{(d)}$ and ${}^{(d)}$ respectively denote the predicted output and the training label of DCNN by input d-th data during training; ${}^{(d)}_i$ and ${}^{(d)}_i$ represent the i-th entry of ${}^{(d)}$ and ${}^{(d)}$, respectively. It should be noted that there are other candidate loss functions for regression task, such as mean-absolute-error (MAE) and smooth MAE. As compare to MAE, the gradient of MSE is dynamic (as the error decreases, so does the gradient), which can accelerate the convergence of the

Table 4 The procedure of off-grid DOA estimation using OGNet.

```
Algorithm 1: Off-grid DOA estimation using OGNet.
Train AE with proper training data and strategies (off-line);
Train DCNN with proper training data and strategies (off-line);
ULA receives data $\boldsymbol{Y}$;
Calculate the sampling covariance $\bar{\boldsymbol{R}}$;
Transform $\bar{\boldsymbol{R}}$ to unitary covariance $\bar{\boldsymbol{R}}_u$ by Eq. (8) and (9);
Get the upper triangular parts of $\bar{\boldsymbol{R}}_u$ by $\boldsymbol{\mu} = utv\{\bar{\boldsymbol{R}}_u\}$;
Input $\boldsymbol{\mu}$ into the trained AE to predict $\boldsymbol{\omega}$;
Obtain $\hat{\boldsymbol{R}}_u$ by Eq. (14);
Input $\hat{\boldsymbol{R}}_u$ into the trained DCNN to predict $\boldsymbol{\zeta}$;
Obtain $P$ spikes in $\boldsymbol{\zeta}$ and corresponding indices by peak searching;
Realize high-precision off-grid DOA estimation according to Eq. (20).
```

function and make the network training faster. Hence, the MSE rather than MAE is chosen as the loss function of DCNN, because the training of CNN values training speed, especially the training of deep CNN.

3.3 Off-Grid DOA Estimation

The training of the networks within OGNet are performed off-line with proper strategies to obtain the trained OGNet. Once the training is completed, the off-grid errors corresponding to sources can be predicted by feed a sampling covariance into the trained OGNet, while the exact DOAs are not estimated yet. The off-grid DOA estimation is realized by a post processing based on the sparsity of , and without a priori on-grid angles estimation. Since is P-sparse, and the index of its each value corresponds to that of on-grid angle on the spatial grid $= \{ \ _1; \ _2; \cdots; \ _G\}$, where there are spikes there're sources impinging from those angles. On the other hand, according to Eq. (16), each value of the spikes in contains the off-grid error information of sources. Hence, the on-grid angles and off-grid errors of sources can be obtained simultaneously by performing peak searching on to find P spikes. Then, the off-grid DOA estimation can be realized by

$$\bar{\ } = \ _{\iota} + \ \frac{\ _{\iota}}{\mathsf{c}} - \frac{\ }{2} \ ; \qquad (20)$$

where $\in 1; 2; \cdots; \mathsf{G}$ denotes the indices corresponding to the P spikes in , $_\iota$ and $_\iota$ denote the -th entry of and , respectively. The procedure of off-grid DOA estimation using OGNet is summarized as in Table 4[†], where the off-line training data and strategies for OGNet are introduced in the following section.

4. Network Training

The training is performed on a Windows PC equipped with 3.2 GHz AMD Ryzen 2700 CPU, 12 GB NVIDIA GeForce RTX 3060 GPU and 48 GB RAM. The generation of training data are performed by MATLAB, and the DL framework for constructing NN, training NN and simulations are based on TensorFlow 2.6.0 plus Python 3.7.12. The ADAM [36] with initial learning rate 0:001 is chosen as the optimizer for all networks within OGNet. In both training and testing phases, a ULA equipped with $\mathsf{M} = 12$ antennas is considered. The data generation and training strategies for the AE and DCNN is described in detail as following.

4.1 Training of AE

To generate the training data of AE, we consider P sources lie in spatial from −60° to 60°. The true DOAs of sources are randomly sampled from the interval [−60°; 60°] with a sampling step of 0:1°, and the angular separation between any two DOAs is greater than 2°. To generate sufficient data for training, 2×10^6 pairs of true DOA are randomly sampled from the interval. At each sample, let the SNR vary randomly in steps of 5 dB between −15 dB to 20 dB and fix the number of snapshots at $\mathsf{T} = 50$. Then, the training data for AE is generated according to Eqs. (4), (5), (6), (9) and (11).

During training phase, the training data of AE is randomly divided into 90% for training and 10% for validation, hence $\mathsf{D_a} = 2 \times 10^6 \times 0{:}9 = 1{:}8 \times 10^6$. The training for AE is carried out 100 epochs with a batch size of 1000. The learning rate drop factor is set as 0:5 with the drop period being 4 epochs.

4.2 Training of DCNN

In generating the training data for DCNN, is considered to be 60°, and is considered to be = 2° for instance. Then, the spatial discrete grid is = {−60°; −58°; −56°; ⋯ ; 0°; ⋯ ; 58°; 60°} and G = 61. Since = 2°, the off-grid error $\mathsf{e}_{g_\mathsf{p}} \subset (-1; 1]$. The constant c is set as c = 10 in this paper, then $\ _{g_\mathsf{p}} \subset (0; 20]$. The number of sources varies from 1 to $\mathsf{P_{max}}$ to enable the DCNN can achieve multi-DOA estimation, where $\mathsf{P_{max}} = 3$ is considered in this paper to relief the memory and system demands. To generate the off-grid DOAs, the on-grid angles of P = 1; 2; 3 sources are firstly generated from all possible combinations in , i.e., we can obtain $C_{61}^1 + C_{61}^2 + C_{61}^3$ pairs of on-grid angles. Then, to release training burden, the off-grid error for each pair of on-grid angle is randomly chosen from " = {−0:9; −0:8; ⋯ ; −0:1; 0; 0:1; ⋯ ; 0:9; 1} without overlap until all off-grid errors in " have been traversed or the rest candidate off-grid errors are not enough to be assigned to the on-grid angles in the angle pair. Then, the total number of pairs of off-grid DOAs for training is $\lfloor 20/1 \rfloor \times C_{61}^1 + \lfloor 20/2 \rfloor \times C_{61}^2 + \lfloor 20/3 \rfloor \times C_{61}^3 = 235460$ with $\lfloor \cdot \rfloor$ denoting the round-down operator. After obtaining the 235460 pairs off-grid DOAs, the corresponding $\boldsymbol{R}_\mathsf{u}$ (the training input for DCNN) with respect to each pair of off-grid DOAs at each SNR in {−15; −10; −5; 0; 5; 10; 15; 20} dB is calculated by (4) and (6), which leads to the total number of training data being 235460 × 8 = 1883680. The output label corresponding to each $\boldsymbol{R}_\mathsf{u}$ is generated during generating the off-grid error according to (15), (16) and (17). For example, if the on-grid angle pair is {−60°; −56°} and the

[†] The number of sources P is assumed to be known previously.

corresponding off-grid error is $\{-0.6°; 0.3°\}$, the output label becomes $= [-0.6\times10+10; 0; 0.3\times10+10; 0; 0; \cdots; 0]^\mathrm{T}$ where the indices of non-zeros elements in is the same as the indices of the position of the corresponding angle pair in .

When training the DCNN, the training data is randomly divided into 90% training set and 10% validation set, hence $D_\mathrm{c} = 1883680 \times 0.9 = 1695312$. The training for DCNN is carried out 200 epochs with a batch size of 512, and the corresponding learning rate drop factor is set as 0.7 with the drop period being 5 epochs.

Remark 2. *It should be noted that the OGNet without AE (i.e., only DCNN) is also capable of achieving off-grid DOA estimation by taking the SUC as the input, even if the DCNN is trained using TUC. However, the estimation performance of DCNN is inferior compared to that of OGNet, as will be demonstrated in the simulation experiments later. On the other hand, the DCNN can be trained under different grid intervals by using the similar data generation and training strategies, and such DCNN still has superior DOA estimation performance, which will also be demonstrated in the simulation experiments later.*

Remark 3. *The computational burden of OGNet is mainly dominated by the floating-point operations (FLOPs) in AE and DCNN when estimating DOA, while the FLOPs of AE and DCNN depend on the number of layers and neurons in each layer. Since AE is consisted by FC layers, its FLOPs is* $2\ {}_{i=1}^{N_\mathrm{a}} I^{(i)}O^{(i)}$, *where* N_a *denotes the number of FC layers in AE, and* $I^{(i)}$ *and* $O^{(i)}$ *represent the number of input and output neurons of each FC layer, respectively. While in DCNN, the FLOPs include two parts: FLOPs of Cov. layers and that of FC layers. The FLOPs of FC layers in DCNN can be similarly calculated as* $2\ {}_{i=1}^{N_\mathrm{d}} I^{(i)}O^{(i)}$ *with* N_d *being the number of FC layers in DCNN. The FLOPs of each Cov. layer in DCNN is* $2C_\mathrm{I}\ {}^2C_\mathrm{O}M^2$ *since all Cov. layers are identical in DCNN, then the total FLOPs of all Cov. layers is* $2C_\mathrm{I}\ {}^2C_\mathrm{O}M^2N_{\mathrm{co}v}$ *where* C_I *and* C_O *represent the input and output channels of each convolution,* k *is the kernel size, and* $N_{\mathrm{co}v}$ *denotes the total number of Cov. layers. Thus the total FLOPs of OGNet is*

$$FLOPs = 2\sum_{i=1}^{N_\mathrm{a}} I^{(i)}O^{(i)} + 2\sum_{i=1}^{N_\mathrm{d}} I^{(i)}O^{(i)} + 2C_\mathrm{I}\ {}^2C_\mathrm{O}M^2N_{\mathrm{co}v}. \tag{21}$$

Based on the similar calculation process, the FLOPs of CNN in [32] can be computed as $2\ {}_{i=1}^{N_\mathrm{d}} I^{(i)}O^{(i)} + 2C_\mathrm{I}\ {}^2C_\mathrm{O}M^2N_{\mathrm{co}v}$. Obviously, according to the specific network parameters of CNN provided in [32], the computational complexity of the OGNet is a little more expensive than that of the CNN. While, since the DOA estimation achieved by NN is only based on simple additions and multiplications in trained networks, it is still within acceptable limits, which will be demonstrated as in the following section.

5. Simulation Experiments and Analyses

Numerous simulation experiments are conducted in this section to evaluate the effectiveness and superiority of the proposed OGNet. During simulations, the ULA is considered to be equipped with $M = 12$ antennas. The random independent far-field narrow-band signal is utilized for simulations. Firstly, the effectiveness of the proposed OGNet under different scenarios is evaluated. Then, the performance superiority of the OGNet is evaluated by comparing with the state-of-art NN-based methods and traditional model-driven methods.

5.1 Effectiveness of the OGNet

Firstly, the effectiveness of the AE within OGNet is evaluated by the difference between the theoretical $\boldsymbol{R}_\mathrm{u}$ and the predicted $\hat{\boldsymbol{R}}_\mathrm{u}$ by AE, which is referred as to predicted difference. The difference between $\boldsymbol{R}_\mathrm{u}$ and the $\bar{\boldsymbol{R}}_\mathrm{u}$, which is called original difference, is calculated for comparison. The number of sources is $P = 2$ with $= [-27.21°; 13.35°]$. The difference is evaluated by

$$\| \ \boldsymbol{R}_\mathrm{u}\|_\mathrm{F} = \frac{1}{N_\mathrm{mc}} \sum_{i=1}^{N_\mathrm{mc}} \|\boldsymbol{R}_\mathrm{u} - \dot{\boldsymbol{R}}_\mathrm{u}^{(i)}\|_\mathrm{F}; \tag{22}$$

where $N_\mathrm{mc} = 10^3$ denotes the total number of Monte Carlo trials, $\dot{\boldsymbol{R}}_\mathrm{u}^{(i)}$ represents $\hat{\boldsymbol{R}}_\mathrm{u}$ or $\bar{\boldsymbol{R}}_\mathrm{u}$ at i-th Monte Carlo trial. The results of $\| \ \boldsymbol{R}_\mathrm{u}\|_\mathrm{F}$ under different SNRs and number of snapshots are given in Fig. 5 with the number of snapshots being $T = 500$ and SNR $= 0$ dB, respectively. As clearly shown in Fig. 5, the difference decreases with SNR increasing. Moreover, the predicted difference is significantly smaller than the original difference in all SNR cases. Similarly, one can also see that the difference decreases with increasing number of snapshots and the predicted difference is much smaller than

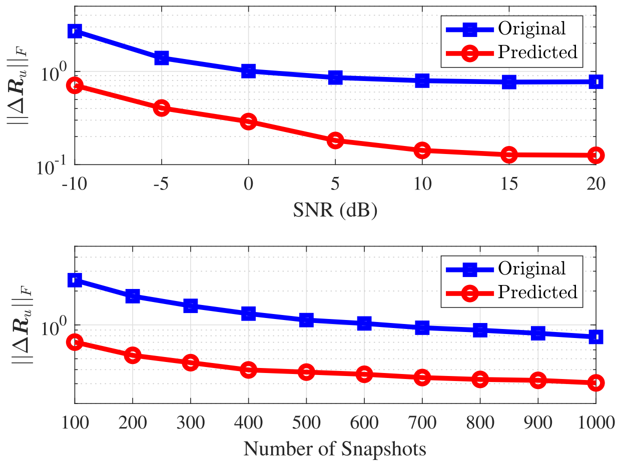


Fig. 5 $\| \ \boldsymbol{R}_\mathrm{u}\|_\mathrm{F}$ under different SNRs and number of snapshots with $P = 2$. Upper: divergence versus SNR. Lower: divergence versus number of snapshots.

the original difference in all snapshot cases. The results in Fig. 5 illustrate that the AE component within OGNet can effectively reduce the difference between the sampled $\bar{\boldsymbol{R}}_u$ and theoretical $\boldsymbol{R}_u$.

As previously claimed in Sect. 4.2, the OGNet without AE (i.e., only DCNN) can also achieve off-grid DOA estimation. Hence, the simulation experiments on DOA estimation RMSE of DCNN and OGNet is conducted to evaluate the effectiveness of the AE and DCNN with $P = 2$ and $= [-27:21°; 13:35°]$. The definition of RMSE is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{mc}P}\sum_{i=1}^{N_{mc}} | \; - \; \bar{}^{(i)}|_2^2}; \qquad (23)$$

where $N_{mc} = 10^3$ denotes the total number of Monte Carlo trials, $'^{(i)}$ denotes the estimated DOAs at i-th Monte Carlo trial. The corresponding results are given in Fig. 6, where the upper is the RMSE versus SNR with $T = 500$, and the lower is the RMSE versus number of snapshots with SNR = 0 dB. As can be seen from the Fig. 6 that the RMSEs of DCNN and OGNet are reasonably decreased with the increasing of SNR and snapshots. While, the RMSE of the OGNet is distinctly much smaller than that of the DCNN, especially at low SNR, which demonstrate that the DCNN can achieve off-grid DOA estimation alone and the AE can effectively improve its performance to achieve high-precision off-grid DOA estimation.

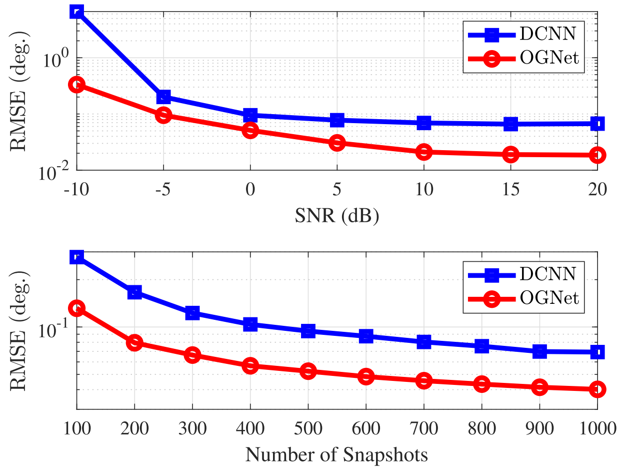


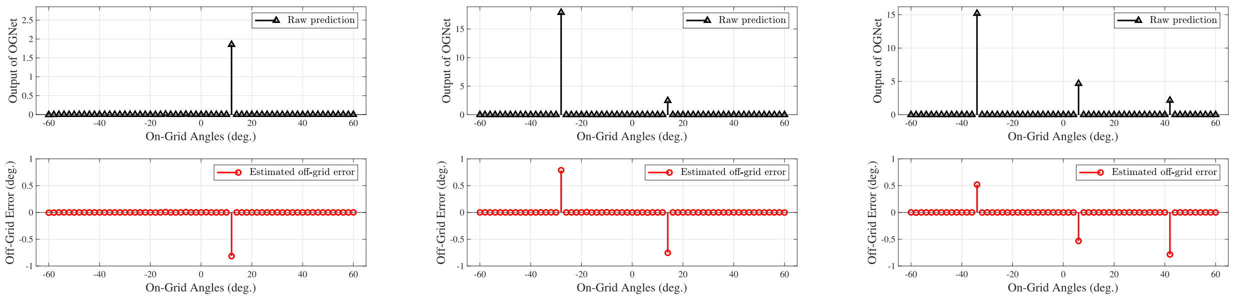
Fig. 6 RMSE of DCNN and OGNet under different SNRs and number of snapshots. Upper: RMSE versus SNR. Lower: RMSE versus number of snapshots.

Further, the effectiveness of the OGNet is evaluated by single-prediction simulation experiments with different number of sources P. When $P = 1$, the true DOA is fixed as $[11:23°]$. When $P = 2$ and $P = 3$, the true DOAs are fixed as $[-27:21°; 13:35°]$ and $[-33:56°; 5:42°; 41:37°]$, respectively. The simulations are conducted under SNR = 0 dB and $T = 500$. The prediction results of OGNet and the corresponding estimated off-grid errors with different number of sources P are shown in Fig. 7, and the corresponding specific numerical results of estimated off-grid errors and DOAs are given in Table 5. As can be seen in Fig. 7, in the case of different P, the raw predictions of OGNet and corresponding estimated off-grid error vector have obvious spikes and are very sparse. On the other hand, the specific numerical results in Table 5 show that the off-grid errors estimated by OGNet are very accurate for different P, thus the corresponding DOA estimation are precise. The mean estimation errors for $P = 1$, $P = 2$ and $P = 3$ reached 0:0448, 0:0531 and 0:0937, respectively, which indicates that the proposed OGNet can effectively achieve high-precision off-grid DOA estimation.

Lastly, different off-grid DOAs are estimated by using the proposed OGNet in different number of sources $P = 1; 2; 3$ to evaluate the effectiveness more widely. For different P, the initial DOA of the first source is fixed as $_1 = -59:52°$, then other initial DOAs are $_2 = \;_1 +$ and $_3 = \;_2 +$ with $= 5:3°$, respectively. Then, DOAs under different P are varies from the initial angles with an increasing step of 1° until all possible angles in the interval of $(-60°; 60°)$ are sampled. The estimation results with SNR = 0 dB and $T = 500$ are shown in Fig. 8. It is clearly demonstrated that the estimated DOAs by OGNet under different number of sources are very close to the true DOAs, which illustrate that the proposed OGNet is valid for different number of sources and different angles.

5.2 Superiority of the OGNet

The effectiveness of the proposed OGNet has been eval-

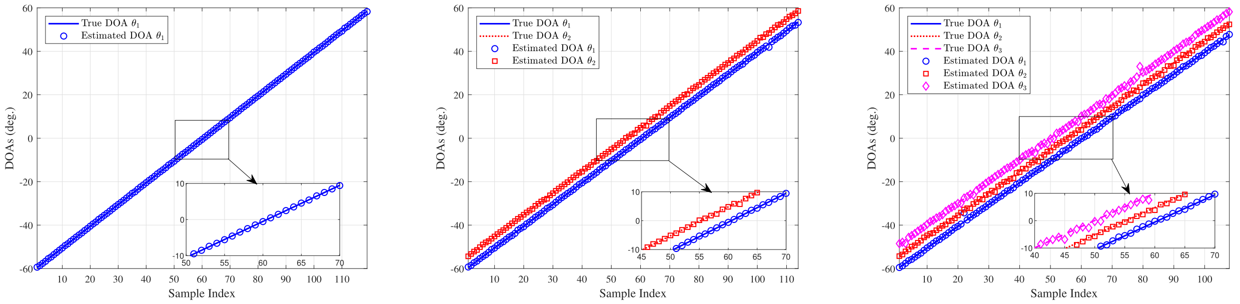


(a) $P = 1$. (b) $P = 2$. (c) $P = 3$.

Fig. 7 Single-prediction results of OGNet and corresponding estimated off-grid errors with different number of sources P.

Table 5 Specific numerical results of estimated off-grid errors and DOAs by OGNet.

Sources	True off-grid errors	Estimated off-grid errors	True DOAs	Estimated DOAs
$P = 1$	[−0.77°]	[−0.8148°]	[11.23°]	[11.1852°]
$P = 2$	[0.79°, −0.65°]	[0.7874°, −0.7536°]	[−27.21°, 13.35°]	[−27.2126°, 13.2464°]
$P = 3$	[0.44°, −0.58°, −0.63°]	[0.5292°, −0.5328°, −0.7848°]	[−33.56°, 5.42°, 41.37°]	[−33.4808°, 5.4672°, 41.2152°]



(a) $P = 1$.

(b) $P = 2$.

(c) $P = 3$.

Fig. 8 Estimation results on different off-grid DOAs with different number of sources P by OGNet.

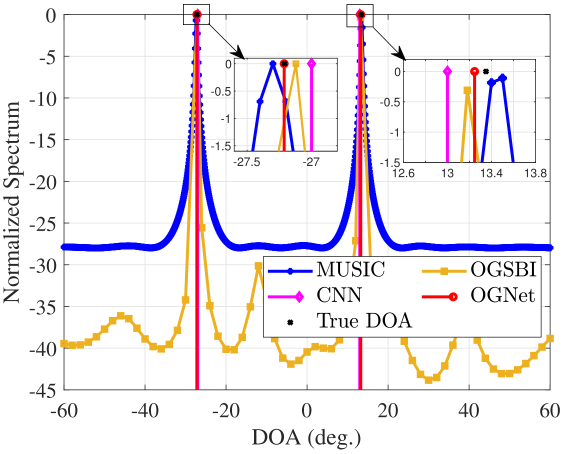


Fig. 9 Normalized spectra of different methods.

uated in the previous subsection. In this subsection, the performance superiority of the proposed OGNet is evaluated by comparing with the sate-of-the-art methods under P = 2. Unless otherwise specified, the corresponding true DOAs of sources are fixed as = [−27:21°; 13:35°]. The methods introduced for comparison include MUSIC [3], off-grid sparse Bayesian inference (OGSBI) [17] and CNN [32]. Additionally, the conditional Cramér-Rao bound (CRB) [37] is calculated for comparison. The evaluation metric is the RMSE defined in Eq. (23).

First of all, the normalized spectra of single DOA estimation for different methods are given in Fig. 9 with SNR = 0 dB and T = 500. Note that the CNN and the proposed OGNet do not really have a spectrum, hence the DOA estimation results of CNN and OGNet are shown as solid lines placed directly in the figure for intuitive comparison. As clearly shown in Fig. 9, the DOAs estimated by the proposed OGNet are closer to the true DOAs than that of the CNN, also than the spectrum peaks of the other comparison methods, which indicate that the proposed OGNet has higher precision of DOA estimation.

Afterward, the computational complexity of different methods are compared by their average time required for a single DOA estimation, the corresponding result is given in Table 6. The results are based on 10^3 independent simulations, T = 500, SNR = 0 dB and = [−27:21°; 13:35°]. As can be seen from Table 6, since the computational complexity of the proposed OGNet is higher than that of the other comparison methods, its average time for a single DOA estimation is reasonably longer than that of its rivals. Despite this, the computational complexity of the proposed OGNet is still within acceptable limits and can fulfill the requirements of real-time estimation.

Then, the simulation experiments for RMSE and probability of successful detection (PSD) of DOA estimation of different methods in terms of SNR are carried out to evaluate the superiority of the OGNet. The criteria for successful detection is $\sqrt{\frac{1}{P}|\quad - \quad '|_2^2} <$ with $= 0{:}3$ in this paper. The results are depicted in Figs. 10 and 11 with T = 500. In Fig. 10, one can find that the RMSE of CNN does not decreased anymore when SNR > 0 dB since it's an on-grid method that have a precision restriction for off-grid angles under a coarse searching grid. Conversely, the RMSE of OGNet continues to decrease as the SNR increases since it can well handle the off-grid error. Meanwhile, the proposed OGNet possesses the lowest RMSE among the other methods. But what cannot be ignored is that when SNR ≥ 10 dB, the RMSE of our proposed method seems to deviate from

Table 6 The average time required for a single DOA estimation based on different methods.

Methods	Average time (sec.)
MUSIC	0.0127
OGSBI	0.0439
CNN	0.0518
OGNet	0.0595

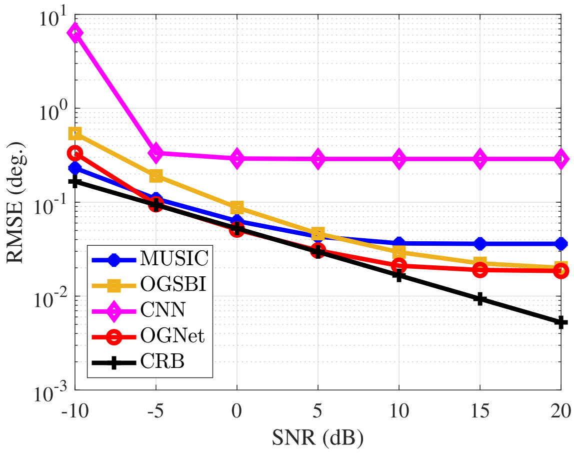


Fig. 10 RMSE of different methods under different SNRs.

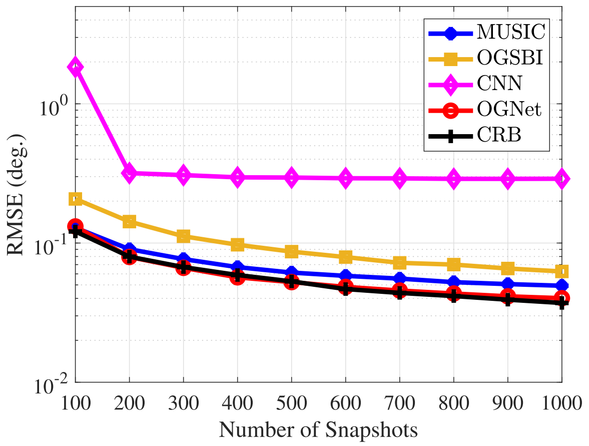


Fig. 12 RMSE of different methods under different number of snapshots.

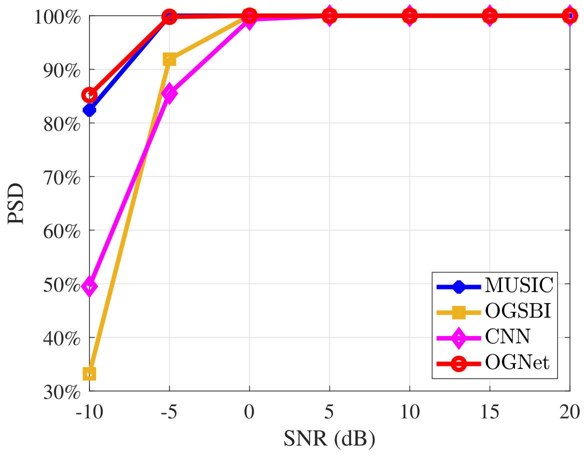


Fig. 11 PSD of different methods under different SNRs.

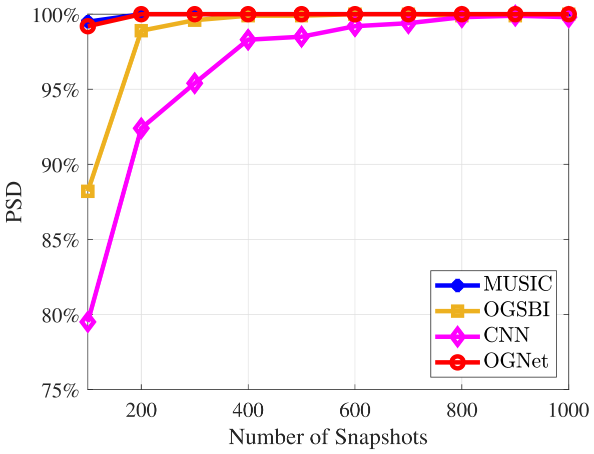


Fig. 13 PSD of different methods under different number of snapshots.

CRB and has a floor effect[†]. On the other hand, it can be easily found in Fig. 11 that the PSD of the proposed OGNet is higher than that of all the comparison methods at low SNRs, and reaches 100% faster than the other methods. These results indicate that the proposed OGNet can achieve high-precision off-grid DOA estimation and has distinct superiority under different SNRs.

Next, the simulation experiments for RMSE and PSD of DOA estimation of different methods in terms of number of snapshots are conducted to evaluate the superiority of the OGNet. The results are given in Figs. 12 and 13 with SNR = 0 dB. From the Fig. 12, we can find that the RMSE of the proposed OGNet is the lowest and is closest to the CRB in all cases of number of snapshots. While, in Fig. 13, the PSD of proposed method is similar to that of MUSIC and higher than that of other methods at small snapshots, and reaches 100% faster than other rivals. The results in Fig. 12 and 13 further illustrate the superiority of the proposed OGNet under different number of snapshots.

Further, the RMSE of different methods under different angular separations are compared. The corresponding result is depicted in Fig. 14, in which T = 500, SNR = 0 dB and the true DOA is set as = [$_1$; $_1$+] with $_1$ = −3:67° and changing from 1° to 5°. As clearly shown in Fig. 14, the proposed OGNet shows lower RMSE as compare to other comparison methods in close proximity, which indicate that the proposed OGNet has obviously performance advantage in terms of resolution.

In the end, the RMSE of different methods in the case of different grid intervals are evaluated. Since the CNN method is an on-grid method whose performance become poor or even invalid for the off-grid angles in coarse grid, only the OGSBI method is introduced to compare with the proposed OGNet. The different grid intervals are set as = {2°;3°;4°;5°}. The data generation and training strategies for OGNet when = 2° have been introduced in detail in previous Sect. 4.2. While, the data generation strategies

[†] This is because that deep networks are biased estimators (this holds for all DL-based estimators and not only the proposed) and the accuracy is limited by some factors such as network architecture and the size of training data-set. With fixed network structure and training data, when the model training is done, there must be an error between its prediction results and labels, resulting in the upper limit of its performance. When the model reaches this upper limit as SNR increases, it is difficult to show further improvements.

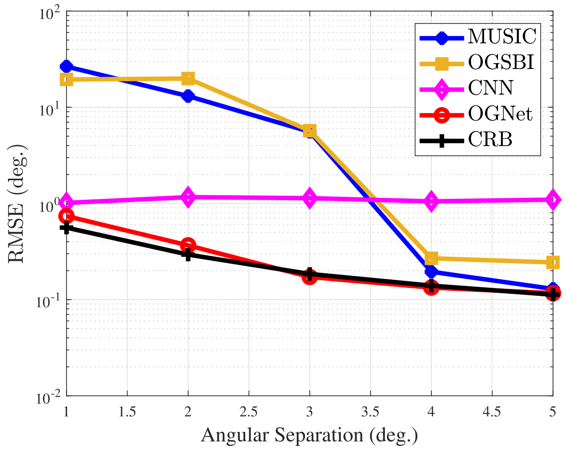


Fig. 14 RMSE of different methods under different angular separations.

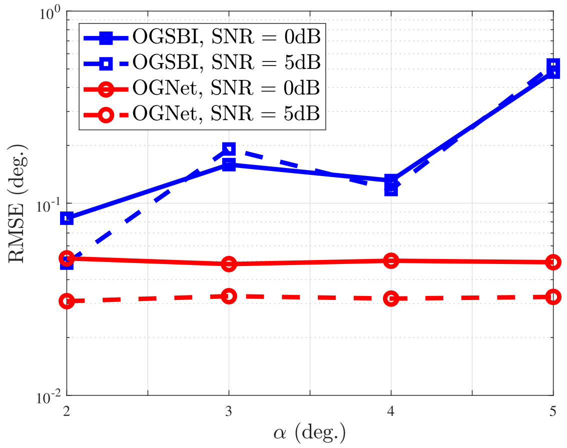


Fig. 15 RMSE of OGSBI and OGNet under different grid intervals.

for = {3°;4°;5°} are the same as that for = 2°, and the training strategies is similar with that for = 2° except for the batch size decreasing to 256. The results is shown in Fig. 15, as in which that the RMSE of the OGSBI increases with the increase of , while the RMSE of OGNet remains stable with the increase of . On the other hand, OGNet shows better performance than OGSBI in different SNRs and grid intervals. The results demonstrate that the proposed OGNet has performance benefits while being robust to different grid intervals.

6. Conclusion

In this paper, a cascade DNN named OGNet is designed for off-grid DOA estimation. Specifically, the upper triangular part of the sampling unitary covariance of array receiving data is firstly taken as the input of AE to reduce the difference between it and the theoretical, then the predicted unitary covariance of AE is input into the DCNN to predict the sparse off-grid error vector. The ultimate DOA estimation is realized by using the sparsity of the output of DCNN, which enables that the proposed OGNet can realize high-precision off-grid DOA estimation without a priori on-grid DOA estimation. The results under various scenarios indicate that the DOA estimation performance and resolution of the OGNet is remarkable and has noticeable advantages over its rivals.

References

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," IEEE Signal Process. Mag., vol.13, no.4, pp.67–94, July 1996.

[2] W. Liu, M. Haardt, M.S. Greco, C.F. Mecklenbräuker, and P. Willett, "Twenty-five years of sensor array and multichannel signal processing: A review of progress to date and potential research directions," IEEE Signal Process. Mag., vol.40, no.4, pp.80–91, 2023.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propag., vol.34, no.3, pp.276–280, March 1986.

[4] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," IEEE Trans. Acoust., Speech, Signal Process., vol.37, no.7, pp.984–995, July 1989.

[5] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," IEEE International Conference on Acoustics, Speech and Signal Processing, pp.336–339, Jan. 1983.

[6] B. Rao and K. Hari, "Performance analysis of root-MUSIC," IEEE Trans. Acoust., Speech, Signal Process., vol.37, no.12, pp.1939–1949, Dec. 1989.

[7] T.J. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," IEEE Trans. Acoust., Speech, Signal Process., vol.33, no.4, pp.806–811, Aug. 1985.

[8] K.C. Huarng and C.C. Yeh, "A unitary transformation method for angle-of-arrival estimation," IEEE Trans. Signal Process., vol.39, no.4, pp.975–977, April 1991.

[9] M. Haardt and J. Nossek, "Unitary ESPRIT: How to obtain increased estimation accuracy with a reduced computational burden," IEEE Trans. Signal Process., vol.43, no.5, pp.1232–1242, May 1995.

[10] D. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory, vol.52, no.4, pp.1289–1306, April 2006.

[11] K. Hayashi, M. Nagahara, and T. Tanaka, "A user's guide to compressed sensing for communications systems," IEICE Trans. Commun., vol.E96-B, no.3, pp.685–712, March 2013.

[12] T. Terada, T. Nishimura, Y. Ogawa, T. Ohgane, and H. Yamada, "DOA estimation for multi-band signal sources using compressed sensing techniques with Khatri-Rao processing," IEICE Trans. Commun., vol.E97-B, no.10, pp.2110–2117, Oct. 2014.

[13] Y. Liu, Z. Zhang, C. Zhou, C. Yan, and Z. Shi, "Robust variational Bayesian inference for direction-of-arrival estimation with sparse array," IEEE Trans. Veh. Technol., vol.71, no.8, pp.8591–8602, 2022.

[14] Z. Yang, J. Li, P. Stoica, and L. Xie, "CHAPTER 11 - Sparse methods for direction-of-arrival estimation," Academic Press Library in Signal Processing, Volume 7, R. Chellappa and S. Theodoridis, eds., pp.509–581, Academic Press, 2018.

[15] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," IEEE Trans. Signal Process., vol.53, no.8, pp.3010–3022, Aug. 2005.

[16] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," J. Mach. Learn. Res., vol.1, no.3, pp.211–244, Sept. 2001.

[17] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse Bayesian inference," IEEE Trans. Signal Process., vol.61, no.1, pp.38–43, Oct. 2012.

[18] J. Dai, X. Bao, W. Xu, and C. Chang, "Root sparse Bayesian learning for off-grid DOA estimation," IEEE Signal Process. Lett., vol.24, no.1, pp.46–50, Jan. 2017.

[19] Z. Yang, L. Xie, and C. Zhang, "A discretization-free sparse and parametric approach for linear array signal processing," IEEE Trans. Signal Process., vol.62, no.19, pp.4959–4973, Oct. 2014.

[20] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol.521, no.7553, pp.436–444, May 2015.
[21] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press., 2016.
[22] Y. Kase, T. Nishimura, T. Ohgane, Y. Ogawa, D. Kitayama, and Y. Kishiyama, “Fundamental trial on DOA estimation with deep learning,” IEICE Trans. Commun., vol.E103-B, no.10, pp.1127–1135, Oct. 2020.
[23] J. Cong, X. Wang, C. Yan, L.T. Yang, M. Dong, and K. Ota, “CRB weighted source localization method based on deep neural networks in multi-UAV network,” IEEE Internet Things J., vol.10, no.7, pp.5747–5759, 2023.
[24] Y. Kase, T. Nishimura, T. Ohgane, Y. Ogawa, T. Sato, and Y. Kishiyama, “Accuracy improvement in DOA estimation with deep learning,” IEICE Trans. Commun., vol.E105-B, no.5, pp.588–599, May 2022.
[25] G.K. Papageorgiou and M. Sellathurai, “Direction-of-arrival estimation in the low-SNR regime via a denoising autoencoder,” IEEE International Workshop on Signal Processing Advances in Wireless Communications, pp.1–5, Aug. 2020.
[26] G.K. Papageorgiou and M. Sellathurai, “Fast direction-of-arrival estimation of multiple targets using deep learning and sparse arrays,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4632–4636, April 2020.
[27] A. Barthelme and W. Utschick, “DoA estimation using neural network-based covariance matrix reconstruction,” IEEE Signal Process. Lett., vol.28, pp.783–787, 2021.
[28] X. Wu, X. Yang, X. Jia, and F. Tian, “A gridless DOA estimation method based on convolutional neural network with Toeplitz prior,” IEEE Signal Process. Lett., vol.29, pp.1247–1251, May 2022.
[29] Z.M. Liu, C. Zhang, and P.S. Yu, “Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections,” IEEE Trans. Antennas Propag., vol.66, no.12, pp.7315–7327, Dec. 2018.
[30] J. Cong, X. Wang, M. Huang, and L. Wan, “Robust DOA estimation method for MIMO radar via deep neural networks,” IEEE Sensors J., vol.21, no.6, pp.7498–7507, March 2021.
[31] L. Wu, Z.M. Liu, and Z.T. Huang, “Deep convolution network for direction of arrival estimation With sparse prior,” IEEE Signal Process. Lett., vol.26, no.11, pp.1688–1692, Nov. 2019.
[32] G.K. Papageorgiou, M. Sellathurai, and Y.C. Eldar, “Deep networks for direction-of-arrival estimation in low SNR,” IEEE Trans. Signal Process., vol.69, pp.3714–3729, June 2021.
[33] H. Wang, X. Wang, M. Huang, L. Wan, and T. Su, “RxCV-based unitary SBL algorithm for off-grid DOA estimation with MIMO radar in unknown non-uniform noise,” Digit. Signal Process., vol.116, p.103119, Sept. 2021.
[34] X.T. Meng, F.G. Yan, B.X. Cao, M. Jin, and Y. Zhang, “Efficient real-valued DOA estimation based on the trigonometry multiple angles transformation in monostatic MIMO radar,” Digit. Signal Process., vol.123, p.103437, 2022.
[35] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” Proc. 32nd International Conference on Machine Learning, F. Bach and D. Blei, eds., Proc. Machine Learning Research, vol.37, pp.448–456, PMLR, July 2015.
[36] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, Dec. 2014.
[37] P. Stoica and A. Nehorai, “Performance study of conditional and unconditional direction-of-arrival estimation,” IEEE Trans. Acoust., Speech, Signal Process., vol.38, no.10, pp.1783–1795, Oct. 1990.

Huafei Wang was born in 1995. He received the B.S. degree and M.S. degrees from Hainan University, Haikou, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree in information and communication engineering at Hainan University. His current research interests include array signal processing, radar signal processing using learning strategy.

Xianpeng Wang was born in 1986. He received the M.S. and Ph.D. degrees from the College of Automation, Harbin Engineering University, Harbin, China, in 2012 and 2015, respectively. He was a full-time Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2015 to 2016. He is currently a Professor with the School of Information and Communication Engineering, Hainan University, Haikou, China. He is the author of more than 100 papers published in related journals and international conference proceedings and was a Reviewer of more than 30 journals. His major research interests include communication system, array signal processing, radar signal processing, compressed sensing, and its applications.

Xiang Lan received the B.S. degree from the Huazhong University of Science and Technology, China, in 2012, and the M.Sc. and Ph.D. degrees from the Department of Electronic and Electrical Engineering, The University of Sheffield, U.K., in 2014 and 2019, respectively. From 2019 to June 2020, he worked as a Research Associate with the Department of Electronic and Electrical Engineering, The University of Sheffield. He is currently a Lecturer with the School of Information and Communication Engineering, Hainan University, China. His research interests include signal processing based on vector sensor arrays (beamforming and DOA estimation with polarized signals) and sparse array processing.

Ting Su received the B.S. degree in communication engineering and the Ph.D. degree in electronic science and technology from the Nanjing University of Science and Technology, Nanjing, China, in 2006 and 2016, respectively. She is currently a Postdoctoral Re-searcher with the Institute of Communications Engineering, Army Engineering University of PLA, Nanjing. Moreover, she is also a Lecturer with Hainan university, Haikou, China. Her research interests include computational electromagnetic, radar signal processing, and wireless communications.

PAPER

SLNR-Based Joint Precoding for RIS Aided Beamspace HAP-NOMA Systems

Pingping JI[†a)], **Lingge JIANG**[†], **Chen HE**[†], **Di HE**[†], ***and*** **Zhuxian LIAN**[††], ***Nonmembers***

SUMMARY High altitude platform (HAP), known as line-of-sight dominated communications, effectively enhance the spectral efficiency of wireless networks. However, the line-of-sight links, particularly in urban areas, may be severely deteriorated due to the complex communication environment. The reconfigurable intelligent surface (RIS) is employed to establish the cascaded-link and improve the quality of communication service by smartly reflecting the signals received from HAP to users without direct-link. Motivated by this, the joint precoding scheme for a novel RIS-aided beamspace HAP with non-orthogonal multiple access (HAP-NOMA) system is investigated to maximize the minimum user signal-to-leakage-plus-noise ratio (SLNR) by considering user fairness. Specifically, the SLNR is utilized as metric to design the joint precoding algorithm for a lower complexity, because the isolation between the precoding obtainment and power allocation can make the two parts be attained iteratively. To deal with the formulated non-convex problem, we first derive the statistical upper bound on SLNR based on the random matrix theory in large scale antenna array. Then, the closed-form expressions of power matrix and passive precoding matrix are given by introducing auxiliary variables based on the derived upper bound on SLNR. The proposed joint precoding only depends on the statistical channel state information (SCSI) instead of instantaneous channel state information (ICSI). NOMA serves multi-users simultaneously in the same group to compensate for the loss of spectral efficiency resulted from the beamspace HAP. Numerical results show the effectiveness of the derived statistical upper bound on SLNR and the performance enhancement of the proposed joint precoding algorithm.

key words: *high altitude platform, signal-to-leakage-plus-noise ratio, reconfigurable intelligent surface, large scale array, non-orthogonal multiple access*

1. Introduction

Quasi-stationary high altitude platform (HAP), as a promising communication technology in beyond 5G/6G networks, is located at an altitude of 17–22 km to provide the stratospheric communication services with large coverage, long flight duration and quick deployment [1]. It is an indispensable component of air-space-ground integrated information communication network, i.e. satellite, stratosphere and terrestrial wireless networks. To date, some works have studied the integration of HAP and various wireless communication technologies to meet the explosive growth in the requirement of high data transmission rate and massive connectivity, such as multiple-input multiple-output (MIMO) [1], artificial intelligence [2], hybrid beamforming [3], non-orthogonal multiple access (NOMA) [4] and reconfigurable intelligent surface (RIS) [5].

The introduction of NOMA to the beamspace HAP can tackle the issue of limited number of users by offering service for multi-users in the same time-frequency-space resource block, and improve the spectral efficiency (SE) by deploying superposition coding at the HAP and successive interference cancellation (SIC) at the terminal users [6]. The HAP realize NOMA in power domain other than time, frequency, or code domain, which can lengthen the endurance time of HAP by lowering the number of radio frequency and ensure the quality of service for HAP by allocating more power to users with poor conditions [4].

Recently, RIS has attracted worldwide attention from the academia and industry due to its characteristics of low cost and portability, which can be directly deployed in existing wireless networks without any other hardwares [7]–[11]. An RIS is a man-made electronmagnetic surface composed of a large number of programmable reconfigurable passive elements and a smart controller [7]. The controller adjusts the passive element, i.e. phase shift, by reflecting the incident signals to desired directions aiming at improving the communication service. In practice, it is prone to suffer severe path loss and serious signal blockage for the link between HAP and terrestrial users, and the RIS can overcome this by assisting the HAP to establish the cascaded-link, i.e. HAP-RIS-user link [5].

The key problem in RIS-aided communication systems is how to solve the joint precoding optimization problem, due to the non-convex objective functions and constraints. To maximize the sum rate in the form of a sum-of-log-ratio, some auxiliary variables has been introduced to derive the closed-form expressions of the passive precoding matrix at RIS and active precoding matrix at the base station [8]. However, each item in the reflecting matrix is highly related with the other items due to the utilization of production between vector and matrix, which incurs a lower complexity at the cost of SE. The first-order Taylor expansion has been used in [9] to transform the non-convex objective function into convex form, which result in performance loss. The maximization of the weighted sum rate has been solved by designing a joint transmit precoding and reflect precoding optimization scheme in [10]. The gradient-projection method has been utilized in the obtainment of the closed-form of passive and active beamforming matrices in [11]. The above works all focus on the instantaneous channel state information (ICSI), which is challenging to be acquired on HAP due to the large distance between HAP and users compared with the terrestrial systems.

Manuscript received March 22, 2024.
Manuscript publicized May 6, 2024.
[†] Shanghai Jiaotong University, China.
[††] Jiangsu University of Science and Technology, China.
a) E-mail: ji_pingping@sjtu.edu.cn
DOI: 10.23919/transcom.2024EBP3045

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

The statistical channel state information (SCSI) has been utilized in [7] and [4]. The closed-form sum rate has been derived in a simple scenery which is not suitable in practice in [7] and the correlation of statistical correlation matrix has been analyzed. The mean-square-error has been researched to reformulate the non-convex ergodic sum rate objective into the convex form in [4] and the partial SCSI is used due to the digital precoder. Obviously, the feedback overhead, i.e. ICSI, becomes exponentially higher with the growth of the number of transmit antennas. Therefore, we design the schemes with SCSI [4] rather than ICSI [6] for the beamspace HAP-NOMA system.

It has been already shown in [9] and [4] that the computation of designing schemes in this downlink scenario is difficult with the signal-to-interference-plus-noise-ratio (SINR) criterion. The adopt of the signal-to-leakage-plus-noise-ratio (SLNR) criterion can reduce the computational complexity by iteratively designing the active beamforming and the power allocation as two disjoint subproblems [12].

In this paper, we study a downlink transmission algorithm for the RIS assisted beamspace HAP-NOMA systems according to SLNR. We aim to maximize the minimum SLNR of all users by jointly designing the passive precoding at the RIS and power allocation on HAP. The main contributions of this paper can be summarized as follows:

- A novel RIS-aided beamspace HAP-NOMA system is proposed. The NOMA combining beamspace HAP improves the achievable sum rate by improving the number of serving users simultaneously. The combination of RIS and beamspace HAP-NOMA makes the users without direct-link from HAP could achieve better service.
- The statistical upper bound on SLNR is derived according to the random matrix theory in large scale antenna. SLNR is used as performance measure for the RIS-aided HAP-NOMA system to reduce the computation complexity by decoupling the power allocation and passive precoding. The minimum SLNR is maximized to consider the user fairness.
- The closed form expressions of power allocation matrix and passive precoding matrix are obtained by introducing a series of auxiliary variables. To be specific, The Lagrange multiplier method is used to solve the power allocation problem, and the bisection method is used to solve the passive precoding problem.
- Numerical results show that the derived upper bound is effective and the proposed algorithm has an distinct performance enhancement.

The rest of this paper is organized as follows. Section 2 introduces the channel model for the novel RIS-aided beamspace HAP-NOMA system. Section 3 presents the main results, where a statistical upper bound on SLNR will be derived and a joint precoding algorithm will be proposed iteratively. Section 4 presents the simulation analysis and Sect. 5 concludes the paper.

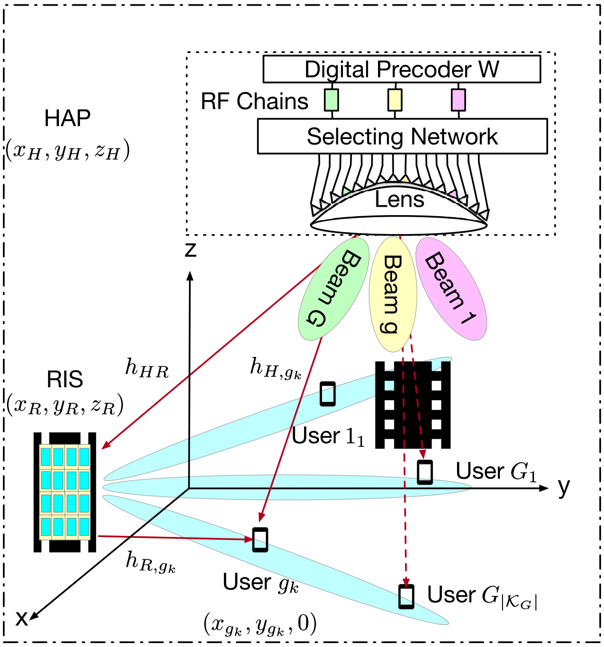


Fig. 1 System model of the proposed RIS-aided beamspace HAP-NOMA.

2. System Model

Consider a downlink RIS-aided beamspace HAP-NOMA system as shown in Fig. 1, where the HAP is equipped with a uniform planar antenna (UPA) of $M = M_v \times M_h$ transmit antennas and G RF chains. The K single-antenna users is served with the help of one RIS, which is equipped with $N = N_v \times N_h$ elements of UPA. The vertical and horizontal antenna spacing $d_0 = \lambda/2$ of RIS are same as that of HAP, λ is the carrier wavelength.

Denote the location of HAP, RIS and user k in group g as (x_H, y_H, z_H), (x_R, y_R, z_R) and $(x_{g_k}, y_{g_k}, 0)$. $\vartheta_{t,r} \in [0, \pi/2)$ and $\varphi_{t,r} \in [-\pi, \pi]$ are the vertical and horizontal angle of departure from transmitter to receiver, $\vartheta_{t,r} = \arctan(\sqrt{(x_t - x_r)^2 + (y_t - y_r)^2}/|z_t - z_r|)$ and $\varphi_{H,g_k} = \arccos((x_r - x_t)/\sqrt{(x_t - x_r)^2 + (y_t - y_r)^2}) \cdot \mathrm{sgn}(y_r - y_t)$, $\mathrm{sgn}(\cdot)$ is a sign function.

Denote the set of all users and the set of users in group g as $\mathcal{K}$ and $\mathcal{K}_g$ respectively. Denote the set of users with weak or no direct-link as $\mathcal{K}'$. Obviously, the number of groups is equal to the number of RF chains as G such that $|\mathcal{K}_g| \geq 1$, $\mathcal{K}_i \cap \mathcal{K}_j = \emptyset$ for $i \neq j$, $\sum_{g=1}^{G} |\mathcal{K}_g| = K$ and $\mathcal{K} = \bigcup_{g=1}^{G} \mathcal{K}_g$.

The user $\{k | k \in (\mathcal{K} - \mathcal{K}')\}$ with direct-link should be assigned to the beam b_{H,g_k} and the user $\{k | k \in \mathcal{K}'\}$ with weak or no direct-link should be in the beam b_{HR}, where $b_{t,r} = M_v(m_{t,r} - 1) + n_{t,r}$ with $m_{t,r} = \frac{M_v}{2} \sin\vartheta_{t,r} \cos\varphi_{t,r} + 1, \varphi_{t,r} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $m_{t,r} = \frac{M_v}{2} \sin\vartheta_{t,r} \cos\varphi_{t,r} + M_v + 1, \varphi_{t,r} \notin [-\frac{\pi}{2}, \frac{\pi}{2}]$, $n_{t,r} = \frac{M_h}{2} \sin\vartheta_{t,r} \sin\varphi_{t,r} + 1, \varphi_{t,r} \in [0, \pi]$ and $n_{t,r} = \frac{M_h}{2} \sin\vartheta_{t,r} \sin\varphi_{t,r} + M_h + 1, \varphi_{t,r} \in [-\pi, 0)$ [4].

The NOMA with successive interference cancelation is utilized in the g-th group and the beamspace HAP-NOMA

Table 1 Definition of parameters used in Eq. (1).

x_{g_k}, y_{g_k}	g_kth transmitted and received signal, $\mathbb{E}[x_{g_k} x_{g_k}^H] = 1$
$\vec{\mathbf{h}}_{g_k}$	g_kth $1 \times G$ beamspace channel vector, $\vec{\mathbf{h}}_{g_k} = \mathbf{h}_{g_k}\mathbf{U}_s$
$\mathbf{h}_{g_k}$	g_kth $1 \times M$ channel vector, $\mathbf{h}_{g_k} = \mathbf{h}_{H,g_k} + \mathbf{h}_{R,g_k}\mathbf{\Phi}\mathbf{h}_{HR}$
$\mathbf{h}_{H/R,g_k}$	channel vector from HAP/RIS to kth user in gth group
$\mathbf{h}_{HR}$	$N \times M$ channel matrix from HAP to RIS
$\mathbf{U}_s$	$M \times G$ selected beam matrix, $\mathbf{U}_s = \mathbf{U}(:, b)_{b\in\mathcal{B}}$
$\mathcal{B}$	the set of selected beams, $\lvert\mathcal{B}\rvert = G$
$\mathbf{w}_g$	gth $G \times 1$ digital precoding vector, $\lvert\mathbf{w}_g\rvert = 1$
$\mathbf{\Phi}$	$N \times N$ diagonal phase shift matrix at RIS
p_{g_k}	g_kth transmitted power, $\sum_{j_i\in\mathcal{K}} p_{j_i} \leq P$
P	total transmitted power
n	additive white Gaussian noise (AWGN), $\mathcal{CN}(0, \sigma^2)$
σ^2	noise variance

channels of users in g-th group satisfy $|\vec{\mathbf{h}}_{g_1}\mathbf{w}_g|^2 \geq \cdots \geq |\vec{\mathbf{h}}_{g_{|\mathcal{K}_g|}}\mathbf{w}_g|^2$. The received signal at the kth user in gth group can be represented as [9]

$$\begin{aligned} y_{g_k} = &\underbrace{\vec{\mathbf{h}}_{g_k}\mathbf{w}_g\sqrt{p_{g_k}}x_{g_k}}_{\text{desired signal}} + \underbrace{\vec{\mathbf{h}}_{g_k}\mathbf{w}_g\sum_{i=1}^{k-1}\sqrt{p_{g_i}}x_{g_i}}_{\text{intra-beam interference}} \\ &+ \underbrace{\vec{\mathbf{h}}_{g_k}\sum_{j\neq g}\sum_{i=1}^{\mathcal{K}_j}\mathbf{w}_j\sqrt{p_{j_i}}x_{j_i}}_{\text{inter-beam interference}} + \underbrace{n_{g_k}}_{\text{noise}}, \end{aligned} \tag{1}$$

where the parameters are defined in Table 1, $\mathbf{h}_{H/R,g_k}$ and $\mathbf{h}_{HR}$ are defined as [1]

$$\begin{aligned} \mathbf{h}_{H,g_k} = &\sqrt{\alpha_{H,g_k}\rho_{H,g_k}}\mathbf{a}(\vartheta_{H,g_k}, \varphi_{H,g_k}, M)+ \\ &\sqrt{\alpha_{H,g_k}(1-\rho_{H,g_k})}\mathbf{h}^w_{H,g_k}\widetilde{\mathbf{R}}^{1/2}_{H,g_k}, \end{aligned} \tag{2a}$$

$$\begin{aligned} \mathbf{h}_{R,g_k} = &\sqrt{\alpha_{R,g_k}\rho_{R,g_k}}\mathbf{a}(\vartheta_{R,g_k}, \varphi_{R,g_k}, N)+ \\ &\sqrt{\alpha_{R,g_k}(1-\rho_{R,g_k})}\mathbf{h}^w_{R,g_k}\widetilde{\mathbf{R}}^{1/2}_{R,g_k}, \end{aligned} \tag{2b}$$

$$\mathbf{h}_{HR} = \sqrt{\alpha_{HR}}\mathbf{a}(\vartheta_{RH}, \varphi_{RH}, N)^H\mathbf{a}(\vartheta_{HR}, \varphi_{HR}, M), \tag{2c}$$

$$\left[\widetilde{\mathbf{R}}_{t,r}\right]_{p,q} = \int_0^{\frac{\pi}{2}}\int_{-\pi}^{\pi} f(\varphi)f(\theta)e^{(j\frac{2\pi}{\lambda}(d_1+d_2))}d\varphi d\theta, \tag{2d}$$

where $f(\varphi) = \frac{e^{(\kappa_{t,r}\cos(\varphi-\mu_{t,r}))}}{2\pi I_0(\kappa)}$, $I_0(\cdot)$ is the zeroth-order modified Bessel function of first kind, $\mu_{t,r} \in [-\pi, \pi]$ is the horizontal AoD, $\kappa_{t,r}$ controls the angular spread (AS), $f(\theta) \propto e^{(-\sqrt{2}|\theta-\theta'_{t,r}|/\delta_{t,r})}$, $\theta'_{t,r}$ and $\delta_{t,r}$ are the mean vertical AoD and AS, $d_1 = (p-q)d_0\sin\theta\cos\varphi$ and $d_2 = (p-q)d_0\sin\theta\sin\varphi$.

Assume $\mathbf{U} = \mathbf{D}(M_v) \otimes \mathbf{D}(M_h) \in \mathbb{C}^{M\times M}$ is the spatial discrete Fourier transformation (DFT) matrix for 3D lens antenna array, where $\mathbf{D}(M) = \frac{1}{\sqrt{M}}[\mathbf{a}_M(0), \mathbf{a}_M(1/M), \cdots, \mathbf{a}_M((M-1)/M)]^H$ and $\mathbf{a}_M(x) = [1, e^{-j2\pi x}, \cdots, e^{-j2\pi(M-1)x}]$,

$$\mathbf{a}(\vartheta, \varphi, N) = \mathbf{a}_{N_v}(\frac{d_0\sin\vartheta\cos\varphi}{\lambda}) \otimes \mathbf{a}_{N_h}(\frac{d_0\sin\vartheta\sin\varphi}{\lambda}), \tag{3}$$

where $\rho_{t,r} = K_{t,r}/(1+K_{t,r})$, $K_{t,r}$ is the Rician factors, the large-scale fading fctor $\alpha_{t,r} = G_{t,r}(4\pi d_{t,r}/\lambda)^{-2}$ [1], $G_{t,r}$ is the effect of antenna gain from transmitter to receiver, $d_{t,r}$ is the distance between the transmitter and receiver. Finally, $\mathbf{h}^w_{H,g_k} \in \mathbb{C}^{1\times M}$ and $\mathbf{h}^w_{R,g_k} \in \mathbb{C}^{1\times N}$ are i.i.d. satisfying the complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I})$.

We further define $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_N]^H$, $c_{g_k} = \boldsymbol{\phi}^H\mathbf{c}_{g_k}$ and $\Xi_{g_k} = \boldsymbol{\phi}\mathbf{\Xi}_{g_k}\boldsymbol{\phi}^H$

$$\begin{aligned} \mathbf{c}_{g_k} = &\sqrt{\alpha_{HR}\alpha_{R,g_k}\rho_{R,g_k}} \\ &\cdot \operatorname{diag}(\mathbf{a}(\vartheta_{R,g_k}, \varphi_{R,g_k}, N))\mathbf{a}(\vartheta_{RH}, \varphi_{RH}, N)^H, \end{aligned} \tag{4a}$$

$$\begin{aligned} \mathbf{\Xi}_{g_k} = &\alpha_{HR}\alpha_{R,g_k}(1-\rho_{R,g_k})\operatorname{diag}(\mathbf{a}(\vartheta_{RH}, \varphi_{RH}, N)) \\ &\widetilde{\mathbf{R}}_{R,g_k}\operatorname{diag}(\mathbf{a}(\vartheta_{RH}, \varphi_{RH}, N)^H). \end{aligned} \tag{4b}$$

The SLNR of the k-th user in g-th group is [12]

$$\text{SLNR}_{g_k} = \frac{|\vec{\mathbf{h}}_{g_k}\mathbf{w}_g|^2 p_{g_k}}{\sum_{j_i\in\mathcal{K}_{-g_k}}|\vec{\mathbf{h}}_{j_i}\mathbf{w}_g|^2 p_{g_k} + \sigma^2}, \tag{5}$$

where $\mathcal{K}_{-g_k} = (\bigcup_{j\neq g}\mathcal{K}_j)\bigcup\{g_1, \cdots, g_{k-1}\}$.

According to generalized eigenvalue decomposition [12] and equation (5), the optimum $\mathbf{w}_g = \frac{\mathbf{w}'_g}{|\mathbf{w}'_g|^2}$ with

$$\mathbf{w}'_g = (\sum_{j_i\in\mathcal{K}_{-g_1}}\vec{\mathbf{h}}^H_{j_i}\vec{\mathbf{h}}_{j_i} + \frac{\sigma^2}{p_{g_1}}\cdot\mathbf{I})^{-1}\vec{\mathbf{h}}^H_{g_1}. \tag{6}$$

3. Joint Precoding Scheme

In this section, we derive the upper bound of the SLNR. Then, the joint precoding algorithm is proposed to maximize the minimum user SLNR.

3.1 The Upper Bound on SLNR

We aim to maximize the SLNR only with the knowledge of SCSI. Assume $b_{HR} = G$ and $\mathcal{K}_{b_{HR}} = \mathcal{K}_G$. We derive the upper bound of SLNR in next theorem, which is proved in Appendix.

Theorem 1. *The upper bound of* SLNR_{g_k} *based on* (5) *can be further given as*

$$\text{SLNR}^u_{g_k} = \frac{\delta'_{g,G}M\alpha_{H,g_k}\rho_{H,g_k}}{\frac{\sigma^2}{p_{g_k}} + \frac{\delta_{g,G}(|\sqrt{\alpha_{H,g_k}\rho_{H,g_k}}+c_{g_k}|^2+\Xi_{g_k})+\delta'_{g,G}(c^H_{g_k}c_{g_k}+\Xi_{g_k})}{\sum\limits_{j_i\in\mathcal{K}_{-g_1}}(\frac{\Xi_{j_i}}{1+m_{j_i}}+\delta'_{j,G}c^H_{j_i}c_{j_i})+\eta+\frac{\sigma^2}{Mp_{g_k}}-\zeta_g}}, \tag{7}$$

where

$$\begin{aligned} \zeta_g &= \textstyle\sum_{l\neq g,G}\frac{\sum\limits_{n\in\mathcal{K}_l}\sqrt{\alpha_{H,l_n}\rho_{H,l_n}}c^H_{l_n}\sum\limits_{n\in\mathcal{K}_l}\sqrt{\alpha_{H,l_n}\rho_{H,l_n}}c_{l_n}}{\sum\limits_{n\in\mathcal{K}_l}\alpha_{H,l_n}\rho_{H,l_n}} \\ &= \boldsymbol{\phi}^H\boldsymbol{\zeta}_g\boldsymbol{\phi}, \end{aligned} \tag{8a}$$

$$\boldsymbol{\zeta}_g = \textstyle\sum_{l\neq g,G}\frac{\sum\limits_{n\in\mathcal{K}_l}\sqrt{\alpha_{H,l_n}\rho_{H,l_n}}\mathbf{c}^H_{l_n}\sum\limits_{n\in\mathcal{K}_l}\sqrt{\alpha_{H,l_n}\rho_{H,l_n}}\mathbf{c}_{l_n}}{\sum\limits_{n\in\mathcal{K}_l}\alpha_{H,l_n}\rho_{H,l_n}}, \tag{8b}$$

$$\eta = \sum_{i=1}^{|\mathcal{K}_G|}|\sqrt{\alpha_{H,G_i}\rho_{H,G_i}} + c_{G_i}|^2, \tag{8c}$$

$$\begin{aligned} m_{g_k} = \Xi_{g_k}/(&\sum_{j_i\in\mathcal{K}_{-g_1}}(\frac{\Xi_{j_i}}{1+m_{j_i}} + \delta'_{j,G}c^H_{j_i}c_{j_i}) \\ &+\eta + \frac{\sigma^2}{Mp_{g_k}} - \zeta_g), \end{aligned} \tag{8d}$$

with the symbol m_{g_k} can be iteratively achieved by $m_{g_k} = \lim_{t\to\infty} m_{g_k}^{(t)}$ and $m_{g_k}^{(0)} = M$, and $\delta_{x,y}$ is a unit-impulse function, the value of which is zero when $x = y$ and one when $x \neq y$, $\delta'_{x,y} = 1 - \delta_{x,y}$.

3.2 Proposed Joint Precoding Scheme

The proposed joint precoding algorithm has been clarified in Algorithm 1. We obtain the $\mathbf{\Phi}, p_{g_k}$ closed-form optimal solution alternately by fixing the other variables.

Under the derivation of Theorem 1, the minimum user SLNR maximization problem can be formulated as

$$\begin{aligned}(\mathcal{P})\max_{p_{g_k},\mathbf{\Phi}}\ & \min\ \mathrm{SLNR}_{g_k}^{u}\\ s.t.\ & C_1:\ \ p_{g_k} \geq 0,\\ & C_2:\ \ \textstyle\sum_{g=1}^{G}\sum_{k=1}^{|\mathcal{K}_g|} p_{g_k} \leq P,\\ & C_3:\ \ |\theta_n| = 1, \forall n,\end{aligned} \tag{9}$$

where C_1 ensures each user can be allocated with non-negative power, C_2 satisfies the maximum total power demand at the HAP, C_3 indicates the constraints of phase shift matrix $\mathbf{\Theta}$.

3.2.1 Algorithm for Optimizing p_{g_k} Given $\boldsymbol{\phi}$

By introducing auxiliary variable t, the problem $\mathcal{P}$ can be transformed as

$$\begin{aligned}&(\mathcal{P}_{p_{g_k}})\ \min_{\{p_{g_k}\}}\ \ t^{-1}\\ &s.t.\ \ C_1, C_2,\\ &C_4: t \leq \mathrm{SLNR}_{g_k}^{u} = A_{g_k}^{(1)} p_{g_k} + \frac{B_{g_k}^{(1)}}{C_{g_k}^{(1)} + \frac{\sigma^2}{p_{g_k}}}\end{aligned} \tag{10}$$

where

$$A_{g_k}^{(1)} = \delta'_{g,G} M \alpha_{H,g_k} \rho_{H,g_k} / \sigma^2, \tag{11a}$$

$$\begin{aligned}B_{g_k}^{(1)} &= \delta_{g,G}(|\sqrt{\alpha_{H,g_k}\rho_{H,g_k}} + c_{g_k}|^2 + \Xi_{g_k})\\ &+\delta'_{g,G}(c_{g_k}^{H} c_{g_k} + \Xi_{g_k}),\end{aligned} \tag{11b}$$

$$C_{g_k}^{(1)} = \sum_{j_i \in \mathcal{K}_{-g_1}} (\tfrac{\Xi_{j_i}}{1+m_{j_i}} + \delta'_{j,G} c_{j_i}^{H} c_{j_i}) + \eta - \zeta_g. \tag{11c}$$

Problem $\mathcal{P}_{p_{g_k}}$ is non-convex resulted from the multi-variable coupling in C_4, another auxiliary variables $\{b_{g_k}\}$ is introduced to make it solvable. And the problem $\mathcal{P}_{p_{g_k}}$ can be equivalently reformulated as

$$\begin{aligned}&(\mathcal{P}'_{p_{g_k}})\ \min_{\{p_{g_k}\}}\ \ t^{-1}\\ &s.t.\ \ C_1, C_2,\\ &C_4^{(1)}: A_{g_k}^{(1)} C_{g_k}^{(1)} p_{g_k}^2 + (A_{g_k}^{(1)}\sigma^2 + B_{g_k}^{(1)}) p_{g_k}\\ &\qquad\qquad \geq C_{g_k}^{(1)} b_{g_k} + \sigma^2 t,\\ &C_4^{(2)}: \begin{bmatrix} t, & \sqrt{b_{g_k}}\\ \sqrt{b_{g_k}}, & p_{g_k}\end{bmatrix} \geq 0,\\ &C_4^{(3)}: t \geq 0.\end{aligned} \tag{12}$$

The Lagrangian function of $(\mathcal{P}'_{p_{g_k}})$ is

$$\begin{aligned}&L(p_{g_k}, t, \boldsymbol{b}, \mu, \boldsymbol{\tau}, \boldsymbol{\kappa})\\ &= t^{-1} + \mu(\textstyle\sum_{g=1}^{G}\sum_{k=1}^{|\mathcal{K}_g|} p_{g_k} - P)\\ &+ \textstyle\sum_{g=1}^{G}\sum_{k=1}^{|\mathcal{K}_g|} \tau_{g_k}(C_{g_k}^{(1)} b_{g_k} + \sigma^2 t - A_{g_k}^{(1)} C_{g_k}^{(1)} p_{g_k}^2\\ &\qquad\qquad + (A_{g_k}^{(1)}\sigma^2 + B_{g_k}^{(1)}) p_{g_k})\\ &+ \textstyle\sum_{g=1}^{G}\sum_{k=1}^{|\mathcal{K}_g|} \kappa_{g_k}(t - b_{g_k} p_{g_k}),\end{aligned} \tag{13}$$

where μ, $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$ are the Lagrange multipliers of C_2, $C_4^{(1)}$ and $C_4^{(2)}$ respectively. Therefore, the optimal p_{g_k} is [15]

$$p_{g_k}^{opt} = \frac{\mu + \tau_{g_k}(A_{g_k}^{(1)}\sigma^2 + B_{g_k}^{(1)}) - \kappa_{g_k} b_{g_k}}{2\tau_{g_k} A_{g_k}^{(1)} C_{g_k}^{(1)}}. \tag{14}$$

3.2.2 Algorithm for Optimizing $\mathbf{\Phi}$ Given p_{g_k}

We use the sub-multiplicativity of norm $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$ and transform the item as $\sqrt{\alpha_{H,g_k}\rho_{H,g_k}} c_{g_k} \leq (\alpha_{H,g_k}\rho_{H,g_k}\|\boldsymbol{c}_{g_k}\|)(\boldsymbol{\phi}\boldsymbol{\phi}^{H})$, then

$$\begin{aligned}\mathrm{SLNR}_{g_k}^{u} \leq\ & \frac{\mathrm{Tr}(\mathbf{A}_{g_k}\boldsymbol{\phi}\boldsymbol{\phi}^{H} + B_{g_k})}{\mathrm{Tr}(\mathbf{C}_{g_k}\boldsymbol{\phi}\boldsymbol{\phi}^{H} + D_{g_k})}\\ &+ \frac{\delta'_{g,G}\alpha_{H,g_k}\rho_{H,g_k}}{\frac{\sigma^2}{M p_{g_k}}},\end{aligned} \tag{15}$$

where

$$\mathbf{A}_{g_k} = (\boldsymbol{c}_{g_k}\boldsymbol{c}_{g_k}^{H} + \mathbf{\Xi}_{g_k} + \delta_{g,G} 2\alpha_{H,g_k}\rho_{H,g_k}\|\boldsymbol{c}_{g_k}\|)\mathbf{I}), \tag{16a}$$

$$B_{g_k} = \delta_{g,G}\alpha_{H,g_k}\rho_{H,g_k}, \tag{16b}$$

$$\mathbf{C}_{g_k} = \sum_{j_i \in \mathcal{K}_{-g_1}} (\tfrac{\mathbf{\Xi}_{j_i}}{1+m_{j_i}} + \delta'_{j,G}\boldsymbol{c}_{j_i}\boldsymbol{c}_{j_i}^{H}) + \sum_{i=1}^{|\mathcal{K}_G|} \boldsymbol{c}_{G_i}\boldsymbol{c}_{G_i}^{H} - \boldsymbol{\zeta}_g, \tag{16c}$$

$$D_{g_k} = \frac{\sigma^2}{M p_{g_k}} + \sum_{i=1}^{|\mathcal{K}_G|} \alpha_{H,G_i}\rho_{H,G_i}. \tag{16d}$$

By introducing two auxiliary variables t_1 and $\mathbf{Z} = \boldsymbol{\phi}\boldsymbol{\phi}^{H}$, and globally relaxing the feasible set of $\mathbf{Z}$ to its convex hull as $\{\mathbf{Z}|\mathrm{Tr}(\mathbf{Z}) = N, \mathbf{0} \leq \mathbf{Z} \leq \mathbf{I}\}$ [14]. The $\mathcal{P}$ is reformulated as

$$\begin{aligned}&(\mathcal{P}_{\boldsymbol{\phi}})\max_{\mathbf{Z}}\ \ t_1\\ &s.t.\ \ C_3,\\ &C_5: \mathrm{Tr}(\mathbf{A}_{g_k}\mathbf{Z} + B_{g_k})/\mathrm{Tr}(\mathbf{C}_{g_k}\mathbf{Z} + D_{g_k}) \geq t_1,\\ &C_6: \mathrm{Tr}(\mathbf{Z}) = N,\\ &C_7: \mathbf{0} \leq \mathbf{Z} \leq \mathbf{I}.\end{aligned} \tag{17}$$

The problem $\mathcal{P}_{\boldsymbol{\phi}}$ is convex when t_1 is fixed. The bisection search method [14] can be utilized by setting

$$t_u = [(\lambda_{\mathbf{A}_{g_k}}^{\max} + B_{g_k})/(\lambda_{\mathbf{C}_{g_k}}^{\min} + D_{g_k})]_{g,k}^{\max}, \tag{18a}$$

$$t_l = [(\lambda_{\mathbf{A}_{g_k}}^{\min} + B_{g_k})/(\lambda_{\mathbf{C}_{g_k}}^{\max} + D_{g_k})]_{g,k}^{\min}, \tag{18b}$$

where $\lambda_{\mathbf{A}}^{\max/\min}$ is the maximum or minimum eigenvalue of matrix $\mathbf{A}$.

Given $\hat{t} \in \mathbb{R}$, the feasible Semi-Definite Programming problem $\mathcal{P}'_{\boldsymbol{\phi}}$ means $\hat{t}$ is a feasible solution of $\mathcal{P}_{\boldsymbol{\phi}}$, then the

Algorithm 1 Proposed Joint Precoding algorithm

```
Input: System parameters and SCSI.
 1: Initialization: $t = 1$, ${p_{g_k}}^0$ and $\mathbf{\Phi}^0$. The maximum iterative number is
    set as $T$. The preset threshold $\xi_t$.
 2: while $t \leq T$ do
 3:     Obtain the optimal $p_{g_k}$ by (14).
 4:     Calculate $t_u$ and $t_l$ by (18).
 5:     while $t_u - t_l > \xi_t$ do
 6:         $t_1 = (t_l + t_u)/2$.
 7:         Solve $\mathcal{P}'_{\boldsymbol{\phi}}$ with $t_1$ to check the feasibility of problem $\mathcal{P}'_{\boldsymbol{\phi}}$.
 8:         if $\mathcal{P}'_{\boldsymbol{\phi}}$ is feasible then
 9:             Obtain $\mathbf{Z}_{opt}$ and $\mathbf{\Phi}_{opt}$ by (19) and (20).
10:             $t_l = t_1$.
11:         else
12:             $t_u = t_1$.
13:         end if
14:     end while
15:     $t = t + 1$.
16: end while
Output: $\mathbf{\Phi}$ and $p_{g_k}, \forall g, k$.
```

optimal solution of $\mathcal{P}_{\boldsymbol{\phi}}$ is in the interval of $[\widehat{t}, t_u]$. Otherwise, the infeasible SDP problem $\mathcal{P}'_{\boldsymbol{\phi}}$ means the optimal solution of $\mathcal{P}_{\boldsymbol{\phi}}$ is in the interval of $[t_l, \widehat{t}]$.

A recursive algorithm is designed from the step 5 to step 14 in Algorithm 1 to find out the optimal solution of $\mathcal{P}_{\boldsymbol{\phi}}$ by gradually narrowing down the feasible interval. Firstly, the initial value of t_1 is given as $(t_l + t_u)/2$, where t_l and t_u are denoted in (18). Next, the feasibility of $\mathcal{P}_{\boldsymbol{\phi}}$ is checked. If it is feasible, t_l is updated as t_1; otherwise, t_u is updated as t_1. Repeat the above steps until $t_u - t_l \leq \xi_t$, where ξ_t is the preset threshold [16].

With given t_1, the problem $\mathcal{P}_{\boldsymbol{\phi}}$ can be transformed as

$$
\begin{aligned}
&(\mathcal{P}'_{\boldsymbol{\phi}}) \text{ find } \mathbf{Z} \\
&s.t. \ \ C_3, C_5, C_6, C_7.
\end{aligned} \tag{19}
$$

Numerical program solvers can be used to solve the convex problem $\mathcal{P}'_{\boldsymbol{\phi}}$, and we can obtain the optimal $\mathbf{Z}_{opt}$. We then construct $\mathbf{\Phi}_{opt}$ based on $\boldsymbol{v}_1$, where $\boldsymbol{v}_1$ is the eigenvector corresponding to the largest eigenvalues of $\mathbf{Z}_{opt}$, as

$$\mathbf{\Phi}_{opt} = \text{diag}(\exp(j \arg(\boldsymbol{v}_1))). \tag{20}$$

3.2.3 Complexity Analysis

In each iteration, μ, $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$ in (13) can be attained with the complexity $O(K^2 \log_2(\delta))$ [15], where δ is denoted as the preset accuracy. The complexity of the obtainment of t_u and t_l in (18) is $O(N^3)$. The worst case to obtain $\mathbf{Z}_{opt}$ and $\mathbf{\Phi}_{opt}$ has $\log_2((t_u - t_l)/\xi_t)$ iterations [14] with the complexity of $O(N^4 + N^2 K)$ in each iteration [17]. The main complexity of the proposed joint precoding algorithm is

$$
\begin{aligned}
C_{pro} = O\{T[&(N^4 + N^2 K)\log_2((t_u - t_l)/\xi_t) \\
&+ K^2 \log_2(\delta)]\}.
\end{aligned} \tag{21}
$$

The main complexity in [9] is

$$
\begin{aligned}
C_1 = O\{T[&\max(3K + 1 + G, KG)^4 \times \sqrt{KG} \log(\tfrac{1}{\xi}) \\
&+ (3K + N + 1)^4 \sqrt{N + 1} \log(\tfrac{1}{\xi}) \\
&+ (3K + N + 2)^4 \sqrt{N + 1} \log(\tfrac{1}{\xi})]\},
\end{aligned} \tag{22}
$$

where ξ is the required accuracy in [9].

The main complexity in [4] is

$$C_2 = O\{T[K^2 \log_2(\delta)]\}, \tag{23}$$

where δ is the required accuracy in [4].

Obviously, the complexity C_2 in (23) of [4] is the lowest as seen from (21)–(23). The second term in (21) is too small that it can be neglected. The t_u and t_l in (21) are defined in (16) and (18). Due to the large-scale fading on HAP, the t_u and t_l are 10^{-5} and 10^{-15} approximately by 10^3 times of repetitive simulations. The value of $\log_2((t_u - t_l)/\xi_t)$ is approximately equal to $\log(\frac{1}{\xi})$, and the value of $N^4 + N^2 K \approx N^4$ because the number of antennas N is too larger than the number of users K in this paper. Finally, we can get that $C_2 < C_{pro} < C_1$.

4. Numerical Results and Analysis

In this section, the performance of the proposed algorithm is evaluated. The simulation parameters of variables for RIS-aided HAP-NOMA system are shown in Table 2 based on [1].

As shown in Fig. 2, we analyze the trends of the sum SLNR of the users $\sum_{g=1}^{G} \sum_{k=1}^{|\mathcal{K}_g|} \text{SLNR}_{g_k}$ with respect to the total transmitted power P. In this simulation, we set $M = 256$. Figure 2 proves the effectiveness of the closed-form expressions which we deduced for the upper bound on SLNR in (7) for the RIS-aided HAP-NOMA systems. The Monte-carlo simulations depicts the $\sum_{g=1}^{G} \sum_{k=1}^{|\mathcal{K}_g|} \text{SLNR}_{g_k}^u$ with respect to P, where $\text{SLNR}_{g_k}^u$ is

Table 2 Simulation parameters.

Variables	Simulation Parameters
Location of HAP	$(x_H, y_H, z_H) = (0, 0, 2 \times 10^4)$
Location of RIS	$(x_R, y_R, z_R) = (0, 5 \times 10^3, 40)$
Frequency	2.4GHz
Bandwidth	10MHz
Noise variance	$\sigma^2 = -80$dBm
Number of antennas on HAP	$M_v = M_h$
Number of antennas on RIS	$N_v = N_h = 8$
Number of users	$K = 20$
Number of no direct-link users	$\lvert\mathcal{K}'\rvert = 10$
Radius of circle, K users randomly distributed	2×10^4
Radius of circle, $\mathcal{K}'$ users randomly distributed	30
Variables in $(2d)$	$\kappa_{t,r} = 5$, $\mu_{t,r} = 0°$, $\theta'_{t.r} = 30°$, $\delta_{t,r} = 10°$
Antenna Gain	$G_{H,g_k} = G_{R,g_k} = 0.5$, $G_{HR} = 1.45$
Iterative number	$T = 20$
Preset threshold	$\xi_t = 10^{-6}$

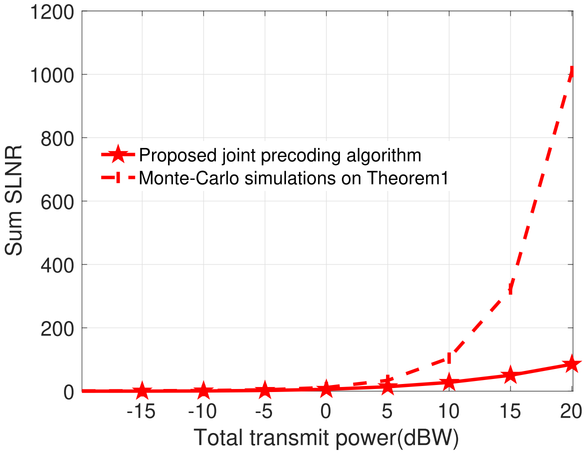


Fig. 2 Sum SLNR versus total transmit power. $M = 256$.

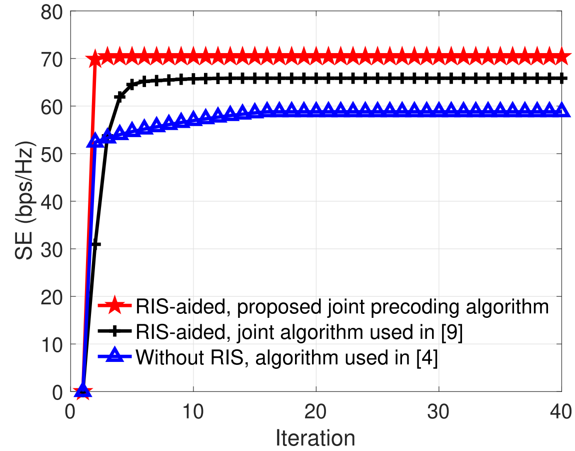


Fig. 3 SE versus iteration. $M = 256$ and $P = 20$ dBW.

defined in (7). And the proposed joint precoding algorithm depicts the $\sum_{g=1}^{G}\sum_{k=1}^{|\mathcal{K}_g|}\mathrm{SLNR}_{g_k}$ with respect to P, where SLNR_{g_k} is defined in (5).

As shown in Fig. 3, the SE performance of the proposed scheme has been evaluated. In this simulation, we set $M = 256$ and $P = 20$dBW. For comparison, we plot the SE achieved by joint algorithm in [9] and without RIS algorithm in [4]. The proposed algorithm outperforms the algorithm in [9] due to the performance loss of the utilization of first-order Taylor expansion in [9]. The worst performance of the algorithm in [4] is because the RIS have not been used. The convergence speed of the proposed algorithm is faster than the other two algorithms due to the closed-form expression on upper bound of SLNR.

Figure 4 shows the SE performance of the proposed scheme with respect to total transmitted power P and the number of antennas on HAP M. As we can observe, the SE obtained by all these algorithms increase with the growth of P and M. In terms of SE, the proposed algorithm with RIS outperforms the other two algorithms, owing to the same reason as mentioned.

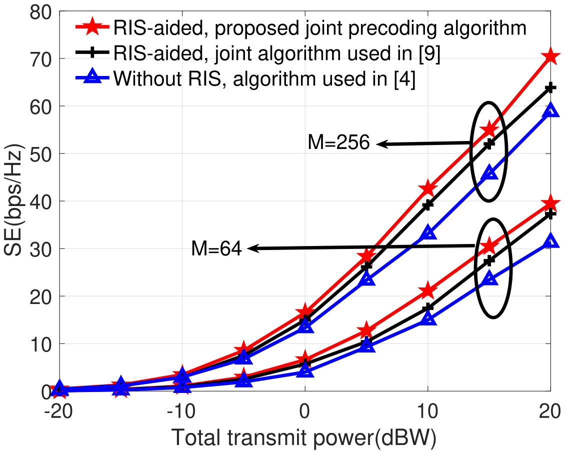


Fig. 4 SE versus total transmit power with respect to different number of antennas M on HAP.

5. Conclusion

In this paper, we have studied a downlink transmission algorithm for the RIS assisted beamspace HAP-NOMA systems according to SLNR. We have aimed to maximize the minimum SLNR of all users by jointly designing the passive precoding at the RIS and power allocation on HAP. The main contributions of this paper could be summarized as follows:

- A novel RIS-aided beamspace HAP-NOMA system has been proposed. The NOMA combining beamspace HAP has improved the achievable sum rate by improving the number of serving users simultaneously. The combination of RIS and beamspace HAP-NOMA has made the users without direct-link from HAP could achieve better service.
- The statistical upper bound on SLNR has been derived according to the random matrix theory in large scale antenna. SLNR has been used as performance measure for the RIS-aided HAP-NOMA system to reduce the computation complexity by decoupling the power allocation and passive precoding. The minimum SLNR has been maximized to consider the user fairness.
- The closed form expressions of power allocation matrix and passive precoding matrix have been obtained by introducing a series of auxiliary variables. To be specific, The Lagrange multiplier method has been used to solve the power allocation problem, and the bisection method has been used to solve the passive precoding problem.
- Numerical results have shown that the derived upper bound is effective and the proposed algorithm has an distinct performance enhancement.

References

[1] Z. Lian, Y. Su, Y. Wang, and L. Jiang, “A non-stationary 3-D wideband channel model for intelligent reflecting surface-assisted HAP-MIMO communication systems,” IEEE Trans. Veh. Technol., vol.71,

no.2, pp.1109–1123, Feb. 2022.

[2] M. Guan, Z. Wu, Y. Cui, X. Cao, L. Wang, J. Ye, and B. Peng, "Efficiency evaluations based on artificial intelligence for 5G massive MIMO communication systems on high-altitude platform stations," IEEE Trans. Ind. Informat., vol.16, no.10, pp.6632–6640, Oct. 2020.

[3] Z. Lian, L. Jiang, C. He, and D. He, "User grouping and beamforming for HAP massive MIMO systems based on statistical-eigenmode," IEEE Wireless Commun. Lett., vol.8, no.3, pp.961–964, June 2019.

[4] P. Ji, L. Jiang, C. He, Z. Lian, and D. He, "Energy-efficient beamforming for beamspace HAP-NOMA systems," IEEE Commun. Lett., vol.25, no.5, pp.1678–1681, May 2021.

[5] N. Gao, S. Jin, X. Li, and M. Matthaiou, "Aerial RIS-assisted high altitude platform communications," IEEE Wireless Commun, Lett., vol.10, no.10, pp.2096–2100, Oct. 2021.

[6] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," IEEE J. Sel. Areas Commun., vol.35, no.10, pp.2370–2382, Oct. 2017.

[7] J. Wang, H. Wang, Y. Han, S. Jin, and X. Li, "Joint transmit beamforming and phase shift design for reconfigurable intelligent surface assisted MIMO systems," IEEE Trans. Cogn. Commun. Netw., vol.7, no.2, pp.354–368, June 2021.

[8] X. Ma, S. Guo, H. Zhang, Y. Fang, and D. Yuan, "Joint beamforming and reflecting design in reconfigurable intelligent surface-aided multi-user communication systems," IEEE Trans. Wireless Commun., vol.20, no.5, pp.3269–3283, May 2021.

[9] P. Liu, Y. Li, W. Cheng, X. Gao, and X. Huang, "Intelligent reflecting surface sided NOMA for millimeter-wave massive MIMO with lens antenna array," IEEE Trans. Veh. Technol., vol.70, no.5, pp.4419–4434, May 2021.

[10] H. Guo, Y. Liang, J. Chen, and E.G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," IEEE Trans. Wireless Commun., vol.19, no.5, pp.3064–3076, May 2020.

[11] C. Pradhan, A. Li, L. Song, B. Vucetic, and Y. Li, "Hybrid precoding design for reconfigurable intelligent surface aided mmWave communication systems," IEEE Wireless Commun. Lett., vol.9, no.7, pp.1041–1045, July 2020.

[12] L. Pang, W. Wu, Y. Zhang, Y. Yuan, Y. Chen, A. Wang, and J. Li, "Joint power allocation and hybrid beamforming for downlink mmWave-NOMA systems," IEEE Trans. Veh. Technol., vol.70, no.10, pp.10173–10184, Oct. 2021.

[13] S. Wagner, R. Couillet, M. Debbah, and D.T.M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," IEEE Trans. Inf. Theory, vol.58, no.7, pp.4509–4537, July 2012.

[14] B. Su, X. Ding, C. Liu, and Y. Wu, "Heteroscedastic max–min distance analysis for dimensionality reduction," IEEE Trans. Image Process., vol.27, no.8, pp.4052–4065, Aug. 2018.

[15] S. Boyd and L. Vandenberghe, Convex Optimication. Cambridge, Cambridge Univ. Press, U.K., 2004.

[16] C. Shen, H. Li, and M. Brooks, "Supervised dimensionality reduction via sequential semidefinite programming," Pattern Recognition, vol.41, no.12, pp.3644–3652, 2008.

[17] K.-Y. Wang, A.M.-C. So, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," IEEE Trans. Signal Process., vol.62, no.21, pp.5690–5705, Nov. 2014.

Appendix: Proof of Theorem 1

We derive Theorem 1 by setting some temporary variables $\mathbf{F}_{g_k} = (\sum_{j_i \in \mathcal{K}_{-g_k}} \vec{\mathbf{h}}_{j_i}^H \vec{\mathbf{h}}_{j_i} + \frac{\sigma^2}{p_{g_k}}\mathbf{I})^{-1}$, $\widetilde{\mathbf{w}}_g = \mathbf{F}_{g_1}\vec{\mathbf{h}}_{g_1}^H$ and

$$\upsilon_{g_k} = \frac{\widetilde{\mathbf{w}}_g^H \vec{\mathbf{h}}_{g_k}^H \vec{\mathbf{h}}_{g_k} \widetilde{\mathbf{w}}_g}{\widetilde{\mathbf{w}}_g^H (\sum_{j_i \in \mathcal{K}_{-g_{k+1}}} \vec{\mathbf{h}}_{j_i}^H \vec{\mathbf{h}}_{j_i} + \frac{\sigma^2}{p_{g_k}}\mathbf{I})\widetilde{\mathbf{w}}_g}. \tag{A$\cdot$ 1}$$

Let $\mathbf{g}_{g_k} = \frac{1}{\sqrt{M}}\mathbf{h}_{H,g_k}^w \sim \mathcal{CN}(\mathbf{0}, \frac{1}{M}\mathbf{I})$, $\mathbf{l}_{g_k} = \frac{1}{\sqrt{N}}\mathbf{h}_{R,g_k}^w \sim \mathcal{CN}(\mathbf{0}, \frac{1}{N}\mathbf{I})$, $\vec{\widehat{\mathbf{h}}}_{g_k} = \sqrt{\alpha_{H,g_k}\rho_{H,g_k}}\mathbf{a}(\vartheta_{H,g_k}, \varphi_{H,g_k}, M) + \sqrt{\alpha_{R,g_k}\rho_{R,g_k}}\mathbf{a}(\vartheta_{R,g_k}, \varphi_{R,g_k}, N)\mathbf{\Phi}\mathbf{h}_{HR}$ and $\vec{\widetilde{\mathbf{h}}}_{g_k} = \sqrt{\alpha_{H,g_k}(1-\rho_{H,g_k})}\mathbf{h}_{H,g_k}^w \widetilde{\mathbf{R}}_{H,g_k}^{1/2} + \sqrt{\alpha_{R,g_k}(1-\rho_{R,g_k})}\mathbf{h}_{R,g_k}^w \widetilde{\mathbf{R}}_{R,g_k}^{1/2}\mathbf{\Phi}\mathbf{h}_{HR}$. We use (2) to get

$$\begin{aligned}
&\vec{\mathbf{h}}_{g_k}\mathbf{F}_{g_k}\vec{\mathbf{h}}_{g_k}^H \\
&= \vec{\widehat{\mathbf{h}}}_{g_k}\mathbf{F}_{g_k}\vec{\widehat{\mathbf{h}}}_{g_k}^H + \vec{\widehat{\mathbf{h}}}_{g_k}\mathbf{F}_{g_k}\vec{\widetilde{\mathbf{h}}}_{g_k}^H + \vec{\widetilde{\mathbf{h}}}_{g_k}\mathbf{F}_{g_k}\vec{\widehat{\mathbf{h}}}_{g_k}^H + \vec{\widetilde{\mathbf{h}}}_{g_k}\mathbf{F}_{g_k}\vec{\widetilde{\mathbf{h}}}_{g_k}^H \\
&\overset{(a)}{=} \vec{\widehat{\mathbf{h}}}_{g_k}\mathbf{F}_{g_k}\vec{\widehat{\mathbf{h}}}_{g_k}^H \\
&+M(\alpha_{H,g_k}(1-\rho_{H,g_k}))\mathbf{g}_{g_k}\widetilde{\mathbf{R}}_{H,g_k}^{1/2}\mathbf{U}_s\mathbf{F}_{g_k}\mathbf{U}_s^H\widetilde{\mathbf{R}}_{H,g_k}^{1/2}\mathbf{g}_{g_k}^H \\
&+N\mathbf{l}_{g_k}\mathbf{\Xi}_{g_k}^{\frac{1}{2}}\boldsymbol{\phi}^H\mathbf{a}(\vartheta_{HR}, \varphi_{HR}, M)\mathbf{U}_s\mathbf{F}_{g_k}\mathbf{U}_s^H \\
&\cdot\mathbf{b}(\vartheta_{HR}, \varphi_{HR}, M)^H\boldsymbol{\phi}\mathbf{\Xi}_{g_k}^{\frac{1}{2}}\mathbf{l}_{g_k}^H \\
&\overset{(b)}{=} \mathrm{Tr}(\vec{\widehat{\mathbf{h}}}_{g_k}^H\vec{\widehat{\mathbf{h}}}_{g_k}\mathbf{F}_{g_k} \\
&+(\alpha_{H,g_k}(1-\rho_{H,g_k}))\mathbf{U}_s^H\widetilde{\mathbf{R}}_{H,g_k}\mathbf{U}_s\mathbf{F}_{g_k} \\
&+\Xi_{g_k}\mathbf{U}_s^H\mathbf{a}(\vartheta_{HR}, \varphi_{HR}, M)^H\mathbf{a}(\vartheta_{HR}, \varphi_{HR}, M)\mathbf{U}_s\mathbf{F}_{g_k}) \\
&\overset{(c)}{=} \mathrm{Tr}(\vec{\mathbf{R}}_{g_k,g_k}\mathbf{F}_{g_k})
\end{aligned} \tag{A$\cdot$ 2}$$

where (a) follows from the Lemma 5 in [13], (b) follows from the Lemma 4 in [13] and (c) follows from the Theorem 1 in [4] and

$$\begin{aligned}
\vec{\mathbf{R}}_{g_k,g_k} = M(&\sqrt{\alpha_{H,g_k}\rho_{H,g_k}}\boldsymbol{\lambda}_G^g + c_{g_k}\boldsymbol{\lambda}_G^G)^H \\
&\cdot(\sqrt{\alpha_{H,g_k}\rho_{H,g_k}}\boldsymbol{\lambda}_G^g + c_{g_k}\boldsymbol{\lambda}_G^G) \\
&\quad +M\Xi_{g_k}\mathbf{\Lambda}_M^{(G,G)}.
\end{aligned} \tag{A$\cdot$ 3}$$

We utilize $(A.1)$, $(A.2)$ and $(A.3)$ to get SLNR_{g_k} as

$$\begin{aligned}
&\mathrm{SLNR}_{g_k} \\
&= \frac{\vec{\mathbf{h}}_{g_1}\mathbf{F}_{g_1}\vec{\mathbf{h}}_{g_k}^H\vec{\mathbf{h}}_{g_k}\mathbf{F}_{g_1}\vec{\mathbf{h}}_{g_1}^H}{(\sum_{j_i \in \mathcal{K}_{-g_k}}\vec{\mathbf{h}}_{g_1}\mathbf{F}_{g_1}\vec{\mathbf{h}}_{j_i}^H\vec{\mathbf{h}}_{j_i}\mathbf{W}_{g_1}\vec{\mathbf{h}}_{g_1}^H + \frac{\sigma^2}{p_{g_k}}\vec{\mathbf{h}}_{g_1}\mathbf{F}_{g_1}^2\vec{\mathbf{h}}_{g_1}^H)} \\
&= \frac{\vec{\mathbf{h}}_{g_1}\mathbf{F}_{g_1}\vec{\mathbf{h}}_{g_k}^H\vec{\mathbf{h}}_{g_k}\mathbf{F}_{g_1}\vec{\mathbf{h}}_{g_1}^H}{\vec{\mathbf{h}}_{g_1}\mathbf{F}_{g_1}(\sum_{j_i \in \mathcal{K}_{-g_{k+1}}}\vec{\mathbf{h}}_{j_i}^H\vec{\mathbf{h}}_{j_i} - \vec{\mathbf{h}}_{g_k}^H\vec{\mathbf{h}}_{g_k} + \frac{\sigma^2}{p_{g_k}}\mathbf{I})\mathbf{F}_{g_1}\vec{\mathbf{h}}_{g_1}^H} \\
&= \upsilon_{g_k}/(1-\upsilon_{g_k}) \\
&\overset{(d)}{\leq} \vec{\mathbf{h}}_{g_k}\mathbf{F}_{g_k}\vec{\mathbf{h}}_{g_k}^H \\
&\overset{(e)}{=} \delta'_{g,G}\mathrm{Tr}(\vec{\mathbf{R}}_{g_k,g_k}\mathbf{F}_{g_1}) \\
&+\delta_{g,G}\mathrm{Tr}(\vec{\mathbf{R}}_{g_k,g_k}(\textstyle\sum_{j_i \in \mathcal{K}}\vec{\mathbf{h}}_{j_i}^H\vec{\mathbf{h}}_{j_i} + \frac{\sigma^2}{p_{g_k}}\mathbf{I})^{-1}) \\
&\overset{(f)}{=} \delta'_{g,G} \\
&\mathrm{Tr}(\vec{\mathbf{R}}_{g_k,g_k}(\sum_{j_i \in \mathcal{K}_{-g_1}}(\tfrac{M\Xi_{j_i}\mathbf{\Lambda}_G^{(G,G)}}{1+m_{j_i}} + \vec{\widehat{\mathbf{h}}}_{j_i}\vec{\widehat{\mathbf{h}}}_{j_i}^H) + \frac{\sigma^2}{p_{g_k}}\mathbf{I})^{-1}) \\
&+\delta_{g,G}\mathrm{Tr}(\vec{\mathbf{R}}_{g_k,g_k}(\sum_{j_i \in \mathcal{K}}(\tfrac{M\Xi_{j_i}\mathbf{\Lambda}_G^{(G,G)}}{1+m_{j_i}} + \vec{\widehat{\mathbf{h}}}_{j_i}\vec{\widehat{\mathbf{h}}}_{j_i}^H) + \frac{\sigma^2}{p_{g_k}}\mathbf{I})^{-1}) \\
&\overset{(g)}{=} \frac{\delta'_{g,G}M\alpha_{H,g_k}\rho_{H,g_k}}{\frac{\sigma^2}{p_{g_k}}} \\
&+\frac{\delta_{g,G}(|\sqrt{\alpha_{H,g_k}\rho_{H,g_k}}+c_{g_k}|^2+\Xi_{g_k})+\delta'_{g,G}(c_{g_k}^H c_{g_k}+\Xi_{g_k})}{\sum\limits_{j_i \in \mathcal{K}_{-g_1}}(\frac{\Xi_{j_i}}{1+m_{j_i}}+\delta'_{j,G}c_{j_i}^H c_{j_i})+\eta+\frac{\sigma^2}{Mp_{g_k}}-\zeta_g}
\end{aligned} \tag{A$\cdot$ 4}$$

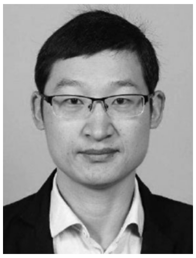
where (d) come from the monotonically increasing function SLNR_{g_k} with respect to $\upsilon_{g_k} \in [0,1)$. Obviously, the optimal $\widetilde{\mathbf{w}}_g^{(opt)}$ can be obtained as $(\sum_{j_i \in \mathcal{K}_{-g_{k+1}}} \vec{\mathbf{h}}_{j_i}^H \vec{\mathbf{h}}_{j_i} + \frac{\sigma^2}{p_{g_k}}\mathbf{I})^{-1} \vec{\mathbf{h}}_{g_k}^H$ to maximize SLNR_{g_k}. We utilize the matrix inversion lemma in (25) and attain the maximum value of $\mathrm{SLNR}_{g_k} = \vec{\mathbf{h}}_{g_k} \mathbf{F}_{g_k} \vec{\mathbf{h}}_{g_k}^H$. And, (e) follows from the rank-1 perturbation lemma in [13], (f) follows the Theorem 1 in [13], (g) follows the Theorem 1 in [4].

Pingping Ji received the B.E. degree in communication engineering and the M.E. degree in communication and information system from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree with the School of Shanghai Jiao Tong University, Shanghai, China. Her research interests include multiple-in-multiple-out (MIMO) technology, reconfigurable intelligent surface (RIS), and precoding for massive MIMO.

Lingge Jiang received the B.E. degree in radio engineering from Southeast University, Nanjing, China, in 1982, and the M.E. degree in electrical engineering, and the Ph.D. degree in electronic system engineering from Tokushima University, Tokushima, Japan, in 1993 and 1996, respectively. In 1996, she joined the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. She is currently a Professor with Shanghai Jiao Tong University. She has authored or coauthored more than 170 journal and conference papers and has 59 awarded patents. Her current research interests include next generation wireless communication systems, wireless sensor networks, cognitive radio networks, cooperative communication, and intelligent information processing. She was the recipient of the 4th Nyoji Yokoyam Excellent Paper Award of Shanghai Jiao Tong University in 2000 and the Technique Invention Prize (second-class) of Shanghai Science and Technology Award in 2008 (20083041-2-R02).

Chen He received the B.E. and the M.E. degrees in electronic engineering from Southeast University of China, Nanjing, China, in 1982 and 1985 respectively, and the Ph.D. degree in electronics system from Tokushima University of Japan, Tokushima, Japan, in 1994. He was with the Department of Electronic Engineering, Southeast University of China, from 1985 to 1990. He joined the Department of Electronic Engineering, Shanghai Jiao Tong University of China, in 1996. He visited Tokushima University of Japan as a foreign researcher from October 1990 to September 1991 and visited the Communication Research Laboratory of Japan from December 1999 to December 2000 as a Research Fellow. He is currently a Professor with Shanghai Jiao Tong University, Shanghai, China. He has authored or coauthored more than 200 journal papers and more than 80 conference papers. His research activities focus on 5G wireless communication systems, including interference cancellation for multi-cells, wireless resource management, and cognitive radio. He received the Best Paper Award in the GLOBECOM 2007.

Di He Di He received the B.E. degree in information engineering from Huazhong University of Science and Technology, Wuhan, China, in 1996, the M.E. degree in communication and information engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 1999, and the Ph.D. degree in circuits and systems from Shanghai Jiao Tong University, Shanghai, China, in 2002. From 2002 to 2004, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Calgary, Canada. He is currently an Associate Professor with Shanghai Jiao Tong University. His research interests include wireless communications and wireless positioning.



Zhuxian Lian received the B.E. degree in communication engineering from Information Engineering University, Zhengzhou, China, in 2011, and the Ph.D. degree from Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai, China, in 2019. He is currently a Lecturer with the Jiangsu University of Science and Technology, Zhenjiang, China. His research interests include channel modeling, MIMO techniques, and precoding for massive MIMO.

PAPER

Throughput Maximization-Based AP Clustering Methods in Downlink Cell-Free MIMO Under Partial CSI Condition*

Daisuke ISHII†a), *Nonmember*, Takanori HARA†, *Member*, *and* Kenichi HIGUCHI†, *Senior Member*

SUMMARY In this paper, we investigate a method for clustering user equipment (UE)-specific transmission access points (APs) in downlink cell-free multiple-input multiple-output (MIMO) assuming that the APs distributed over the system coverage know only part of the instantaneous channel state information (CSI). As a beamforming (BF) method based on partial CSI, we use a layered partially non-orthogonal zero-forcing (ZF) method based on channel matrix muting, which is applicable to the case where different transmitting AP groups are selected for each UE under partial CSI conditions. We propose two AP clustering methods. Both proposed methods first tentatively determine the transmitting APs independently for each UE and then iteratively update the transmitting APs for each UE based on the estimated throughput considering the interference among the UEs. One of the two proposed methods introduces a UE cluster for each UE into the iterative updates of the transmitting APs to balance throughput performance and scalability. Computer simulations show that the proposed methods achieve higher geometric-mean and worst user throughput than those for the conventional methods.

key words: *MIMO, cell-free MIMO, inter-base station cooperation, transmitter selection, beamforming, partial CSI*

1. Introduction

Cooperative multiple-input multiple-output (MIMO) [1]–[4], which employs MIMO transmission to multiple sets of user equipment (UEs) in coordination among base stations (BSs), eliminates the cell boundaries between coordinated BSs. This approach increases the throughput by utilizing an increased number of transmitter and receiver antennas at the network level. Furthermore, distributed antenna systems (DASs) [5], [6], in which BS antennas are distributed throughout the system coverage area, are effective in reducing the area of insensitivity, reducing the required transmission power, and increasing the total throughput compared to when many antennas are locally deployed. In this context, cell-free MIMO [7]–[9], which is one type of DAS and distributes antennas, such as access points (APs), over the system coverage, has recently been actively researched and developed for application to the 5th generation mobile communication system new radio (NR) and later systems [10]–[12].

Beamforming (BF) is essential to achieve high transmission capacity in MIMO transmission. It takes advantage of the high spatial degrees of freedom of a MIMO channel to increase the received signal power and to suppress interference between spatially multiplexed UEs. A sophisticated BF scheme requires instantaneous channel state information (CSI), which corresponds to the instantaneous complex channel coefficients, at the AP.

However, it is impractical for each AP to obtain ideally the instantaneous CSI for all UEs. For example, in a time division duplex (TDD) system, although the AP estimates the instantaneous CSI using the uplink reference signal, such a CSI estimate for UEs that are far from the AP tends to be inaccurate. Moreover, the CSI estimation and the calculation of sophisticated BF vectors for all UEs require a high level of computational complexity when cell-free MIMO supports a large number of UEs.

To overcome these issues, our research group previously reported a partially non-orthogonal zero-forcing (ZF) method for MIMO with partial CSI knowledge [13]. This method designs BF vectors based on a muting method that inserts zeros into small channel coefficients. The BF enables spatial multiplexing of a group of UEs with different instantaneous CSI knowledge where each AP knows only some instantaneous CSI. Since the degree of freedom of the MIMO channel is used effectively, this method increases user throughput compared to the conventional method [14], [15] that allocates different orthogonal channels to UEs with different instantaneous CSI knowledge.

In addition to the BF design with partial CSI knowledge, approaches to enhance the scalability of cell-free MIMO have been investigated [9], [16], [17]. Although there are many approaches, this paper focuses on a method that determines the transmitting AP groups, which is also known as *clustering*[**]. Methods for AP clustering each UE independently from other UEs have been proposed [9], [18]–[21]. In [19], an AP clustering method was proposed that determines the transmitting AP groups while taking into account the received signal power using the partially non-orthogonal ZF method as a BF method. Moreover, in [20] and [21], this method was extended to consider the interference to other UEs. In contrast, the methods proposed in [9] and [18] rely on path loss to determine the transmitting AP groups. However, independent determination of the

Manuscript received December 8, 2023.
Manuscript revised February 28, 2024.
Manuscript publicized May 6, 2024.

†Graduate School of Science and Technology, Tokyo University of Science, Noda-shi, 278-8510 Japan.

[*] The material in this paper was presented in part at the 97th IEEE Vehicular Technology Conference (VTC2023-Spring), 2023.

a) E-mail: 7322510@ed.tus.ac.jp

DOI: 10.23919/transcom.2023EBT0008

[**] To reduce the required computational complexity for the BF, a local BF method in which each AP individually computes the BF vector has been investigated [16]. Besides, the authors of [17] have proposed the method to categorize APs into the ones using a sophisticated BF, such as ZF, and the others using a simple BF.

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

transmitting APs for a specific UE induces interference to other UEs, leading to degradation in the system throughput. Therefore, an alternative AP clustering method that takes into account the effects among UEs is desirable to enhance further the performance of downlink cell-free MIMO with partial instantaneous CSI knowledge. In addition to improving the system throughput, the scalability of the AP clustering method is important for accommodating a system with numerous UEs and APs.

In this paper, we propose two AP clustering methods for each UE in downlink cell-free MIMO based on a layered partially non-orthogonal ZF method. Both proposed methods first tentatively determine the transmitting APs independently for each UE and then iteratively update the transmitting APs for each UE using the geometric-mean throughput as a metric. These steps allow the proposed methods to determine the APs for each UE considering the mutual effects among UEs. Moreover, one of the two proposed methods focuses on balancing throughput performance and scalability by introducing a UE cluster into the determination of transmitting APs for each UE. Computer simulations show that the proposed methods achieve higher throughput than that for the conventional methods.

The rest of this paper is organized as follows. Section 2 describes the model of cell-free MIMO with partial CSI knowledge and the layered partially non-orthogonal ZF method. Subsequently, the conventional UE-independent AP clustering methods are reviewed in Sect. 3. Section 4 presents the proposed AP clustering methods. Section 5 shows the simulation results and Sect. 6 concludes the paper.

2. System Model

2.1 Cell-Free MIMO Model

We consider a downlink cell-free MIMO system that comprises APs with N transmitter antennas and single-antenna UEs. Let $\mathcal{L}$ and $\mathcal{K}$ be the set of APs and UEs in the system coverage, respectively. The set of APs that are candidates for transmitting APs to UE k is denoted by $\mathcal{L}_k \subseteq \mathcal{L}$. Throughout the paper, we set the number of APs in $\mathcal{L}_k$ to T for all UEs, and $\mathcal{L}_k$ comprises APs with the T lowest path loss between it and UE $k \in \mathcal{K}$.

In this paper, the instantaneous CSI is assumed to be estimated based on the uplink reference signal in a TDD system. To consider the partial CSI condition, we define $\mathcal{S}_k \subseteq \mathcal{L}_k$ as the set of APs that know the instantaneous CSI of UE k, as shown in Fig. 1. Set $\mathcal{S}_k$ comprises APs whose path loss between it and UE k is less than $G_{k,\min} + \Delta_{\text{CSI}}$ dB, where $G_{k,\min}$ is the minimum path loss between UE k and the APs and Δ_{CSI} is a parameter that represents the ease of ascertaining the instantaneous CSI. Namely, $\mathcal{S}_k = \{l \mid \beta_{k,l} \leq G_{k,\min} + \Delta_{\text{CSI}}\}$ with $\beta_{k,l}$ denoting the path loss between UE k and AP l.

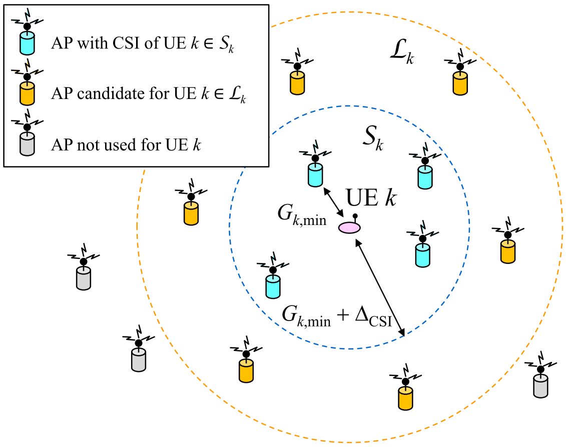

AP with CSI of UE $k \in \mathcal{S}_k$
AP candidate for UE $k \in \mathcal{L}_k$
AP not used for UE k
$\mathcal{L}_k$
$\mathcal{S}_k$
UE k
$G_{k,\min}$
$G_{k,\min} + \Delta_{\text{CSI}}$

Fig. 1 System model.

2.2 Layered Partially Non-Orthogonal ZF Method

In this subsection, we review the layered partially non-orthogonal ZF method [19], which is the BF method used in this paper. Although this method utilizes the block diagonalization (BD) method [22] to encompass the case where a UE has multiple antennas, BD is replaced by ZF in the description since this paper assumes that all UEs have a single antenna.

Let $\mathcal{G}(\mathcal{S}_k)$ be the set of UEs other than UE k for which at least one of AP $l \in \mathcal{S}_k$ knows the instantaneous CSI. Thus, $\mathcal{G}(\mathcal{S}_k) = \{i\colon \mathcal{S}_i \cap \mathcal{S}_k \neq \varnothing, i \neq k \in \mathcal{K}\}$. Moreover, the $1 \times N$-dimensional channel matrix between AP l and UE k is denoted by $\mathbf{H}_{k,l}$. Note that AP $l \notin \mathcal{S}_k$ does not know $\mathbf{H}_{k,l}$ due to the partial CSI condition.

Let us consider BF used for data transmission to UE k. As a baseline of the layered BD method, it is reasonable that the transmitting APs for UE k are set to $[\mathcal{S}_k]_1, \ldots, [\mathcal{S}_k]_{|\mathcal{S}_k|}$ where $[\mathcal{S}_k]_l$ represents the l-th element of $\mathcal{S}_k$. Term $\mathbf{H}_{\mathcal{S}_k}$ denotes the $|\mathcal{G}(\mathcal{S}_k)| \times |\mathcal{S}_k|N$-dimensional channel matrix between the AP set $\mathcal{S}_k$ and the UEs in $\mathcal{G}(\mathcal{S}_k)$. If all the elements of $\mathcal{G}(\mathcal{S}_k)$ are known, the BF vector to UE k can be obtained from the Moore-Penrose generalized inverse of $\mathbf{H}_{\mathcal{S}_k}$ based on the principle of the ZF method. However, due to the partial CSI condition, $\mathbf{H}_{\mathcal{S}_k}$ may contain unknown elements. Therefore, we use a method to determine the BF vector based on a muted channel matrix, e.g., as in [13]. The muting method [13] first performs a muting operation that sets all $\mathbf{H}_{i,l}$, i.e., $i \in \mathcal{G}(\mathcal{S}_k)$, $l \notin \mathcal{S}_i$, to zero matrix $\mathbf{O}$, since $\mathbf{H}_{i,l}$ ($i \in \mathcal{G}(\mathcal{S}_k)$, $l \notin \mathcal{S}_i$) is not known at the AP. Term $\tilde{\mathbf{H}}_{\mathcal{S}_k}$ is the $|\mathcal{G}(\mathcal{S}_k)| \times |\mathcal{S}_k|N$-dimensional matrix obtained after applying the muting operation to $\mathbf{H}_{\mathcal{S}_k}$. The BF vector to UE k can be obtained from the Moore-Penrose generalized inverse of $\tilde{\mathbf{H}}_{\mathcal{S}_k}$. The idea of the muting method is also used in the study of cell-free MIMO in [18].

In this case, the data transmitted by AP $[\mathcal{S}_k]_l$ to UE k do not interfere with other UEs for which AP $[\mathcal{S}_k]_l$ knows the instantaneous CSI. On the other hand, the transmitted data to UE k interfere with UEs whose instantaneous CSI is unknown to AP $[\mathcal{S}_k]_l$, resulting in partially non-orthogonal ZF. However, since the path loss of channels with unknown instantaneous CSI is high, the interference power is expected to be largely suppressed by the channel.

Given $\mathcal{S}_k$, $F(\mathcal{S}_k)$ is defined as

$$F(\mathcal{S}_k) = |\mathcal{S}_k| N - |\mathcal{G}(\mathcal{S}_k)|. \tag{1}$$

This indicates the order of the received signal power gain that UE k can obtain when BF based on ZF is performed. The non-positive value of $F(\mathcal{S}_k)$ implies that all the degrees of freedom of the MIMO channel are used for interference suppression. Thus, for data transmission to UE k using AP group $\mathcal{S}_k$, $F(\mathcal{S}_k)$ must be greater than one. If $F(\mathcal{S}_k)$ is less than or close to one, the layered ZF method considers adding to the transmitting APs for UE k to increase the received signal power gain. Since the instantaneous CSI between the added APs and UE k is unknown, the use of additional APs does not directly contribute to the transmission quality of UE k. However, the growth in the number of transmitting APs for UE k increases the channel degrees of freedom that can be used to null out interference in other UEs. As a result, an increase in the received signal power gain can be expected.

Let $\mathcal{D}_k$ and $\mathbf{H}_{\mathcal{S}_k \cup \mathcal{D}_k}$ be the AP group used for data transmission to UE k together with APs in $\mathcal{S}_k$ ($\mathcal{S}_k \cap \mathcal{D}_k = \varnothing$) and let $\mathbf{H}_{\mathcal{S}_k \cup \mathcal{D}_k}$ be the $|\mathcal{G}(\mathcal{S}_k \cup \mathcal{D}_k)| \times |\mathcal{S}_k \cup \mathcal{D}_k| N$-dimensional channel matrix between AP $[\mathcal{S}_k \cup \mathcal{D}_k]_1, \ldots, [\mathcal{S}_k \cup \mathcal{D}_k]_{|\mathcal{S}_k \cup \mathcal{D}_k|}$ and the UEs in $\mathcal{G}(\mathcal{S}_k \cup \mathcal{D}_k)$, respectively. Matrix $\tilde{\mathbf{H}}_{\mathcal{S}_k \cup \mathcal{D}_k}$ is obtained by muting operation on $\mathbf{H}_{\mathcal{S}_k \cup \mathcal{D}_k}$ as described above. Provided that $F(\mathcal{S}_k \cup \mathcal{D}_k)$ is greater than or equal to one, the BF vector to UE k can be obtained from the Moore-Penrose generalized inverse matrix of $\tilde{\mathbf{H}}_{\mathcal{S}_k \cup \mathcal{D}_k}$.

The $|\mathcal{J}_k| N$-dimensional BF vector for UE k, whose norm is normalized, is denoted by $\mathbf{b}_k(\mathcal{J}_k)$ with $\mathcal{J}_k$ denoting the transmitting AP group used for transmission to UE k. Let $\tilde{\mathbf{H}}_k(\mathcal{J}_k)$ be a $1 \times |\mathcal{J}_k| N$-dimensional muting channel matrix where the unknown term of instantaneous CSI between UE k and AP group $\mathcal{J}_k$ is set to zero. Then, the effective channel gain for UE k is expressed as

$$\lambda_k(\mathcal{J}_k) = \|\tilde{\mathbf{H}}_k(\mathcal{J}_k) \mathbf{b}(\mathcal{J}_k)\|^2. \tag{2}$$

3. Conventional AP Clustering Methods

This section describes the conventional AP clustering methods proposed in [18] and [21], which will be evaluated in Sect. 5 along with the proposed methods. In this paper, we refer to them as the *path-loss based method* and the *signal-to-leakage-and-noise ratio (SLNR)-based method*, respectively.

Let $\mathcal{J}_k$ be the set of APs used for data transmission to UE $k \in \mathcal{K}$. The path-loss based method [18] independently determines $\mathcal{J}_k$ by selecting APs with the U lowest path loss between it and UE $k \in \mathcal{K}$ from the candidate APs. We note that the candidate APs comprise the APs in $\mathcal{L}_k$, unlike in [18]. Although the process of the path-loss based method is obviously simple, it determines a group of APs for transmission to each UE without considering the mutual effects among other UEs.

On the other hand, the SLNR-based method [21] determines $\mathcal{J}_k$ based on maximization of the metric below.

$$\mathcal{J}_k = \arg \max_{\mathcal{S}_k \cup \mathcal{D}_k} \frac{\lambda_k(\mathcal{S}_k \cup \mathcal{D}_k)}{z_k(\mathcal{S}_k \cup \mathcal{D}_k)}, \tag{3}$$

where $z_k(\mathcal{S}_k \cup \mathcal{D}_k)$ is the sum of the estimated interference power that the data transmitted from AP group $\mathcal{S}_k \cup \mathcal{D}_k$ to UE k gives to all other UEs and is given below.

$$z_k(\mathcal{S}_k \cup \mathcal{D}_k) = \sum_{i \neq k} \sum_{l \in (\mathcal{S}_k \cup \mathcal{D}_k) \setminus \mathcal{S}_i} \beta_{i,l} \frac{p_k}{|\mathcal{S}_k \cup \mathcal{D}_k|}, \tag{4}$$

where p_k is the transmission power of the data transmitted to UE k. The estimated interference power is calculated using the path loss under the assumption that the selected APs transmit data to the UE using the identical transmission power. The numerator of the metric in (3) is calculated using (2) and corresponds to the desired received signal power of UE k. The SLNR-based method is expected to improve the throughput at the system level by taking into account the amount of interference that the selected AP group may cause to other UEs. However, this method does not attempt to maximize the system throughput directly, rather it maximizes the SLNR for each UE individually.

In light of the above, the conventional AP clustering methods do not fully consider the mutual effect of the AP clustering for one UE on the other UEs. Furthermore, these methods do not select the transmitting AP groups based on the achievable throughput.

4. Proposed AP Clustering Methods

In this paper, we propose two AP clustering methods to improve the geometric-mean user throughput over the entire system coverage area. The first one achieves the highest throughput since it considers the mutual effects among all UEs. The second one considers scalability in addition to the achievable throughput. Each of the proposed methods is described below.

4.1 Iterative Update Method

We propose an AP clustering method that determines the transmitting AP group for each UE based on an iterative algorithm. Hereafter, this method is referred to as the *iterative update method*. We note that this method was originally proposed in [23], in which the estimated values of the average user throughput and worst user throughput that is defined by the minimum throughput among all UEs are utilized as a metric. In this paper, the geometric-mean throughput is used as a metric to guarantee user fairness. The iterative update method considers the mutual effects among all UEs by estimating the throughput. The flow of the iterative update method is described as follows.

Let $\mathcal{J}_k^{(r)}$ be the transmitting AP group for UE k in the r-th iteration. The iterative update method initializes $\mathcal{J}_k^{(r=0)}$ using the SLNR-based method as

$$\mathcal{J}_k^{(0)} = \arg\max_{\mathcal{S}_k\cup\mathcal{D}_k} \frac{\lambda_k(\mathcal{S}_k\cup\mathcal{D}_k)}{z_k(\mathcal{S}_k\cup\mathcal{D}_k)}. \tag{5}$$

After determining $\mathcal{J}_k^{(0)}$, the transmitting AP groups for all UEs are iteratively updated using the AP groups obtained in the previous iteration. Specifically, in the r-th iteration, transmitting AP groups $\mathcal{J}_k^{(r)}$ for each UE k are updated based on transmitting AP groups $\{\mathcal{J}_i^{(r-1)}\}$ for all other UEs obtained in the $r-1$-th iteration. The updates are performed sequentially for each UE as follows.

$$\mathcal{J}_k^{(r)} = \arg\max_{\mathcal{J}_k}\left(\prod\nolimits_{i\in\mathcal{K}} C_i^{(r)}\right)^{\frac{1}{|\mathcal{K}|}}, \tag{6}$$

where $C_k^{(r)}$ is the estimated throughput of UE k in the AP clustering process for UE k in the r-th iteration and is calculated by

$$C_k^{(r)} = \log_2\left(1 + \frac{\lambda_k\left(\mathcal{J}_k^{(r)}\right)p_k}{\sum\nolimits_{i=1}^{k-1} w_{i,k}\left(\mathcal{J}_i^{(r)}\right) + \sum\nolimits_{i=k+1}^{|\mathcal{K}|} w_{i,k}\left(\mathcal{J}_i^{(r-1)}\right) + N_0}\right), \tag{7}$$

where N_0 denotes the noise power and

$$w_{i,k}\left(\mathcal{J}_i^{(r)}\right) = \sum_{l\in(\mathcal{S}_i\cup\mathcal{D}_i)\backslash\mathcal{S}_k} \beta_{k,l}\frac{p_i}{|\mathcal{S}_i\cup\mathcal{D}_i|}. \tag{8}$$

Note that $w_{i,k}(\mathcal{J}_i^{(r)})$ implies the estimated interference power, which is calculated in the same manner as (4), to UE k when the transmitting AP group for UE i is $\mathcal{J}_i^{(r)}$.

Without loss of generality, it is assumed that the updates are performed in the order of UE number $1,\ldots,|\mathcal{K}|$ as shown in Fig. 2. The iterative update method calculates (7) for all patterns of the transmitting AP group for each UE. Therefore, its computational complexity depends on $|\mathcal{K}|$ and the number of candidates for the AP group, which is determined by T.

The above iterative process is repeated until r reaches R, which is the maximum number of iterations.

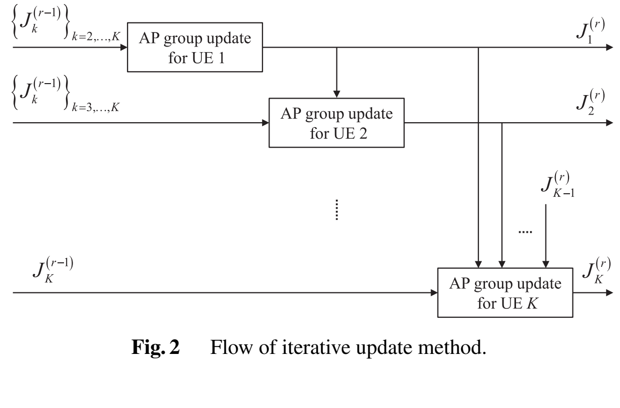


Fig. 2 Flow of iterative update method.

4.2 Scalable Iterative Update Method

The iterative update method fully considers the effect from the determination of the transmitting AP group for other UEs. Meanwhile, to calculate the metric in (6), the iterative update method needs to calculate the effective channel gain, which requires generating the BF vector for UE k using the set of all other UEs whose instantaneous CSI is known at the AP group for UE k. The method also necessitates calculating the estimated interference power from all UEs in the system coverage area, leading to the difficulty in achieving scalability. Moreover, since the iterative update method updates the AP group in order for each UE, the processing time per iteration depends on the number of UEs in the system coverage area. Hence, we also propose the *scalable iterative update method* to consider throughput and scalability.

The scalable iterative update method determines the UE cluster for each UE and updates the transmitting AP group using the metric based on UEs in the UE cluster. The UE cluster for UE k, which is defined as $\mathcal{Q}_k$, comprises the UEs that are expected to be nearby UE k. The scalable iterative update method reselects the transmitting AP group for each UE to maximize the metric calculated based on such a UE cluster. The flow of the scalable iterative update method is described as follows.

The scalable iterative update method is constituted by R iterations the same as in the iterative update method. This method generates a UE cluster for each UE and determines $\mathcal{J}'^{(0)}_k$ using the SLNR-based method based on UE cluster $\mathcal{Q}_k$. Cluster $\mathcal{Q}_k$ is constituted by any number of other UEs in order of decreasing path loss for UE k. The SLNR-based method using the UE cluster is determined as

$$\mathcal{J}'^{(0)}_k = \arg\max_{\mathcal{S}_k\cup\mathcal{D}_k} \frac{\lambda'_k(\mathcal{S}_k\cup\mathcal{D}_k)}{z'_k(\mathcal{S}_k\cup\mathcal{D}_k)}, \tag{9}$$

where $\lambda'_k(\mathcal{S}_k\cup\mathcal{D}_k)$ is the effective channel gain of UE k. It is calculated by (2) using the BF vector generated by considering only the UEs that are both in $\mathcal{Q}_k$ and $\mathcal{S}_k\cup\mathcal{D}_k$. Term $z'_k(\mathcal{S}_k\cup\mathcal{D}_k)$ is the sum of the estimated interference power that the data transmitted from AP group $\mathcal{S}_k\cup\mathcal{D}_k$ to UE k given to all UEs in $\mathcal{Q}_k$ and is given by

$$z'_k(\mathcal{S}_k\cup\mathcal{D}_k) = \sum_{i\in\mathcal{Q}_k}\sum_{l\in(\mathcal{S}_k\cup\mathcal{D}_k)\backslash\mathcal{S}_i}\beta_{i,l}\frac{p_i}{|\mathcal{S}_k\cup\mathcal{D}_k|}. \tag{10}$$

After initialization, the transmitting APs for all UEs are iteratively updated. In the r-th iteration, transmitting APs $\mathcal{J}'^{(r)}_k$ for each UE k are updated based on the transmitting APs of all UEs obtained in the $r-1$-th iteration, such as $\{\mathcal{J}'^{(r-1)}_i\}$. The updates are performed independently for each UE as

$$\mathcal{J}'^{(r)}_k = \arg\max_{\mathcal{J}_k}\left(\prod\nolimits_{i\in\mathcal{Q}_k\cup\{k\}} C'^{(r)}_i\right)^{\frac{1}{|\mathcal{Q}_k|+1}}, \tag{11}$$

where $C'^{(r)}_k$ is the estimated throughput of UE k in the AP clustering process for UE k in the r-th iteration and is calculated by

$$C_k'^{(r)} = \log_2\left(1 + \frac{\lambda_k'\left(\mathcal{J}_k'^{(r)}\right)p_k}{\sum_{i\in\mathcal{Q}_k} w_{i,k}\left(\mathcal{J}_i'^{(r-1)}\right) + N_0}\right). \qquad (12)$$

The numerator in (12) is the interference power from other UEs that is calculated using $\{\mathcal{J}_k'^{(r-1)}\}$. This indicates that the calculation of (12), namely the estimated throughput, relies only on the results obtained in the previous iteration, unlike (7). Therefore, the scalable iterative method can update the transmitting AP groups of all UEs in parallel, as shown in Fig. 3.

4.3 Computational Complexity

In this subsection, we discuss the computational complexity of the two proposed AP clustering methods. Table 1 gives the computational complexity of the iterative update and scalable iterative update methods. Term $\mathcal{G}'(\mathcal{S}_k)$ is the set of UEs that AP $l \in \mathcal{S}_k$ knows the instantaneous CSI in $\mathcal{Q}_k$, $\mathcal{G}'(\mathcal{S}_k) = \{i\colon \mathcal{S}_i \cap \mathcal{S}_k \neq \emptyset, i \in \mathcal{Q}_k\}$. In this comparison, it is assumed that the number of elements in set $\mathcal{G}'(\mathcal{S}_k)$ is larger than $|\mathcal{S}_k \cup \mathcal{D}_k|N$. As shown in Table 1, the scalable iterative update method especially reduces the computational complexity required for the calculation of the Moore-Penrose general inverse matrix to obtain the BF vector.

In the following, we discuss the computational complexity in the path-loss based and SLNR-based methods. The path-loss based method does not need the complex calculations given in Table 1 to determine $\mathcal{J}_k$ since it independently selects APs with the U lowest path loss between it and UE k from the candidate APs. Therefore, the computational complexity level of this method is much lower than that for the proposed methods. On the other hand, the SLNR-based method needs to calculate the estimated interference power that the signal transmitted from the AP group to UE k gives to other UEs, the Moore-Penrose generalized inverse matrix to obtain the BF vector, and the metric, similar to the iterative update method. However, this method does not calculate the estimated interference power of UE k and does not iterate the above three calculations. This contrasts with the iterative update method, resulting in a lower computational complexity level. Although the proposed methods require a higher computational complexity level than that for the conventional methods, they increase the user throughput, which will be confirmed in Sect. 5.

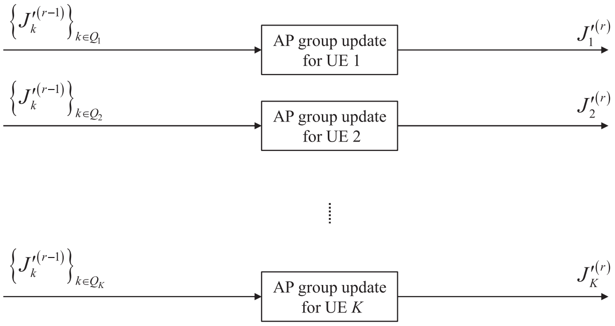

$\{J_k'^{(r-1)}\}_{k\in Q_1}$
AP group update for UE 1
$J_1'^{(r)}$
$\{J_k'^{(r-1)}\}_{k\in Q_2}$
AP group update for UE 2
$J_2'^{(r)}$
$\{J_k'^{(r-1)}\}_{k\in Q_K}$
AP group update for UE K
$J_K'^{(r)}$

Fig. 3 Flow of scalable iterative update method.

Table 1 Comparison of computational complexity.

Calculation	Complexity	
	Iterative update method	Scalable iterative update method
Estimated interference power that signal transmitted from AP group to UE k gives to other UEs	$\sum_{i\neq k, i\in\mathcal{K}} \lvert S_k \cap D_k \setminus S_i \rvert$	$\sum_{i\in Q_k} \lvert S_k \cap D_k \setminus S_i \rvert$
Moore-Penrose generalized inverse matrix	$O\left(\lvert S_k \cup D_k \rvert N \lvert \mathcal{G}(S_k \cup D_k)\rvert^2\right)$	$O\left(\lvert S_k \cup D_k \rvert N \lvert \mathcal{G}'(S_k \cup D_k)\rvert^2\right)$
Estimated interference power of UE k	$\sum_{i\neq k, i\in\mathcal{K}} \lvert S_i \cap D_i \setminus S_k \rvert$	$\sum_{i\in Q_k} \lvert S_i \cap D_i \setminus S_k \rvert$
Metric calculation per UE	$\lvert\mathcal{K}\rvert$	$\lvert\mathcal{Q}_k\rvert + 1$

5. Numerical Results

5.1 Simulation Parameters

The user throughput is evaluated by computer simulation according to Table 2. APs and UEs are randomly placed in a wraparound square system coverage with a side of 2 km according to the Poisson point process. The number of AP antennas is set to $N = 1$, as considered in related cell-free MIMO literature such as [7]–[9], [18]. The transmission bandwidth is set to 100 MHz, and the transmission signal power per UE is set to 40 dBm. The propagation model simulates the distance attenuation, shadowing, and instantaneous fading as given in Table 2, and the noise power density at the UE is set to −165 dBm/Hz. Term Δ_{CSI} in the partial CSI model described in Sect. 2 is parameterized. In the proposed methods, the number of iterations is set to two. In addition to the proposed methods, the scalable iterative update method that only performs initialization ($r = 0$), path-loss based method, and SLNR-based method are evaluated. Throughput is calculated based on Shannon's capacity formula. If $F(\mathcal{J}_k)$ is less than one, the throughput of UE k is set to zero.

Table 2 Simulation parameters.

Number of antennas per AP, N		1
Number of antennas per UE		1
Number of candidate APs per UE, T		15
System bandwidth		100 MHz
Node density	AP	50 / km^2
	UE	10 / km^2
Transmission power		40 dBm per UE
Distance-dependent path loss (including antenna gain)		114.1+37.6log$_{10}$(d), d: kilometers
Shadowing		Lognormal shadowing with standard deviation of 8 dB and inter-site correlation of 0.5
Instantaneous fading		Independent block Rayleigh
Receiver noise power density		−165 dBm/Hz
Number of iterations		2

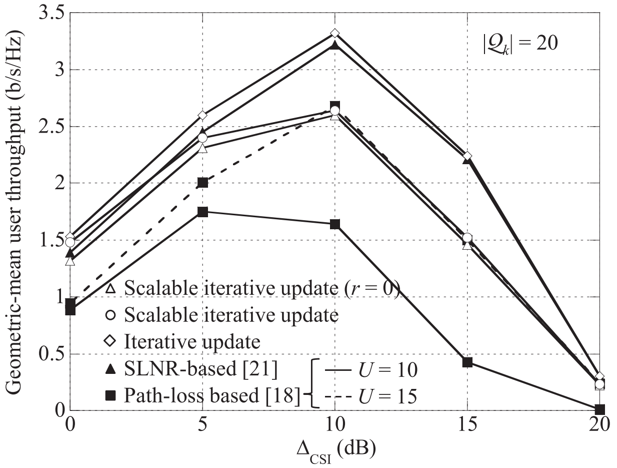


Fig. 4 Geometric-mean user throughput as a function of Δ_{CSI}.

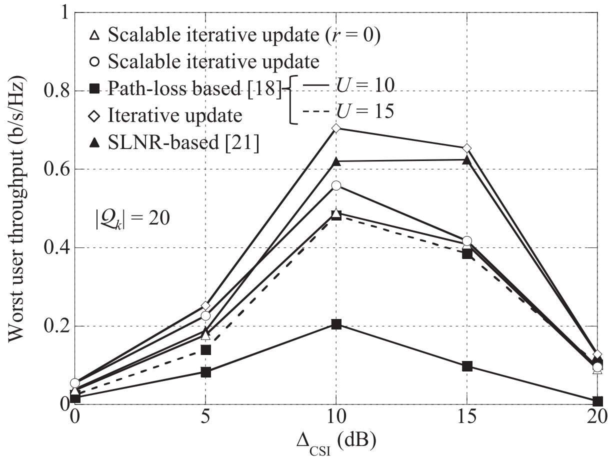


Fig. 5 Worst user throughput as a function of Δ_{CSI}.

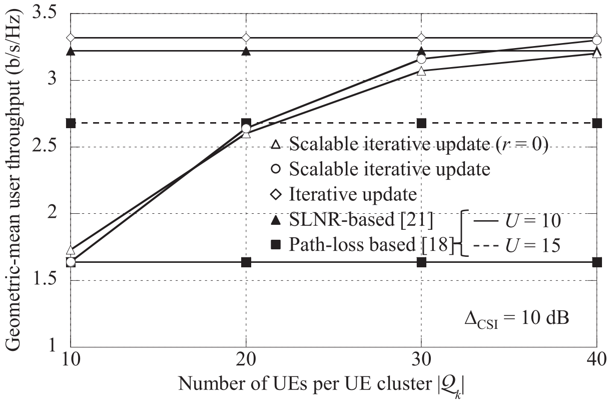


Fig. 6 Geometric-mean user throughput as a function of $|\mathcal{Q}_k|$.

5.2 Simulation Results

Figures 4 and 5 show the geometric-mean and worst user throughput as a function of Δ_{CSI}, respectively. In the scalable iterative update method, the number of UEs per UE cluster $|\mathcal{Q}_k|$ is set to 20, which is half the number of UEs $|\mathcal{K}|$ in the simulated system coverage area. The geometric-mean and worst user throughput for all the methods tend to decrease when Δ_{CSI} is greater than 10 dB. This is because the number of other UEs that need to be nulled for data transmission to a UE tends to increase when Δ_{CSI} is large. If the number of such UEs is close to the total number of transmitting AP antennas, the received signal power gain by the BF is liable to be small. In addition, data transmission is impossible in the extreme case where there are more UEs that need to be nulled for data transmission to a UE than the number of antennas of all transmitting APs, resulting in a lack of spatial degrees of freedom. As shown in Figs. 4 and 5, most of the methods achieve the highest throughput when Δ_{CSI} is 10 dB. Meanwhile, the large value of Δ_{CSI} necessitates many APs estimating the CSI using the reference signal with a low received signal power. This results in non-negligible CSI estimation errors and is thus infeasible in real systems. It is shown that cell range expansion, which biases the received signal power in the handover criteria and is deployed in 4G and 5G cellular systems, can compensate for the difference in the channel conditions by setting its bias value to 10 dB [24]. Since this bias value is equivalent to Δ_{CSI}, it indicates the validity of using $\Delta_{\mathrm{CSI}} = 10$ dB. Therefore, Δ_{CSI} is set to 10 dB in the subsequent evaluations.

The iterative update method achieves higher geometric-mean and worst user throughput than other methods. This indicates that the iterative update method selects transmitting AP groups considering the effect among all UEs. The throughput of the pass-loss based method with $U = 15$ is higher than that with $U = 10$, while it is lower than the SLNR-based method.

The scalable iterative update method outperforms the path-loss based method and improves the user throughput by updating with iterations. Moreover, this method achieves geometric-mean and worst user throughput levels of greater than 60% for each of the SLNR-based and iterative update methods. This implies that the scalable iterative update method can select the appropriate transmitting AP group since updating them has a great impact on the throughput of UEs that neighbor a target UE.

Figures 6 and 7 show geometric-mean and worst user throughput as a function of $|\mathcal{Q}_k|$, respectively. From the figures, we see that the scalable iterative update method selects the transmitting AP group that improves both geometric-mean and worst user throughput as $|\mathcal{Q}_k|$ increases. This is because the AP group selection in this method can appropriately take into account the effects among UEs thanks to an increase in UEs that are considered in the calculation of the effective channel gain and interference power. When $|\mathcal{Q}_k|$ is greater than 30, the throughput of the scalable iterative update method is close to that of the SLNR-based and iterative update methods. In addition, it achieves the same level of throughput at $|\mathcal{Q}_k| = 20$ compared to the path-loss based method with $U = 15$.

Finally, Fig. 8 shows the cumulative probability of the number of APs used per UE. Both of the proposed meth-

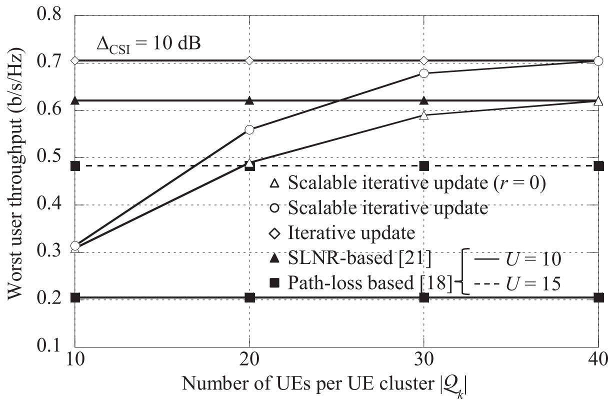


Fig. 7 Worst user throughput as a function of $|Q_k|$.

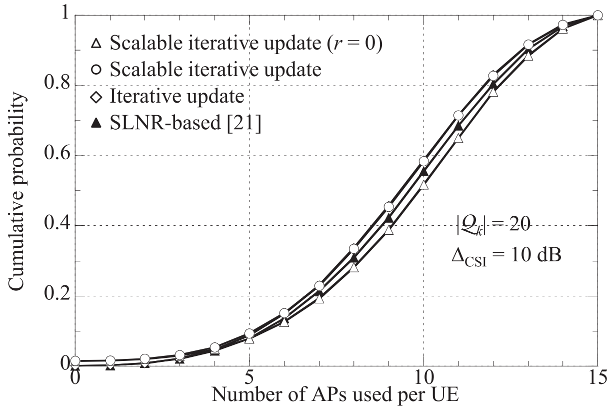


Fig. 8 Cumulative probability of the number of APs per UE.

ods achieve comparable throughput to the path-loss based and SLNR-based methods while reducing the number of transmitting APs per UE. Therefore, the proposed methods achieve good throughput while maintaining the same number of APs per UE, which may lead to reduced complexity of the system configuration.

6. Conclusion

We considered a downlink cell-free MIMO system with partial instantaneous CSI knowledge and proposed two AP clustering methods that consider the throughput and interference among UEs. The first proposed method, the iterative update method, fully considers the effect among all other UEs in the system coverage and successively updates transmitting APs for each UE with the calculation of the estimated throughput. The other proposed method, the scalable iterative update method, determines the UE cluster for each UE, which comprises the UEs considered in the AP clustering process, and updates the AP groups independently, reducing the computational complexity level. The proposed AP clustering methods improve the geometric-mean and worst user throughput compared to those for the conventional methods while avoiding an excessive increase in the number of transmitting APs per UE. Moreover, the scalable iterative update method achieves comparable throughput to the iterative update and SLNR-based methods while reducing the number of UEs considered in the AP clustering process. In the future, we plan to investigate ways to reduce the load on the coordination process required for the AP clustering in realistic environments where coordination among APs via a central processing unit is limited. Moreover, we will consider improving the layered BF method based on channel-matrix muting and extending the method to cope with the limitation of the transmission power per AP.

References

[1] H. Zhang and H. Dai, “Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks,” EURASIP J. Wirel. Commun. Netw., vol.2004, no.2, pp.222–235, 2004.

[2] S. Jing, D.N.C. Tse, J.B. Soriaga, J. Hou, J.E. Smee, and R. Padovani, “Multicell downlink capacity with coordinated processing,” EURASIP J. Wirel. Commun. Netw., vol.2008, pp.1–19, 2008.

[3] D. Gesbert, S. Hanly, H. Huang, S. Shamai, O. Simeone, and W. Yu, “Multi-cell MIMO cooperative networks: A new look at interference,” IEEE J. Sel. Areas Commun., vol.28, no.9, pp.1380–1408, Dec. 2010.

[4] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, “Coordinated multipoint transmission/reception techniques for LTE-Advanced,” IEEE Wireless Commun., vol.17, no.3, pp.26–34, June 2010.

[5] W. Choi and G. Andrews, “Downlink performance and capacity of distributed antenna systems in a multicell environment,” IEEE Trans. Wireless Commun., vol.6, no.1, pp.69–73, Jan. 2007.

[6] J. Park, E. Song, and W. Sung, “Capacity analysis for distributed antenna systems using cooperative transmission schemes in fading channels,” IEEE Trans. Wireless Commun., vol.8, no.2, pp.586–592, Feb. 2009.

[7] H.Q. Ngo, A. Ashikhmin, H. Yang, E.G. Larsson, and T.L. Marzetta, “Cell-free massive MIMO versus small cells,” IEEE Trans. Wireless Commun., vol.16, no.3, pp.1834–1850, March 2017.

[8] E. Nayebi, A. Ashikhmin, T.L. Marzetta, H. Yang, and B.D. Rao, “Precoding and power optimization in cell-free massive MIMO system,” IEEE Trans. Wireless Commun., vol.16, no.7, pp.4445–4459, July 2017.

[9] E. Bjorson and L. Sanguinetti, “Scalable cell-free massive MIMO systems,” IEEE Trans. Commun., vol.68, no.7, pp.4247–4261, July 2020.

[10] E. Dahlman, S. Parkvall, and J. Sköld, 5G NR: The Next Generation Wireless Access Technology, Academic Press, 2018.

[11] NTT DOCOMO, “White paper: 5G evolution and 6G,” Jan. 2022.

[12] C. Pan, M. Elkashlan, J. Wang, J. Yuan, and L. Hanzo, “User-centric C-RAN architecture for ultra-dense 5G networks: Challenges and methodologies,” IEEE Commun. Mag., vol.56, no.6, pp.14–20, June 2018.

[13] Y. Tajika, H. Taoka, and K. Higuchi, “Partially non-orthogonal block diagonalization-based precoding in downlink multiuser MIMO with limited channel state information feedback,” IEICE Trans. Commun., vol.E94-B, no.12, pp.3280–3288, Dec. 2011.

[14] S. Shim, J.S. Kwak, R.W. Heath, and J.G. Andrews, “Block diagonalization for multi-user MIMO with other-cell interference,” IEEE Trans. Wireless Commun., vol.7, no.7, pp.2671–2681, July 2008.

[15] W.W.L. Ho, T.Q.S. Quek, and S. Sun, “Decentralized base station processing for multiuser MIMO downlink CoMP,” Proc. IEEE VTC2010-Spring, May 2010.

[16] G. Interdonato, M. Karlsson, E. Björnson, and E. Larsson, “Local partial zero-forcing precoding for cell-free massive MIMO,” IEEE Trans. Wireless Commun., vol.19, no.7, pp.4758–4774, July 2020.

[17] L. Du, L. Li, H.Q. Ngo, and M. Matthaiou, "Cell-free massive MIMO: Joint maximum-ratio and zero-forcing precoder with power control," IEEE Trans. Commun., vol.69, no.6, pp.3741–3756, June 2021.

[18] M. Mojahedian and A. Lozano, "Subset regularized zero-forcing precoders for cell-free C-RANs," Proc. IEEE 2021 29th EUSIPCO, Aug. 2021.

[19] K. Higuchi, "Layered block diagonalization for base station cooperated multiuser MIMO with partial channel state information feedback," Proc. IEEE ICNC2012, Jan.-Feb. 2012.

[20] Y. Oshima, A. Benjebbour, and K. Higuchi, "Throughput performance of layered partially non-orthogonal block diagonalization with adaptive interference admission control in distributed antenna system," Proc. IEEE ICCS2012, Nov. 2012.

[21] Y. Oshima, A. Benjebbour, and K. Higuchi, "A novel adaptive interference admission control method for layered partially non-orthogonal block diagonalization for base station cooperative MIMO," IEICE Trans. Commun., vol.E97-B, no.1, pp.155–163, Jan. 2014.

[22] Q.H. Spencer, A.L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," IEEE Trans. Signal Process., vol.52, no.2, pp.461–471, Feb. 2004.

[23] D. Ishii, T. Hara, N. Nonaka, and K. Higuchi, "Clustering method in downlink cell-free MIMO using layered partially non-orthogonal ZF-based beamforming," Proc. IEEE VTC2023-Spring, June 2023.

[24] M. Shirakabe, A. Morimoto, and N. Miki, "Performance evaluation of inter-cell interference coordination and cell range expansion in heterogeneous networks for LTE-Advanced downlink," Proc. IEEE ISWCS 2011, Nov. 2011.

Daisuke Ishii received the B.E. and M.E. degrees from Tokyo University of Science, Noda, Japan in 2022 and 2024, respectively. In 2024, he joined NTT East Corporation. His research interests include wireless communications.

Takanori Hara received the B.E., M.E., and Ph.D. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2017, 2019, and 2022, respectively. Since April 2022, he has been with the Department of Electrical Engineering, at Tokyo University of Science, Chiba, Japan, where he is currently an Assistant Professor. His current research interests are grant-free access, compressed sensing, and MIMO technologies.

Kenichi Higuchi received the B.E. degree from Waseda University, Tokyo, Japan, in 1994, and received the Dr.Eng. degree from Tohoku University, Sendai, Japan in 2002. In 1994, he joined NTT Mobile Communications Network, Inc. (now, NTT DOCOMO, INC.). While with NTT DOCOMO, INC., he was engaged in the research and standardization of wireless access technologies for wideband DS-CDMA mobile radio, HSPA, LTE, and broadband wireless packet access technologies for systems beyond IMT-2000. In 2007, he joined the faculty of the Tokyo University of Science and currently holds the position of Professor. His current research interests are in the areas of wireless technologies and mobile communication systems, including advanced multiple access such as non-orthogonal multiple access (NOMA), radio resource allocation, inter-cell interference coordination, multiple-antenna transmission techniques, signal processing such as interference cancellation and turbo equalization, and issues related to heterogeneous networks using small cells. He was a co-recipient of the Best Paper Award of the International Symposium on Wireless Personal Multimedia Communications in 2004 and 2007, the Best Paper Award from the IEICE in 2021, a recipient of the Young Researcher's Award from the IEICE in 2003, the 5th YRP Award in 2007, the Prime Minister Invention Prize in 2010, and the Invention Prize of Commissioner of the Japan Patent Office in 2015. He is a senior member of the IEEE.

PAPER

Peak Cancellation Signal Generation Considering Variance in Signal Power among Transmitter Antennas in PAPR Reduction Method Using Null Space in MIMO Channel for MIMO-OFDM Signals*

Jun SAITO†, Nobuhide NONAKA††, *Members*, *and* Kenichi HIGUCHI†a), *Senior Member*

SUMMARY We propose a novel peak-to-average power ratio (PAPR) reduction method based on a peak cancellation (PC) signal vector that considers the variance in the average signal power among transmitter antennas for massive multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) signals using the null space in a MIMO channel. First, we discuss the conditions under which the PC signal vector achieves a sufficient PAPR reduction effect after its projection onto the null space of the MIMO channel. The discussion reveals that the magnitude of the correlation between the PC signal vector before projection and the transmission signal vector should be as low as possible. Based on this observation and the fact that to reduce the PAPR it is helpful to suppress the variation in the transmission signal power among antennas, which may be enhanced by beamforming (BF), we propose a novel method for generating a PC signal vector. The proposed PC signal vector is designed so that the signal power levels of all the transmitter antennas are limited to be between the maximum and minimum power threshold levels at the target timing. The newly introduced feature in the proposed method, i.e., increasing the signal power to be above the minimum power threshold, contributes to suppressing the transmission signal power variance among antennas and to improving the PAPR reduction capability after projecting the PC signal onto the null space in the MIMO channel. This is because the proposed method decreases the magnitude of the correlation between the PC signal vectors before its projection and the transmission signal vectors. Based on computer simulation results, we show that the PAPR reduction performance of the proposed method is improved compared to that for the conventional method and the proposed method reduces the computational complexity compared to that for the conventional method for achieving the same target PAPR.

key words: *OFDM, PAPR reduction, MIMO, null space, peak cancellation, computational complexity reduction*

1. Introduction

The combination of downlink massive multiple-input multiple-output (MIMO) [1], [2], in which the number of transmitter antennas is very large, and orthogonal frequency division multiplexing (OFDM) signals achieves wide-coverage enhanced mobile broadband communications. However, the peak-to-average power ratio (PAPR) of the OFDM signals is high. In a massive MIMO scenario, the PAPR of the OFDM signals may be further increased due to the variation in the transmission signal power levels among antennas through the beamforming (BF) process. In a massive MIMO environment, a power amplifier with relatively low power consumption needs to be used for each of a large number of transmitter antennas. Therefore, in massive MIMO-OFDM transmission, PAPR reduction is a crucial issue that must be addressed.

In downlink massive MIMO, the number of transmission antennas at a base station is in general much larger than that for receiver antennas at the user terminal. Under this assumption, joint optimization of BF, OFDM modulation, and PAPR reduction was studied in [3]. In [4], a PAPR reduction method was reported in which some of the transmission antennas are exclusively used to reduce the PAPR. In this method, the in-band interference due to the signal for PAPR reduction is eliminated on the receiver side. However, this method decreases the BF gain of the data streams due to the decrease in the number of transmitter antennas used for transmitting the data streams.

In [5]–[16], a PAPR reduction method was discussed that uses the null space in a MIMO channel. This method limits the signal for PAPR reduction transmitted to only the null space in the MIMO channel by applying BF to remove the in-band interference due to the PAPR reduction signal on the receiver side. Since all the transmission antennas are fully utilized for transmission of the data streams, the achievable BF gain of this method is greater than that for the method in [4].

Members of our research group reported a PAPR reduction algorithm based on the peak cancellation (PC) signal in order to reduce the computational complexity in the PAPR reduction method using the null space of a MIMO channel [17]–[19]. This algorithm is referred to as PC with a channel-null constraint (PCCNC). The original idea for the method using the PC signal was reported in [20], [21]. The PC signal is designed so that it has a single dominant peak and satisfies the requirement for out-of-band radiation. By directly adding the PC signal to the transmission signal in the time domain at each transmission antenna, the PAPR of

Manuscript received July 30, 2023.
Manuscript revised March 31, 2024.
Manuscript publicized May 6, 2024.

†Graduate School of Science and Technology, Tokyo University of Science, Noda-shi, 278-8510 Japan.
††6G-IOWN Promotion Department, NTT DOCOMO, INC., Yokosuka-shi, 239-8536 Japan.

*The material in this paper was presented in part at IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, Scotland, March 2023.

a) E-mail: higuchik@rs.tus.ac.jp

DOI: 10.23919/transcom.2023EBP3125

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

the transmission signal is suppressed. PCCNC performs PC signal-based PAPR reduction jointly considering all transmission signals for all antennas. Thus, the PC signal is constructed in vector form. The BF vector of the PC signal is set orthogonal to the MIMO channel. In other words, the BF vector of the PC signal is restricted within the null space in the MIMO channel. With this restriction, the interference due to the PC signal that is imparted to the data streams is removed on the receiver side.

The original PCCNC in [17]–[19] first generates the PC signal vector which suppresses all the peak signal components whose power levels are beyond the maximum power threshold in the transmission signals of all transmitter antennas at the target timing. Then, the generated PC signal vector is projected onto the null space in the MIMO channel to calculate the final version of the PC signal vector. However, the PAPR reduction effect of the PC signal vector is degraded to some extent by the projecting operation of the PC signal vector onto the null space in the MIMO channel. Therefore, there is room for improvement in the generation method of the PC signal that suppresses the degradation of the PAPR reduction effect due to projection onto the null space in the MIMO channel. In the following, the original PCCNC in [17] is referred to as conventional PCCNC.

This paper proposes a novel generation method for the PC signal vector for PCCNC. First, we discuss the conditions under which the PC signal vector achieves a sufficient PAPR reduction effect after its projection onto the null space of the MIMO channel. The discussion reveals that the magnitude of the correlation between the PC signal vector before projection onto the null space and the transmission signal vector should be as low as possible. Based on this observation and the fact that it is helpful for PAPR reduction to suppress the variation in the transmission signal power among antennas, which may be enhanced by BF, we propose a novel generation method for the PC signal vector. The proposed PC signal vector is designed so that the signal power levels of all the transmitter antennas are limited to be between the maximum and minimum power threshold levels at the target timing. The newly introduced feature in the proposed method, i.e., increasing the signal power to be above the minimum power threshold, contributes to suppressing the transmission signal power variance among antennas, which directly contributes to the PAPR reduction among antennas. Furthermore, by balancing the PC signal components that reduce the peak signal power and that increase the signal power in the transmission signal vector, the proposed PC signal vector can be uncorrelated with the transmission signal vector. Therefore, the PAPR reduction capability after the projection of the PC signal vector onto the null space in the MIMO channel can be enhanced compared to that for the conventional PCCNC. Based on computer simulation results, we show that the PAPR reduction performance of the proposed PCCNC is improved compared to that of the conventional one and the proposed PCCNC reduces the computational complexity compared to that for the conventional one for achieving the same target PAPR by reducing the required number of iterations in the PCCNC process. We note that the contents of this paper are based on [22], but include enhanced evaluations such as the complexity analysis, evaluation of the complementary cumulative distribution function (CCDF) of the PAPR and throughput, and performance evaluation of the proposed method with various numbers of transmitter antennas.

The remainder of the paper is organized as follows. First, Sect. 2 describes the PCCNC using the conventional PC signal vector generation method, and discusses the conditions for the PC signal vector to achieve sufficient PAPR reduction. Section 3 describes the proposed PC signal vector generation method. Section 4 shows the numerical results based on computer simulations. Finally, Sect. 5 concludes the paper.

2. Conventional PCCNC and Conditions for PC Signal Vector to Achieve Sufficient PAPR Reduction

2.1 Conventional PCCNC

Let us consider MIMO multiplexing where the number of transmitter antennas at the base station is $N_{\rm tx}$ and the number of user terminals each having a single receiver antenna is $N_{\rm rx}$. We set $N_{\rm tx} > N_{\rm rx}$ assuming a downlink massive MIMO scenario. In this paper, we assume that the MIMO channel is not frequency selective for simplicity. The $N_{\rm rx} \times N_{\rm tx}$-dimensional channel matrix is denoted as $\mathbf{H}$. Since $N_{\rm tx}$ is greater than $N_{\rm rx}$, we have $N_{\rm tx} \times (N_{\rm tx} - N_{\rm rx})$-dimensional matrix $\mathbf{K}$ that satisfies $\mathbf{HK} = \mathbf{O}$. All the $N_{\rm tx} - N_{\rm rx}$ column vectors in $\mathbf{K}$ are orthonormalized to each other. Matrix $\mathbf{K}$ corresponds to the null space in MIMO channel $\mathbf{H}$.

Hereafter, we explain the conventional PCCNC [17]. Figure 1 shows the block diagram of a base station transmitter with PCCNC. First, the OFDM data stream signal is generated. Then, BF is applied to the OFDM data stream signal to generate the $N_{\rm tx}$-dimensional transmission signal vector of data streams. After that, PCCNC is processed for a given transmission signal vector in order to reduce the PAPR of the transmission signal vector.

The $N_{\rm rx}$-dimensional data stream vector at discrete time t ($t = 0, \ldots, F-1$: F is the number of inverse fast Fourier transform (IFFT) points in OFDM signal generation) is denoted as $\mathbf{s}[t]$. The $N_{\rm tx} \times N_{\rm rx}$-dimensional BF matrix, $\mathbf{M}$, is multiplied to $\mathbf{s}[t]$ to generate the $N_{\rm tx}$-dimensional time-domain transmission signal vector before PAPR reduction, $\mathbf{x}[t]$. Vector $\mathbf{x}[t]$ is represented as

$$\mathbf{x}[t] = \begin{bmatrix} x_1[t] & \cdots & x_{N_{\rm tx}}[t] \end{bmatrix}^T = \mathbf{Ms}[t], \qquad (1)$$

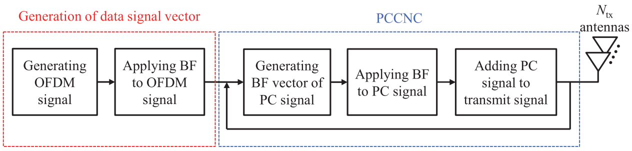

Generation of data signal vector
PCCNC
$N_{\rm tx}$ antennas
Generating OFDM signal
Applying BF to OFDM signal
Generating BF vector of PC signal
Applying BF to PC signal
Adding PC signal to transmit signal

Fig. 1 Block diagram of base station transmitter with PCCNC.

where $x_n[t]$ is the n-th element of $\mathbf{x}[t]$, which is transmitted from transmitter antenna n.

PCCNC reduces the PAPR by iteratively repeating the direct addition of the PC signal vectors to transmission signal vector $\mathbf{x}[t]$. The PC signal vector is generated by multiplying BF vector, $\mathbf{m}_{\mathrm{pc}}$, which is directed to the null space in the MIMO channel, to the time-shifted version of basic time-domain signal, $b[t]$, which is a sinc function whose dominant peak amplitude at $t = 0$ is set to 1 and whose bandwidth is equal to the signal transmission bandwidth.

The time-domain transmission signal vector at the j-th iteration of the PCCNC process is denoted as $\mathbf{x}^{(j)}[t] = \left[x_1^{(j)}[t] \quad \cdots \quad x_{N_{\mathrm{tx}}}^{(j)}[t]\right]^T$. We assume $\mathbf{x}^{(1)}[t] = \mathbf{x}[t]$ as the initial setting. At the j-th iteration, PCCNC tries to reduce the PAPR observed in $\mathbf{x}^{(j)}[t]$ at target time index $\tau^{(j)}$. Target time index $\tau^{(j)}$ is determined based on the peak observation in $\mathbf{x}^{(j)}[t]$. More specifically, $\tau^{(j)}$ is set to the time index where $x_n^{(j)}[t]$ has the maximum amplitude for all $n = 1, \ldots, N_{\mathrm{tx}}$ and $t = 0, \ldots, F - 1$. In the j-th iteration of the conventional PCCNC, the PC signal vector, $\mathbf{c}^{(j)}[t]$, which is represented in the form of (2), is added to $\mathbf{x}^{(j)}[t]$.

$$\mathbf{c}^{(j)}[t] = \mathbf{m}_{\mathrm{pc}}^{(j)} b\left[t - \tau^{(j)}\right]. \tag{2}$$

$$\mathbf{x}^{(j+1)}[t] = \mathbf{x}^{(j)}[t] + \mathbf{c}^{(j)}[t]. \tag{3}$$

The purpose of $\mathbf{c}^{(j)}[t]$ in the conventional PCCNC is to suppress the peak signal observed in $\mathbf{x}^{(j)}[\tau^{(j)}]$ using the dominant peak signal portion of $b[0]$.

An important component of $\mathbf{c}^{(j)}[t]$ is its BF vector, $m_{\mathrm{pc}}^{(j)}$, and investigating its generation method is the purpose of this paper. The N_{tx}-dimensional vector, $\mathbf{m}_{\mathrm{pc}}^{(j)}$, should lie within null space $\mathbf{K}$ in the MIMO channel. PC signal vector $\mathbf{c}^{(j)}[t]$ is generated by multiplying $\mathbf{m}_{\mathrm{pc}}^{(j)}$ to the $\tau^{(j)}$-time-shifted version of $b[t]$. Thus, PC signal vector $\mathbf{c}^{(j)}[t]$ meets the requirement of out-of-band radiation and does not interfere with the data streams. This is because the PC signal is transmitted to only the null space in the given MIMO channel and does not appear on the receiver side. In generating $\mathbf{m}_{\mathrm{pc}}^{(j)}$, the first step is to find BF vector $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ of the PC signal vectors that are desirable from the viewpoint of PAPR suppression. Next, $\mathbf{m}_{\mathrm{pc}}^{(j)}$ is generated by projecting $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ onto null space $\mathbf{K}$ in the MIMO channel.

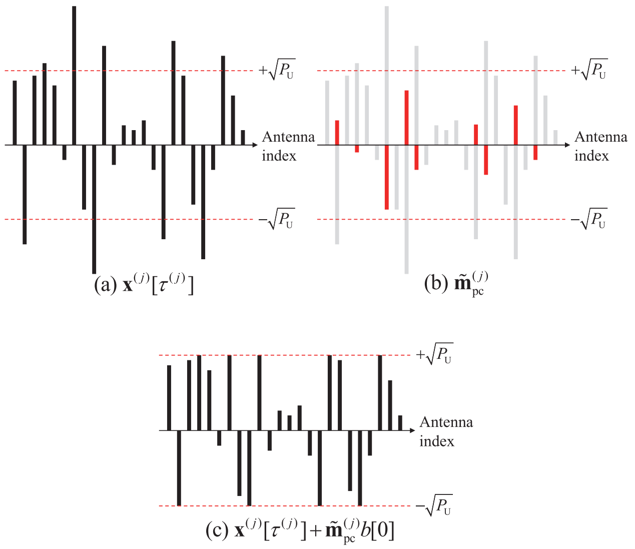
Figure 2 shows the conventional method for PC signal generation described in [17]. In [17], $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ is generated using the following formula.

$$\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)} = \left[\tilde{m}_{\mathrm{pc},1}^{(j)} \quad \cdots \quad \tilde{m}_{\mathrm{pc},N_{\mathrm{tx}}}^{(j)}\right]^T,$$

where

$$\tilde{m}_{\mathrm{pc},n}^{(j)} = \begin{cases} \sqrt{P_{\mathrm{U}}} e^{j\theta_n^{(j)}[\tau^{(j)}]} - x_n^{(j)}\left[\tau^{(j)}\right], & \left|x_n^{(j)}\left[\tau^{(j)}\right]\right|^2 > P_{\mathrm{U}} \\ 0, & \text{Otherwise} \end{cases}. \tag{4}$$

Here, P_{U} is the maximum power threshold and $\theta_n^{(j)}[\tau^{(j)}]$ is the phase of $x_n^{(j)}[\tau^{(j)}]$. Vector $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ is the ideal peak reduction vector in the sense that if $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ is used as $\mathbf{m}_{\mathrm{pc}}^{(j)}$ in (2), the transmission signal power levels for all N_{tx} antennas at time $\tau^{(j)}$ (thus, $|x_n^{(j)}[\tau^{(j)}]|^2$ for all n) can simultaneously be suppressed to equal to or lower than maximum power threshold P_{U} as shown in Fig. 2.



(a) $\mathbf{x}^{(j)}[\tau^{(j)}]$

(b) $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$

(c) $\mathbf{x}^{(j)}[\tau^{(j)}]+\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}b[0]$

Fig. 2 Conventional PC signal vector.

However, $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ has an element that is orthogonal to the null space in the MIMO channel, which leads to interference to the data streams. Therefore, $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ is projected onto null space $\mathbf{K}$ in the MIMO channel to obtain $\mathbf{m}_{\mathrm{pc}}^{(j)}$ in (2).

$$\mathbf{m}_{\mathrm{pc}}^{(j)} = \mathbf{K}\mathbf{K}^H\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}. \tag{5}$$

2.2 Discussions on Conditions for PC Signal Vector to Achieve Sufficient PAPR Reduction Effect

The conventional method for generating the BF vector of the PC signal before projection $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$ in [17] is based on a standard approach in PAPR reduction, i.e., clipping of the high peak power components. However, as revealed below, there is room for improvement in terms of the PAPR reduction effect after its projection onto the null space $\mathbf{K}$ in the MIMO channel. Hereafter, we analyze the impact of the projection of the PC signal vector onto the null space in the MIMO channel on the PAPR reduction capability. For simplicity of notation, we consider the $j = 1$-st PAPR reduction process of PCCNC and denote $\mathbf{x}^{(1)}[t] = \mathbf{x}[t]$, $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(1)}$, and $\mathbf{m}_{\mathrm{pc}}^{(1)}$ by $\mathbf{x}$, $\tilde{\mathbf{m}}_{\mathrm{pc}}$, and $\mathbf{m}_{\mathrm{pc}}$, respectively.

PCCNC must have a large correlation between $\tilde{\mathbf{m}}_{\mathrm{pc}}$ and $\mathbf{m}_{\mathrm{pc}}$, which is the projection of $\tilde{\mathbf{m}}_{\mathrm{pc}}$ onto the null space $\mathbf{K}$ in the MIMO channel, in order to obtain a sufficient PAPR reduction effect, assuming that $\tilde{\mathbf{m}}_{\mathrm{pc}}$ is designed to achieve a good PAPR reduction capability. The correlation between $\mathbf{m}_{\mathrm{pc}}$ and $\tilde{\mathbf{m}}_{\mathrm{pc}}$, $R(\mathbf{m}_{\mathrm{pc}})$, is represented as

$$R\left(\mathbf{m}_{\mathrm{pc}}\right) = \frac{\tilde{\mathbf{m}}_{\mathrm{pc}}^H\mathbf{m}_{\mathrm{pc}}}{\left\|\tilde{\mathbf{m}}_{\mathrm{pc}}\right\| \cdot \left\|\mathbf{m}_{\mathrm{pc}}\right\|} = \frac{\tilde{\mathbf{m}}_{\mathrm{pc}}^H\mathbf{K}\mathbf{K}^H\tilde{\mathbf{m}}_{\mathrm{pc}}}{\left\|\tilde{\mathbf{m}}_{\mathrm{pc}}\right\| \cdot \left\|\mathbf{m}_{\mathrm{pc}}\right\|} = \frac{\left\|\mathbf{K}^H\tilde{\mathbf{m}}_{\mathrm{pc}}\right\|^2}{\left\|\tilde{\mathbf{m}}_{\mathrm{pc}}\right\| \cdot \left\|\mathbf{m}_{\mathrm{pc}}\right\|}. \tag{6}$$

From (6), $R(\mathbf{m}_{\text{pc}})$ is always a non-negative real number and is determined by $\|\mathbf{K}^H\tilde{\mathbf{m}}_{\text{pc}}\|$. Therefore, it is important to prevent $\mathbf{K}$ and $\tilde{\mathbf{m}}_{\text{pc}}$ from being orthogonal as much as possible.

Clearly, the BF vector of the PC signal before projection $\tilde{\mathbf{m}}_{\text{pc}}$ is generated based on given transmission signal vector $\mathbf{x}$. BF matrix $\mathbf{M}$ for data streams to generate $\mathbf{x}$ as $\mathbf{x} = \mathbf{Ms}$ is orthogonal to null space $\mathbf{K}$ in the MIMO channel. This is because $\mathbf{M}$ aims to send data stream $\mathbf{s}$ to the destination user terminal regardless of the BF criteria such as zero forcing (ZF), minimum mean squared error (MMSE), or an eigenmode MIMO-based criterion. Therefore,

$$\mathbf{K}^H\mathbf{x} = \mathbf{K}^H\mathbf{Ms} = \mathbf{0}. \tag{7}$$

Thus, transmission signal vector $\mathbf{x}$ is orthogonal to $\mathbf{K}$.

The conventional method generates $\tilde{\mathbf{m}}_{\text{pc}}$ so that it eliminates the signal elements above maximum power threshold P_{U}. Therefore, $\tilde{\mathbf{m}}_{\text{pc}}$ and transmission signal vector $\mathbf{x}$ have a negative correlation as illustrated in Fig. 2. If we assume the extreme condition of $P_{\text{U}} = 0$, $\tilde{\mathbf{m}}_{\text{pc}}$ becomes $-\mathbf{x}$ and $R(\mathbf{m}_{\text{pc}})$ becomes

$$R\left(\mathbf{m}_{\text{pc}}\right) = \frac{\left\|\mathbf{K}^H\tilde{\mathbf{m}}_{\text{pc}}\right\|^2}{\left\|\tilde{\mathbf{m}}_{\text{pc}}\right\| \cdot \left\|\mathbf{m}_{\text{pc}}\right\|} = \frac{\left\|-\mathbf{K}^H\mathbf{x}\right\|^2}{\left\|\tilde{\mathbf{m}}_{\text{pc}}\right\| \cdot \left\|\mathbf{m}_{\text{pc}}\right\|} = 0. \tag{8}$$

This means that after projecting $\tilde{\mathbf{m}}_{\text{pc}}$ onto null space $\mathbf{K}$ in the given MIMO channel, $\mathbf{m}_{\text{pc}}$ becomes $\mathbf{0}$ and no PAPR reduction is achieved.

Based on the above discussion, the BF vector of the PC signal before projection, $\tilde{\mathbf{m}}_{\text{pc}}$, should be uncorrelated with transmission signal vector $\mathbf{x}$ as much as possible. Then, null space $\mathbf{K}$ and $\tilde{\mathbf{m}}_{\text{pc}}$ will no longer be orthogonal, so $R(\mathbf{m}_{\text{pc}})$ will increase and the PAPR reduction effect can be expected to increase.

3. Proposed PC Signal Vector Generation Method

Based on the discussion in Sect. 2, we propose a novel method for generating the PC signal vector, more specifically, a new method for generating $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$.

Conventional $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ calculated using (4) has a negative correlation with transmission signal vector $\mathbf{x}^{(j)}[\tau^{(j)}]$. If the magnitude of the correlation between $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ and $\mathbf{x}^{(j)}[\tau^{(j)}]$ can be decreased by adding a component that has a positive correlation with transmission signal vector $\mathbf{x}^{(j)}[\tau^{(j)}]$ with respect to conventional $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$, the PAPR reduction capability after projecting $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ onto the null space in the MIMO channel can be enhanced.

On the other hand, the PAPR in MIMO transmission also increases due to the variance in the average transmission signal power among transmission antennas due to BF. Here, if a signal component that increases the transmission signal power for an antenna with a low transmission signal power is included in $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ in order to decrease the variance of the average transmission signal power levels among transmission antennas, such a signal component can be expected to have a positive correlation with $\mathbf{x}^{(j)}[\tau^{(j)}]$.

Based on the above analysis, we propose a new generation method for the PC signal vector. In the proposed method, the BF vector of the PC signal before projection, $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$, is designed so that the signal power levels of all the transmission antennas at the target timing are restricted to be between the maximum and minimum power threshold levels. In the proposed method, we define the minimum power threshold, P_{L}, in addition to the maximum power threshold P_{U}, where P_{U} is greater than P_{L}. Proposed $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ is represented as

$$\tilde{\mathbf{m}}_{\text{pc}}^{(j)} = \left[\tilde{m}_{\text{pc},1}^{(j)} \quad \cdots \quad \tilde{m}_{\text{pc},N_{\text{tx}}}^{(j)}\right]^T,$$

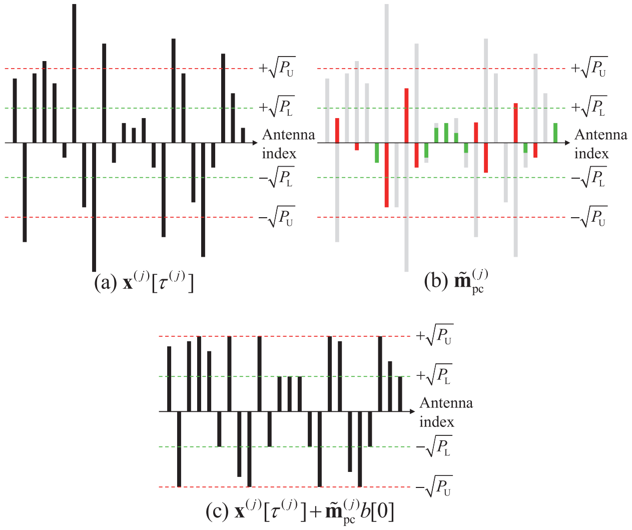
where

$$\tilde{m}_{\text{pc},n}^{(j)} = \begin{cases} \sqrt{P_{\text{U}}}e^{j\theta_n^{(j)}[\tau^{(j)}]} - x_n^{(j)}\left[\tau^{(j)}\right], & \left|x_n^{(j)}\left[\tau^{(j)}\right]\right|^2 > P_{\text{U}} \\ \sqrt{P_{\text{L}}}e^{j\theta_n^{(j)}[\tau^{(j)}]} - x_n^{(j)}\left[\tau^{(j)}\right], & \left|x_n^{(j)}\left[\tau^{(j)}\right]\right|^2 < P_{\text{L}} \\ 0, & \text{Otherwise} \end{cases}. \tag{9}$$

Figure 3 shows the proposed PC signal generation. The first component of $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ shown in the first line in (9), which is shown in red in Fig. 3(b), is the same as in the conventional method using (4). With this component, the transmission signal power levels for all N_{tx} antennas at time $\tau^{(j)}$ (thus, $|x_n^{(j)}[\tau^{(j)}]|^2$ for all n) can simultaneously be set equal to or lower than maximum power threshold P_{U}. This component has a negative correlation with $\mathbf{x}^{(j)}[\tau^{(j)}]$.

The second component of $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ shown in the second line in (9), which is shown in green in Fig. 3(b), is newly introduced in the proposed method. With this component, the transmission signal power levels for all N_{tx} antennas at time $\tau^{(j)}$ (thus, $|x_n^{(j)}[\tau^{(j)}]|^2$ for all n) can simultaneously be set equal to or higher than minimum power threshold P_{L}. This component has a positive correlation with $\mathbf{x}^{(j)}[\tau^{(j)}]$.

By balancing the two components, the correlation between $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$ and $\mathbf{x}^{(j)}[\tau^{(j)}]$ can be close to 0. As a consequence,



(a) $\mathbf{x}^{(j)}[\tau^{(j)}]$ (b) $\tilde{\mathbf{m}}_{\text{pc}}^{(j)}$

(c) $\mathbf{x}^{(j)}[\tau^{(j)}] + \tilde{\mathbf{m}}_{\text{pc}}^{(j)}b[0]$

Fig. 3 Proposed PC signal vector.

$R\left(\mathbf{m}_{\mathrm{pc}}\right) = \|\mathbf{K}^H\tilde{\mathbf{m}}_{\mathrm{pc}}\|^2/\|\tilde{\mathbf{m}}_{\mathrm{pc}}\|\cdot\|\mathbf{m}_{\mathrm{pc}}\|$ can be increased, and the PAPR reduction effect after projection onto the null space in the MIMO channel can be enhanced. The proposed method also reduces the PAPR by directly reducing the variance in the average transmission signal power levels among transmitter antennas.

In the proposed method, the final PC signal vector, $\mathbf{m}_{\mathrm{pc}}^{(j)}$, is generated by applying the null restriction using (5) to the generated $\tilde{\mathbf{m}}_{\mathrm{pc}}^{(j)}$, which is the same as in the conventional PCCNC. Since the proposed method only adds the process of the minimum power threshold to the conventional method, the computational cost such as the required number of real multiplications per iteration for the proposed method is approximately the same as in the conventional method.

4. Numerical Results

The performance of the proposed PCCNC is evaluated based on computer simulations and compared to the conventional PCCNC [17]. The number of transmitter antennas, N_{tx}, is parameterized. The number of receiving user terminals, N_{rx}, is set to 4. The ZF-based BF is applied to actualize multiuser MIMO. The number of subcarriers in the OFDM signal is 64. The number of FFT/IFFT points is set to 256, which corresponds to 4-times oversampling in the time domain in order to measure satisfactorily accurate PAPR levels [23]. For general evaluation, we assume that the signal constellation of each subcarrier follows an independent standard complex Gaussian distribution. Flat Rayleigh fading is assumed as the channel model, which is independent between transmitter antennas and between receiving users. The signal-to-noise ratio (SNR) is set to 20 dB. In the proposed PCCNC and the conventional PCCNC, we assume that after J iterations, per-antenna PC-based PAPR reduction (PAPC) [20], [21] is applied with J_{add} iterations in order to achieve a lower PAPR at the cost of reduced throughput. Maximum power threshold P_{U} and minimum power threshold P_{L} are defined as the signal power threshold normalized by the signal power per antenna averaged over the channel realizations. The PAPR is defined as the ratio of the peak signal power to the average signal power across all the transmitter antennas per OFDM symbol. The sum throughput of N_{rx} streams (users) is measured based on the Shannon formula taking into account the Bussgang's theorem [24].

Figure 4 shows the average $R(\mathbf{m}_{\mathrm{pc}})$ as a function of P_{L} in the proposed PCCNC. The number of transmitter antennas, N_{tx}, is set to 100. Threshold P_{U} is 6 dB and iterations of the proposed PCCNC, J, is parameterized from 1 to 20. Overall, the average $R(\mathbf{m}_{\mathrm{pc}})$ increases by setting P_{L} to be larger than P_{L} of −100 dB, which corresponds to the conventional PCCNC. This suggests that the degradation in the PAPR reduction effect accompanying the projection processing onto the null space in the MIMO channel of the PC signal vector can be alleviated using an appropriate P_{L} in the proposed PCCNC compared to that in the conventional PCCNC. As J is increased, the P_{L} that maximizes the average $R(\mathbf{m}_{\mathrm{pc}})$ is decreased. This is because the transmission signal vector at J-th iteration $\mathbf{x}^{(J)}[t]$ contains the PC signal vectors that are added at up to $J-1$ iterations, which increases the correlation between null space $\mathbf{K}$ in the MIMO channel and $\mathbf{x}^{(J)}[t]$. Therefore, the necessity for the PC signal components to increase the power level of the signal below minimum power threshold P_{L} that has a positive correlation with $\mathbf{x}[t]$ decreases, and a lower P_{L} is considered to be optimum from the viewpoint of maximizing the average $R(\mathbf{m}_{\mathrm{pc}})$.

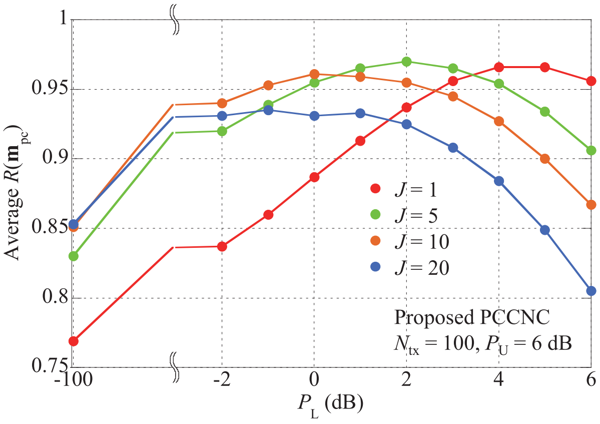


Fig. 4 Average $R(\mathbf{m}_{\mathrm{pc}})$ as a function of P_{L}.

To estimate the degree of improvement in the PAPR reduction effect per added PC signal in the proposed method, we evaluated the peak power reduction ratio, G_{peak}, which is defined as

$$G_{\mathrm{peak}} = \left(\frac{A_{\mathrm{after}} - \sqrt{P_{\mathrm{U}}}}{A_{\mathrm{before}} - \sqrt{P_{\mathrm{U}}}}\right)^2. \tag{10}$$

Here, A_{before} and A_{after} are the amplitudes of the peak signal component before and after addition of the PC signal at the J-th iteration, respectively. Ratio G_{peak} is a measure of how much the addition of the PC signal reduces the signal power above maximum power threshold P_{U}. Figure 5 shows the average G_{peak} as a function of P_{L} for the proposed PCCNC. Term N_{tx} is set to 100. Threshold P_{U} is 6 dB and J is parameterized. The average G_{peak} is decreased by appropriately setting the P_{L} levels for the respective J. From Figs. 4 and 5, we see that the best P_{L} level that maximizes the average $R(\mathbf{m}_{\mathrm{pc}})$ coincides with the P_{L} level that minimizes the average G_{peak} for the respective J. This confirms the validity of the analysis regarding the conditions for the PC signal vector to achieve a sufficient PAPR reduction capability after projection onto the null space in the MIMO channel in this paper.

In the proposed PCCNC, another factor that further reduces the PAPR compared to the conventional PCCNC is the reduction of the variance in the average transmission signal power among transmitter antennas after addition of the PC signal. To assess this effect, the reduction rate of the variance in the average transmission signal power among transmission antennas, G_{ant}, is evaluated. Term G_{ant} is defined as

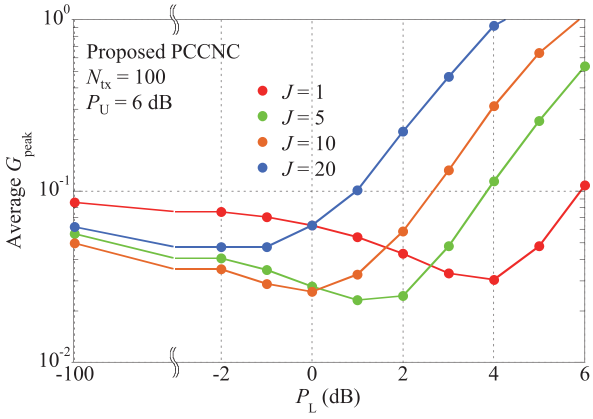


Fig. 5 Average G_{peak} as a function of P_{L}.

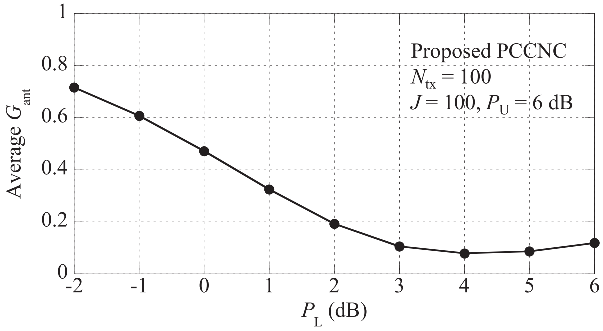


Fig. 6 Average G_{ant} as a function of P_{L}.

$$G_{\text{ant}} = \frac{\text{Var}\left[\overline{P}_n^{\text{prop.}}\right]}{\text{Var}\left[\overline{P}_n^{\text{conv.}}\right]}. \tag{11}$$

Here, $\text{Var}\left[\overline{P}_n^{\text{prop.}}\right]$ and $\text{Var}\left[\overline{P}_n^{\text{conv.}}\right]$ are the variance in the average transmission signal power among N_{tx} transmitter antennas when the proposed PCCNC and the conventional PCCNC are used, respectively. Figure 6 shows the average G_{ant} as a function of P_{L}. Term N_{tx} is set to 100. Number of iterations J is set to 100 and P_{U} is 6 dB. From Fig. 6, the variance in the average transmission signal power among transmitter antennas is suppressed by using the proposed PCCNC with an appropriate setting for P_{L}.

Figures 7 and 8 show the CCDF of the PAPR and throughput, respectively. The proposed PCCNC with P_{L} as a parameter and the conventional PCCNC are tested. Term N_{tx} is set to 100. Maximum power threshold P_{U} and J are set to 6 dB and 100, respectively. Term J_{add} is set to 60. Based on Fig. 7, the PAPR distribution of the proposed PCCNC is improved as P_{L} is set larger. This is because the proposed PCCNC increases the PAPR reduction effect after the projection of the PC signal onto the null space in the MIMO channel and reduces the average transmission power variance among the transmitter antennas. From Fig. 8, the throughput of the proposed PCCNC at the CCDF of 0.5 is most improved when P_{L} is set between −2 dB and 2 dB.

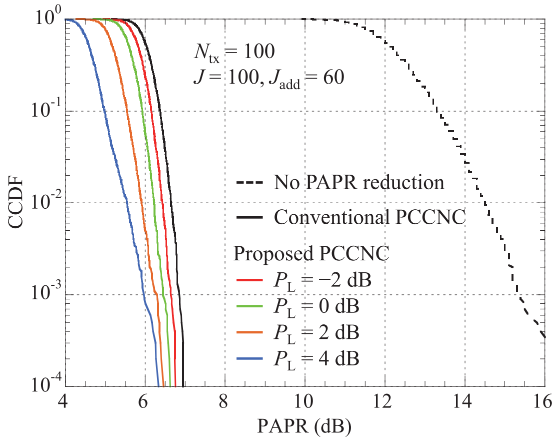


Fig. 7 CCDF of PAPR.

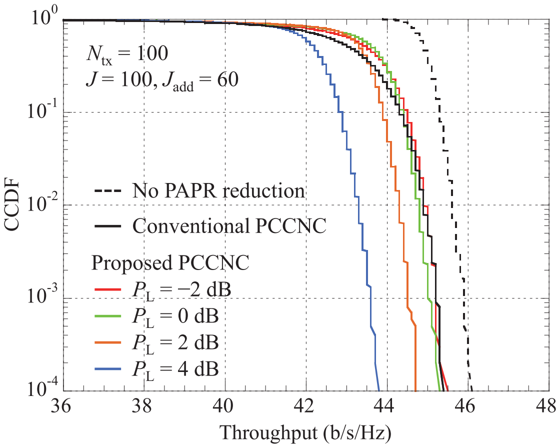


Fig. 8 CCDF of throughput.

With this P_{L} setting, the throughput with the proposed PCCNC is higher than that with the conventional PCCNC. This is because the proposed PCCNC can sufficiently suppress the peak power, and the subsequent PAPC does not need to generate much of a PC signal to suppress the residual peaks. Therefore, in the proposed method, interference caused by the PC signal of PAPC can be decreased compared to that for the conventional PCCNC. On the other hand, the throughput of the proposed PCCNC with the relatively high P_{L} value of 4 dB is degraded. This is because the transmission power for data signal transmission is reduced since a large fraction of the total transmission power is consumed for the PC signal transmission to increase the transmission power of the signal component whose power is below minimum power threshold P_{L}.

Figure 9 shows the average throughput as a function of the average PAPR. The proposed PCCNC with P_{L} as a parameter and the conventional PCCNC are tested. Term N_{tx} is set to 100. The numbers of iterations of PCCNC and PAPC, J and J_{add}, are set to 100 and 60, respectively, for both proposed and conventional PCCNCs. The relationship between the average PAPR and the average throughput is varied by changing P_{U} for the respective methods. Figure 9

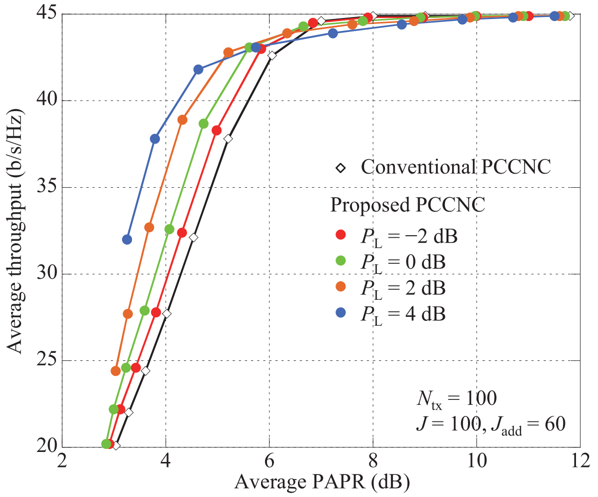


Fig. 9 Average throughput as a function of average PAPR.

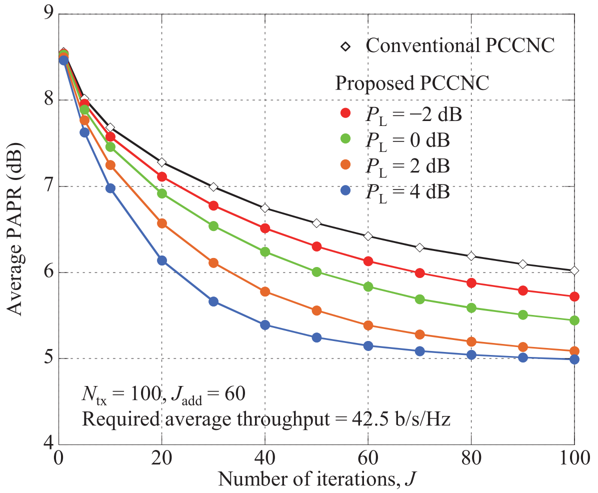


Fig. 10 Average PAPR for required average throughput as a function of J.

shows that the proposed method increases the throughput compared to that for the conventional PCCNC for the same required PAPR level, especially when the required PAPR is small. This is because the proposed method achieves a sufficient PAPR reduction more efficiently using the PC signal directed to the null space in the MIMO channel by decorrelating the PC signal and transmission data signal. Furthermore, the proposed method suppresses the variance in the average transmission signal power among transmitter antennas, which directly contributes to the reduction in PAPR.

Figure 10 shows the average PAPR for the required average throughput of 42.5 b/s/Hz as a function of the number of iterations, J. The proposed PCCNC with P_L as a parameter and the conventional PCCNC are tested. Term N_{tx} is set to 100. The numbers of iterations for PAPC, J_{add}, is fixed at 60 for all methods. For each J of the respective methods, P_U is adjusted so that the required average throughput of 42.5 b/s/Hz is satisfied. The proposed PCCNC decreases the average PAPR compared to that for the conventional PCCNC for the same J. This is because the proposed PCCNC achieves a superior PAPR reduction effect by decorrelating the PC signal and transmission data signal, which also contributes to reducing the variance in the average transmission signal power among the transmitter antennas. As mentioned in Sect. 3, the computational cost such as the required number of real multiplications per iteration for the proposed PCCNC is approximately the same as that for the conventional one, since the proposed PCCNC only adds the process of the minimum power threshold to the conventional one. Therefore, the reduction in the required number of iterations for achieving the same PAPR shown in Fig. 10 indicates that the proposed PCCNC reduces the computational cost compared to the previous PCCNC for achieving the same PAPR.

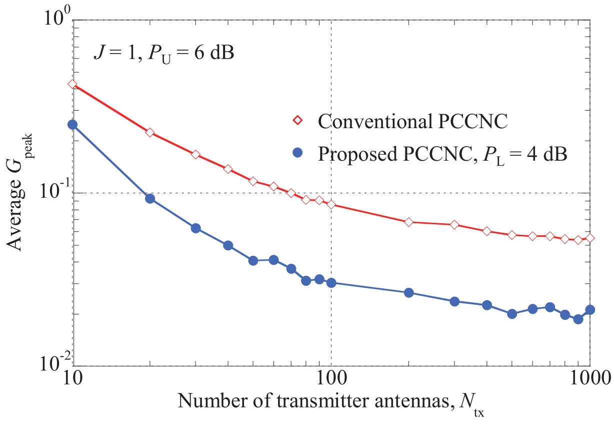


Fig. 11 Average G_{peak} as a function of N_{tx}.

In the following, performance evaluations with various N_{tx} are presented to assess the impact of the dimensions of the null space in the MIMO channel on the performance of the proposed method. Figure 11 shows the average G_{peak} as a function of N_{tx}. Terms J and P_U are set to 1 and 6 dB, respectively. The proposed PCCNC with the P_L value of 4 dB and conventional PCCNC are tested. Both the proposed and conventional PCCNCs achieve lower average G_{peak} as N_{tx} is increased, thanks to the increased dimensions of the null space in the MIMO channel. The proposed PCCNC achieves a lower G_{peak} than the conventional PCCNC regardless of the N_{tx} value.

Figure 12 shows the average G_{ant} as a function of N_{tx}. Term J is set to 100. Thresholds P_U and P_L are set to 6 dB and 4 dB, respectively. The average G_{ant} is reduced as N_{tx} is increased. Thus, the performance gain by using the proposed PCCNC in terms of the reduction in the variance in the average transmission signal power among transmitter antennas is increased as the dimensions of the null space in the MIMO channel increases.

Figure 13 shows the average throughput as a function of the average PAPR. Term N_{tx} is parameterized from 20 to 1000. The proposed PCCNC with the P_L value of 4 dB and conventional PCCNC are tested. The relationship between the average PAPR and the average throughput is varied by changing P_U for the respective methods. Terms J and J_{add}

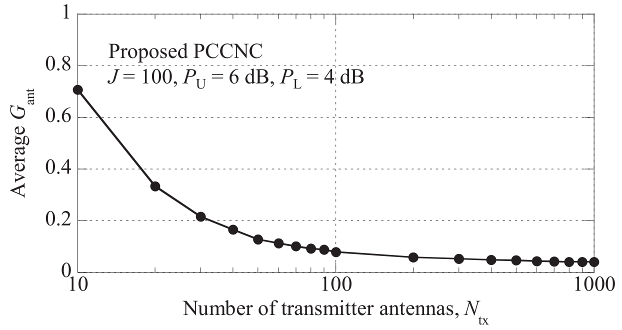


Fig. 12 Average G_{ant} as a function of N_{tx}.

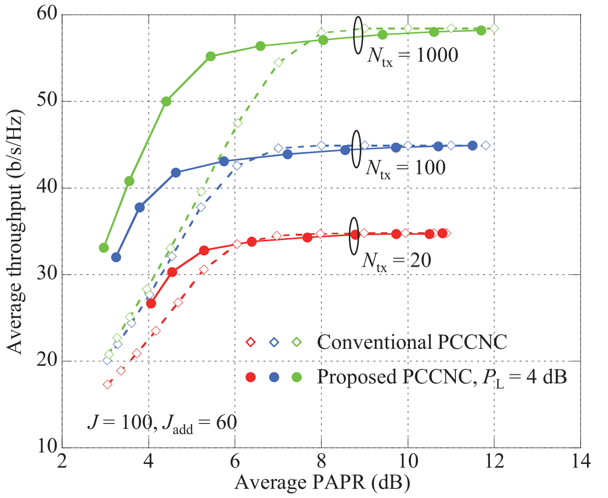


Fig. 13 Average throughput as a function of average PAPR for various N_{tx}.

are set to 100 and 60, respectively. Figure 13 shows that the proposed PCCNC increases the throughput compared to that for the conventional PCCNC for the same required PAPR level, especially when the required PAPR is small. The throughput gain by using the PCCNC compared to that for the conventional one is increased as N_{tx} is increased. This confirms that the proposed PCCNC is effective especially when the dimensions of the null space in the MIMO channel are large.

Figure 14 shows the average PAPR for the required average throughput as a function of N_{tx}. The required average throughput for each N_{tx} is defined as the 95% value of the achievable throughput when no PAPR reduction is performed. The proposed PCCNC with the P_L value of 4 dB and conventional PCCNC are tested. Terms J and J_{add} are set to 100 and 60, respectively, for both methods. The proposed PCCNC achieves a lower PAPR than that for the conventional one for all N_{tx}. The conventional PCCNC increases the PAPR as N_{tx} is increased. This is mainly due to the increased variance in the average transmission signal power among the transmitter antennas. On the other hand, the proposed PCCNC significantly reduces the PAPR by maintaining the PAPR reduction capability after projecting the PC signal vector onto the channel null and suppressing the variance in the average transmission signal power among the transmitter antennas when N_{tx} is very large.

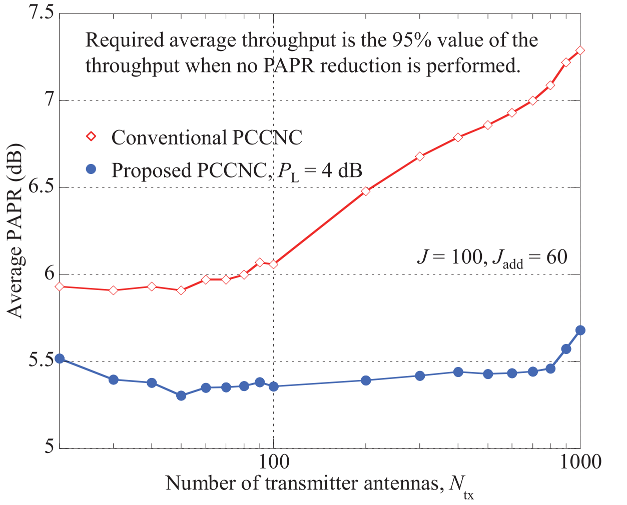


Fig. 14 Average PAPR for required average throughput as a function of N_{tx}.

5. Conclusion

This paper proposed a new generation method for the PC signal vector, which has a component that increases the transmission power of the transmission signal so that the transmission signal power levels for all transmitter antennas at the target timing can simultaneously be set equal to or higher than the minimum power threshold. This is completely different from the conventional standard concept of PAPR reduction, i.e., the clipping of the high peak power components. This new idea decorrelates the PC signal vector and transmit data signal vector, and this enables more PC signal components for PAPR reduction to be emitted to the null space of the given MIMO channel. Furthermore, the variance in the average transmission signal power levels among antennas, which is one cause for the PAPR increase in MIMO transmission with BF, can be suppressed. As a result, the proposed method achieves a greater PAPR reduction effect than the conventional method. This paper assumes a frequency nonselective channel. The extension of the proposed method to accommodate a frequency selective channel is possible by utilizing the approach in [18] for example. Detailed investigation on extending the proposed method to accommodate a frequency selective channel as well as investigating a more power efficient construction for the PC signal vector are left for future study.

References

[1] T.L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," IEEE Trans. Wireless Commun., vol.9, no.11, pp.3590–3600, Nov. 2010.

[2] H. Papadopoulos, C. Wang, O. Bursalioglu, X. Hou, and Y. Kishiyama, "Massive MIMO technologies and challenges towards 5G," IEICE Trans. Commun., vol.E99-B, no.3, pp.602–621, March

2016.

[3] C. Studer and E.G. Larsson, "PAR-aware large-scale multi-user MIMO-OFDM downlink," IEEE J. Sel. Areas Commun., vol.31, no.2, pp.303–313, Feb. 2013.

[4] H. Prabhu, O. Edfors, J. Rodrigues, L. Liu, and F. Rusek, "A low-complex peak-to-average power reduction scheme for OFDM based massive MIMO systems," Proc. ISCCSP2014, Athens, Greece, May 2014.

[5] M. Iwasaki and K. Higuchi, "Clipping and filtering-based PAPR reduction method for precoded OFDM-MIMO signals," Proc. IEEE VTC2010-Spring, Taipei, Taiwan, May 2010.

[6] Y. Sato, M. Iwasaki, S. Inoue, and K. Higuchi, "Clipping and filtering-based adaptive PAPR reduction method for precoded OFDM-MIMO signals," IEICE Trans. Commun., vol.E96-B, no.9, pp.2270–2280, Sept. 2013.

[7] S. Inoue, T. Kawamura, and K. Higuchi, "Throughput/ACLR performance of CF-based adaptive PAPR reduction method for eigenmode MIMO-OFDM signals with AMC," IEICE Trans. Commun., vol.E96-B, no.9, pp.2293–2300, Sept. 2013.

[8] R. Kimura, Y. Tajika, and K. Higuchi, "CF-based adaptive PAPR reduction method for block diagonalization-based multiuser MIMO-OFDM signals," Proc. IEEE VTC2011-Spring, Budapest, Hungary, May 2011.

[9] Y. Matsumoto, K. Tateishi, and K. Higuchi, "Performance evaluations on adaptive PAPR reduction method using null space in MIMO channel for eigenmode massive MIMO-OFDM signals," Proc. APCC2017, Perth, Australia, Dec. 2017.

[10] A. Ivanov, A. Volokhatyi, D. Lakontsev, and D. Yarotsky, "Unused beam reservation for PAPR reduction in massive MIMO system," Proc. IEEE VTC2018-Spring, Porto, Portugal, June 2018.

[11] R. Zayani, H. Shaiek, and D. Roviras, "PAPR-aware massive MIMO-OFDM downlink," IEEE Access, vol.7, pp.25474–25484, Feb. 2019.

[12] S. Shin, N. Nonaka, and K. Higuchi, "User group selection method in multiuser MIMO-OFDM transmission with adaptive PAPR reduction using null space in MIMO channel," Proc. IEEE VTC2020-Fall, Virtual Conference, Nov.-Dec. 2020.

[13] L. Hua, Y. Wang, Z. Lian, Y. Su, and Z. Xie, "Low-complexity PAPR-aware precoding for massive MIMO-OFDM downlink systems," IEEE Wireless Commun. Lett., vol.11, no.7, pp.1339–1343, July 2022.

[14] Y. Sekiguchi, N. Nonaka, and K. Higuchi, "PAPR reduction of OFDM signals using null space in MIMO channel for MIMO amplify-and-forward relay transmission," IEICE Trans. Commun., vol.E105-B, no.9, pp.1078–1086, Sept. 2022.

[15] A. Kakehashi, N. Nonaka, and K. Higuchi, "PAPR reduction using null space in MIMO channel based on signal processing at base station for downlink AF-based relaying MIMO-OFDM signals," Proc. IEEE VTC2022-Fall, London and Beijing, Sept. 2022.

[16] A. Kakehashi, T. Hara, and K. Higuchi, "Base station-driven PAPR reduction method utilizing null space for MIMO-OFDM systems with amplify-and-forward relaying," IEEE Access, vol.12, pp.24714–24724, Feb. 2024.

[17] T. Suzuki, M. Suzuki, Y. Kishiyama, and K. Higuchi, "Complexity-reduced adaptive PAPR reduction method using null space in MIMO channel for MIMO-OFDM signals," IEICE Trans. Commun., vol.E103-B, no.9, pp.1019–1029, Sept. 2020.

[18] L. Yamaguchi, N. Nonaka, and K. Higuchi, "PC-signal-based PAPR reduction using null space in MIMO channel for MIMO-OFDM signals under frequency-selective fading channel," Proc. IEEE VTC2020-Fall, Virtual Conference, Nov.-Dec. 2020.

[19] T. Suzuki, M. Suzuki, and K. Higuchi, "Parallel peak cancellation signal-based PAPR reduction method using null space in MIMO channel for MIMO-OFDM transmission," IEICE Trans. Commun., vol.E104-B, no.5, pp.539–549, May 2021.

[20] T. Hino and O. Muta, "Adaptive peak power cancellation scheme under requirements of ACLR and EVM for MIMO-OFDM systems," Proc. IEEE PIMRC2012, Sydney, Australia, Sept. 2012.

[21] T. Kageyama, O. Muta, and H. Gacanin, "An adaptive peak cancellation method for linear-precoded MIMO-OFDM signals," Proc. IEEE PIMRC2015, Hong Kong, Aug.-Sept. 2015.

[22] J. Saito, N. Nonaka, and K. Higuchi, "PAPR reduction using null space in MIMO channel considering signal power difference among transmitter antennas," Proc. IEEE WCNC2023, Glasgow, Scotland, UK, March 2023.

[23] M. Sharif, M. Gharavi-Alkhansari, and B.H. Khalaj, "On the peak-to-average power of OFDM signals based on oversampling," IEEE Trans. Commun., vol.51, no.1, pp.72–78, Jan. 2003.

[24] H. Ochiai and H. Imai, "Performance analysis of deliberately clipped OFDM signals," IEEE Trans. Commun., vol.50, no.1, pp.89–101, Jan. 2002.

Jun Saito received the B.E. and M.E. degrees from Tokyo University of Science, Noda, Japan in 2022 and 2024, respectively. In 2024, he joined IBM Japan, Ltd. His research interests include wireless communications.

Nobuhide Nonaka received the B.E. and M.E. degrees in electronic engineering from Tokyo University of Science, Japan, in 2013 and 2015, respectively. Since April 2015, he has been with NTT DOCOMO, INC. Since April 2018, he has been engaged in the research of next generation radio access technologies. He received the Young Researcher's Award from the IEICE in 2019. He is a member of the IEICE.

Kenichi Higuchi received the B.E. degree from Waseda University, Tokyo, Japan, in 1994, and received the Dr.Eng. degree from Tohoku University, Sendai, Japan in 2002. In 1994, he joined NTT Mobile Communications Network, Inc. (now, NTT DOCOMO, INC.). While with NTT DOCOMO, INC., he was engaged in the research and standardization of wireless access technologies for wideband DS-CDMA mobile radio, HSPA, LTE, and broadband wireless packet access technologies for systems beyond IMT-2000. In 2007, he joined the faculty of the Tokyo University of Science and currently holds the position of Professor. His current research interests are in the areas of wireless technologies and mobile communication systems, including advanced multiple access such as non-orthogonal multiple access (NOMA), radio resource allocation, inter-cell interference coordination, multiple-antenna transmission techniques, signal processing such as interference cancellation and turbo equalization, and issues related to heterogeneous networks using small cells. He was a co-recipient of the Best Paper Award of the International Symposium on Wireless Personal Multimedia Communications in 2004 and 2007, the Best Paper Award from the IEICE in 2021, a recipient of the Young Researcher's Award from the IEICE in 2003, the 5th YRP Award in 2007, the Prime Minister Invention Prize in 2010, and the Invention Prize of Commissioner of the Japan Patent Office in 2015. He is a senior member of the IEEE.

PAPER

Global Navigation Satellite System Signal Phase Combining and Performance of Distributed Antenna Arrays

Wenfei GUO[†,††], Jun ZHANG[†], Chi GUO[†a)], *and* Weijun FENG[†], *Nonmembers*

SUMMARY Low signal power and susceptibility to interference cause difficulties for traditional global navigation satellite system (GNSS) receivers in tracking weak signals. Extending coherent integration time is a common approach for enhancing signal gain. However, coherent integration time cannot be indefinitely increased owing to navigation bit sign transition, receiver dynamics, and clock noises. This study proposes a cross-correlation phase combining (CPC) algorithm suitable for distributed multi-antenna receivers to improve carrier-tracking performance in weak GNSS signal conditions. This algorithm cross-correlates each antenna's intermediate frequency (IF) signal and local carrier to detect the phase differences. Subsequently, the IF signals are weighted to achieve phase alignment and coherently combined. The carrier-to-noise ratio (CNR) and carrier phase equation of the combined signal were derived for the CPC algorithm. Global positioning system (GPS) signals received by distributed antenna array with six elements were used to validate the performance of the algorithm. The results demonstrated that the CPC algorithm could effectively achieve signal phase alignment at 32 dB-Hz, resulting in a combined-signal CNR enhancement of 6 dB. The phase-tracking error variance was reduced by 72% at 30 dB-Hz compared with that of a single-antenna signal. The algorithm exhibited low phased array calibration requirements, overcoming the limitations associated with coherent integration time and effectively enhancing tracking performance in weak-signal environments.

key words: *weak GNSS signal, coherent integration, distributed antenna arrays, multi-antenna signal combining*

1. Introduction

Global navigation satellite system (GNSS) provides users with positioning, navigation, and timing functions. GNSS has steadily developed over the past decades, and satellite navigation receiver applications have increased across various fields [1]. However, the satellite navigation signal power reaching the Earth's surface is weak owing to high free-space loss and attenuation while traveling through the atmosphere and ionosphere. Even in open outdoor environments, the signal strength is only approximately −130 dBm [2]. Receivers exhibit a decline in signal power in harsh environments. For example, trees in forests and buildings in urban areas obstruct signals and weaken signal strength. In addition, these signals may be subjected to cross-correlation interference from powerful satellites or interfering signals [3] and multipath interference [4]. In such environments, the received signal-to-noise ratio (SNR) can be attenuated by 10–30 dB [5], resulting in severe signal deterioration.

The low signal power in weak-signal environments causes the SNR of the signal input to the tracking loop phase discriminator to be extremely low, leading to considerable errors in the output of the carrier-tracking loop or even a loss of lock. Conventional carrier-tracking loops cannot adapt to weak-signal environments. Local and international researchers have extensively studied high-sensitivity carrier-tracking technologies, including loop optimization, coherent integration extension, external assistance, vector-tracking loop, and high-sensitivity tracking algorithms [6], [7]. Coherent integration extension is the most direct and commonly used method for enhancing SNR gain. However, this approach has various limitations, such as the bit sign transition of navigation data, receiver and satellite dynamics, and frequency stability of crystal oscillators [8]–[10]. For the commonly used global positioning system (GPS) L1 C/A signal, noncoherent integration can eliminate navigation data but introduces quadratic loss [1]. The acceleration caused by the receiver and satellite motions and the frequency drift of the crystal oscillator of the receiver limit the coherent integration gain. Scholars have proposed methods for estimating satellite dynamics based on satellite orbit approximation [11]. The receiver dynamics and frequency drift of the crystal oscillator are crucial for long-term coherent integration, significantly affecting the performance of the receiver [12].

Compared with single-antenna signals, combining antenna array signals provides numerous advantages, such as stable system performance, effective enhancement of signal reception quality, high antenna utilization, cost savings, and ease of maintenance [13]. The wireless channel is coherent for satellite navigation signals. This implies that the antenna array elements receiving the signal differ only in amplitude and phase when multipath effects are disregarded [14], [15]. Although the signal is distorted owing to the nonlinearity of the satellite devices, this distortion does not change the coherence of the wireless channel itself. Because the signals received by individual antenna array elements are correlated, and the noise remains mutually independent, the signals from different propagation channels can be constructively combined in the spatial domain. Adjusting the array signal phase weights and coherently adding these signals leads to signal gain, enhancing the SNR [16].

Array antenna systems are typically categorized as phased or distributed. Unlike their phased array counterparts, distributed array antennas do not have strict half-

Manuscript received November 28, 2023.
Manuscript revised February 23, 2024.
Manuscript publicized May 14, 2024.

[†] Wuhan University, Wuhan, 430072, China.
[††] Institute of Electronic Information, Wuhan Qingchuan University, Wuhan, 430204, China.
a) E-mail: guochi@whu.edu.cn

DOI: 10.23919/transcom.2023EBP3191

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

wavelength requirements for the position of the array elements. The gain of the signal combined in the distributed array antennas is unrelated to the arrival direction, but only to the received SNR and carrier phase. This phenomenon confers distributed arrays with heightened flexibility, maneuverability, and a high gain advantage [17]. Currently, the algorithms for signals combining distributed array antennas include SIMPLE [18], SUMPLE [19], EIGEN [20], and LSFIT [21]. Although the LSFIT algorithm boasts minimal combining loss, it requires cross-correlation functions among all antenna signals, resulting in the calculation amount being proportional to the square of the antenna count. By contrast, the SUMPLE algorithm requires a number of calculations proportional to the number of antennas. This value closely matches that of the SIMPLE algorithm; however, it delivers a performance comparable to that of the LSFIT algorithm. Therefore, the SUMPLE algorithm is widely used for combining multi-antenna signals.

Rogstad originally introduced the SUMPLE algorithm in 2005. Subsequently, he described the SUMPLE algorithm comprehensively and analyzed its performance from various aspects. Xu et al. conducted a comparative study between the SUMPLE and SIMPLE algorithms and found the SUMPLE algorithm superior to the SIMPLE algorithm [22]. Chen et al. used a minimum mean-squared error estimator to amplify the performance of the SUMPLE algorithm [23]. Wang et al. improved the combined performance of the SUMPLE algorithm by identifying and excluding low-quality signal paths by assessing gain factors [24]. Yan et al. applied the SUMPLE algorithm to large-scale antenna arrays on a CPU-GPU heterogeneous system without experimental verification or algorithm performance improvement [25]. Li et al. enhanced the SNR of the input signal by resampling and used the SUMPLE algorithm for signal combination but without any algorithm enhancements [26]. These studies have extensively analyzed and improved the SUMPLE algorithm from different perspectives and applied it to various domains; however, the application of the algorithm in GPS systems remains unexplored. Currently, the G-STAR developed by Lockheed Martin [27]and the Integrated GPS Anti-Interference System (IGAS) developed by Rockwell Collins [28] are both mature satellite navigation antenna array receivers. However, the digital signal combining technology that has been maturely applied in existing products is non-blind beamforming. Non-blind beamforming requires the use of external information sources such as inertial navigation systems (INS) to calibrate the position and elevation of the antenna array [29]. However, in normal scenarios, it is difficult to bear the cost of adding an inertial navigation system. Therefore, a signal-combining technology suitable for distributed arrays that does not require DOA information and reducing dependence on external information sources is an important development trend for GPS signal combing.

This study introduced a cross-correlation phase combining (CPC) algorithm to address the difficulty of GPS tracking in a weak-signal environment and overcome the challenges of existing methods, primarily single-antenna GPS systems. During the GPS signal tracking process, the phase-locked loop (PLL) generates a local carrier matching the signal carrier of the antenna. This local carrier is considered as a beneficial factor because it can be employed as a stable phase input to enhance the phase stability of the combined signal and can also serve as an independent signal input for phase-offset determination, enhancing accuracy. In this algorithm, one antenna signal of the array is selected in turn, and the weighted sum of the remaining antenna signals and local carrier are the reference signals. The selected antenna and reference signals are cross-correlated to calculate the phase offset between them, which is then used for phase compensation of the antenna signal. Owing to the inclusion of a fixed-phase component of the local carrier in the reference signal, the antenna signal phase converges toward the local carrier. In the next iteration, as the signal tends to be coherently added, the SNR of the reference signal increases, and the phase offset compensation value approaches the actual value. Consequently, after a sufficient number of iterations, the combined-signal phase gradually aligns with the local carrier phase.

This study analyzed the CNR and carrier phase of the signals using the CPC algorithm. GPS signals received by distributed antenna array with six elements were used to evaluate the performance of the algorithm. The remainder of this paper is structured as follows: Section 2 discusses the distributed antenna array signal model; Section 3 elaborates on the CPC algorithm and introduces the structure of the multi-antenna signal-combining system; Section 4 analyzes the CNR and carrier phase of the combined signals using the CPC algorithm, delineates the factors influencing the signal-combining performance, and presents the simulation experiment; Section 5 presents the simulation and performance analysis results; and Sect. 6 summarizes the conclusions.

2. Distributed Array Antenna Signal and System Model

Figure 1 shows the structure of the signal combination system. Consider L antennas in a distributed antenna system situated at diverse locations but working in the same frequency band. Generally, the distance between the distributed antenna array elements is greater than half the wavelength, and the antenna is freely distributed in three-dimensional (3D) space [30].

In practice, factors, such as various spatial positions of antennas in a distributed array and inconsistencies in the phase attributes of the radio frequency front-end (RFFE) and signal transmission channels, contribute to the phase offsets among signals received by different antennas. Therefore, before signal combination, a phase alignment should be performed to increase signal coherence. Phase alignment involves using multi-antenna signal correlation processing to estimate the phase offsets among distinct intermediate frequency (IF) signals and achieve signal carrier phase consistency through the phase rotation factor. The phase rotation

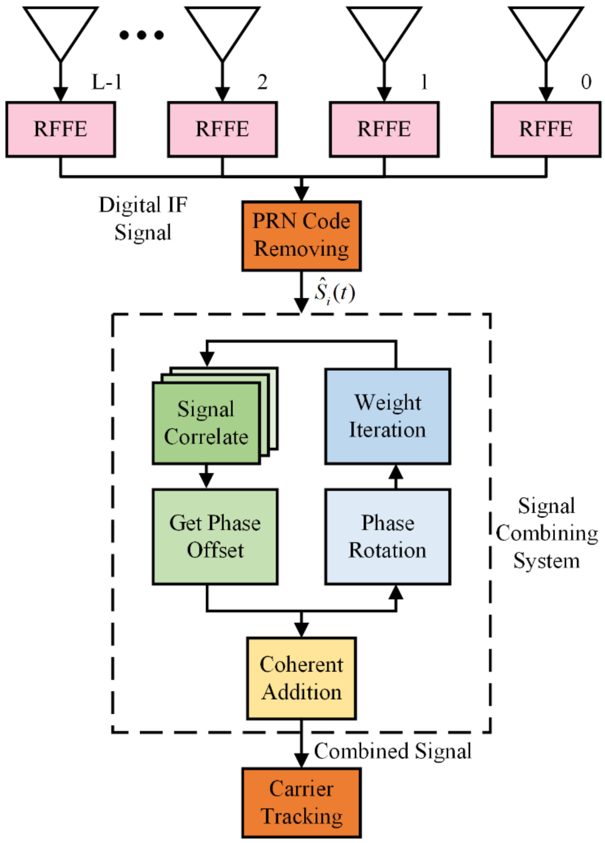


Fig. 1 Array antenna signal-combining system.

factor is defined as the phase weight and is iteratively updated continuously. Section 3 presents a detailed analysis. Upon successful phase alignment, the coherently combined signals are output to the PLL loop, and the combined signal is carrier-tracked.

The GPS signal received by the antenna includes signals from all satellites in the sky. Thus, when estimating the phase offset of a particular satellite signal through signal correlation, other satellite signals are considered as interference. This can seriously affect the accuracy of the phase-offset estimation. Therefore, in this study, the signal-combining system processed the pseudo-code-stripped IF signal. Pseudocode is stripped through native pseudocode. The local pseudocode, like the local carrier, is fed back from the subsequent tracking process. The IF signals addressed in this study only include navigation signals received by antenna from a single satellite signal.

After down conversion by the RFFE and the removal of the pseudo-random noise codes, the digital IF signal from antenna 0 can be expressed as follows:

$$S_0(t) = \alpha_0 D_0(t) sin\,(\omega_{IF} * t + \theta_0) + n_0(t), \tag{1}$$

where α_0 represents the amplitude coefficient of the received signal, $D_0(t)$ represents the data level containing the navigation information, ω_{IF} represents the carrier frequency after mixing under RFFE, θ_0 represents the carrier phase of the *0-th* antenna, $n_0(t)$ represents the noise component in the reference signal. Considering the phase offsets between the various antenna signals relative to the reference antenna, the received signals from the remaining $L-1$ antennas can be expressed as follows:

$$\begin{aligned} S_i(t) &= \alpha_i D_i(t) sin\,(\omega_{IF} * (t + \tau_i) + \theta_0) + n_i(t), \\ &\quad i = 1, 2, \ldots, L-1, \end{aligned} \tag{2}$$

where the subscript i denotes the antenna number, and τ_i represents the time delay of the signal from the $L-1$ antennas relative to the reference antenna signal; the time delay τ_i encompasses differences in antenna circuitry, RFFE channel hardware delays, and transmission path delays due to varying antenna positions, leading to IF carrier phase offset. No correlation was assumed between the navigation signal and noise among the various antennas.

Complex signals are vital to signal-combining systems owing to the need for cross-correlation operations. By modifying Euler's formula, the received signal of the entire distributed antenna system can be defined as follows [31]:

$$\hat{S}_i(t) = \hat{s}_i(t) e^{j\omega_{IF}\tau_i} + \hat{n}_i(t), \tag{3}$$

where the hat symbol $\widehat{(\cdot)}$ above the signal and noise indicates that they are complex signals, and $\hat{s}_i(t) = \alpha_i D_i(t) e^{j[\omega_{IF}*t+\theta_0]}$ represents the signal component. Without signal tracking, prior information on the satellite signal incident angle, and uncertainty in amplitude and phase caused by antenna deployment and radio frequency channels, the phase offsets result in an inequality in the carrier phase among various antenna signals, resulting in a CNR loss [32], [33]. Therefore, obtaining phase offsets among various antenna signals and compensating for this difference to align the signals are crucial for signal phase combination.

3. Cross-Correlation Phase Combining Algorithm

Aligning the phase delays of signals is essential for coherently combining antenna signals in a distributed antenna array system. When estimating the phase offsets between signals in an antenna array, the relative phase offset can be estimated by correlating the signals. If the CNR of each antenna signal is sufficiently high, all antenna pairs can exhibit strong correlations, no special processing will be required, and the phase offsets derived from the correlation can be directly used to align the signal [13]. By contrast, when the CNR is low, other methods must be employed. For example, If the user is in an urban area, the geometric structure of the building, wall thickness, etc. can attenuate the signal by 10–25 dB, and the signal power attenuation caused by the indoor environment can reach 25–30 dB [34]. In weak signal environment, it is necessary to use all possible antenna pairs to improve the phase alignment performance, reducing the CNR loss caused by phase estimation errors.

Figure 2 shows the structure of the multi-antenna signal-combining system. The IF signal represents the IF digital signal that completes the pseudo-code stripping. The

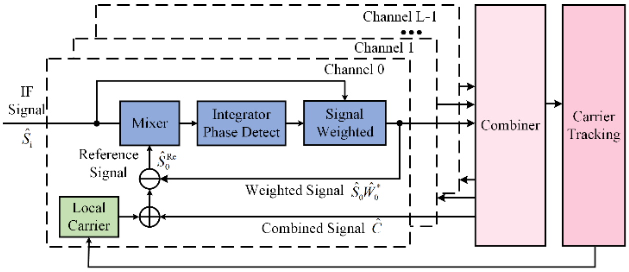


Fig. 2 Structure of the multi-antenna signal-combining system.

IF signal of each channel is cross-correlated with the reference signal, and the phase detector outputs the phase offset between the two, which is the phase weight. The IF signal is phase-compensated through the phase weight, and then the signals of all the antennas are input to the combiner and added. The reference signal is generated by adding the combined signal and local carrier and then subtracting the IF signal. The local carrier of each channel is consistent and fed back by the PLL. Following multiple iterations, the phase detector output gradually converges to zero, completing the phase alignment. After the phase alignment, a coherent combination enhances the combined-signal CNR. The calculation amount of the CPC algorithm is proportional to the number of array elements, requires L correlators, and is slightly more complex than the SUMPLE algorithm. Compared with the SUMPLE algorithm, the CPC algorithm requires L additional adders.

Figure 3 shows a flowchart of the CPC algorithm. For each antenna channel, the antenna signal is correlated with the reference signal to generate phase weights. Each L-antenna IF signal sequentially performs correlation and weighting operations, and the weighted IF signals are coherently combined to form an iteration. The phase weights $\hat{W}_i$ begin with a phase-zero unit vector. After each iteration, a new weight replaces the previous weight, and this process is repeated. The algorithm converges after several iterations. After achieving phase alignment for all antenna signals, they are coherently combined to produce the output.

In the CPC algorithm, the local carrier enhancing the CNR of the reference signal and improving algorithm performance. Simultaneously, during the signal phase alignment process, the phase center of the local carrier remains unchanged, no phase jitter occurs, and the local carrier participates in all reference signals, which is equivalent to introducing a fixed component into the reference signal. Throughout each iteration, the process gradually aligns the phase of each antenna signal toward this fixed component, ultimately converging to it. Thus, the CPC algorithm effectively estimates the phase offset and retains the advantages of the conventional algorithms.

From Eq. (3), the *i-th* antenna signal of the antenna array and weight phase in the correlation-weighting process can be expressed as follows:

$$\hat{S}_i(t) = \hat{s}_i(t) + \hat{n}_i^s(t), \tag{4}$$

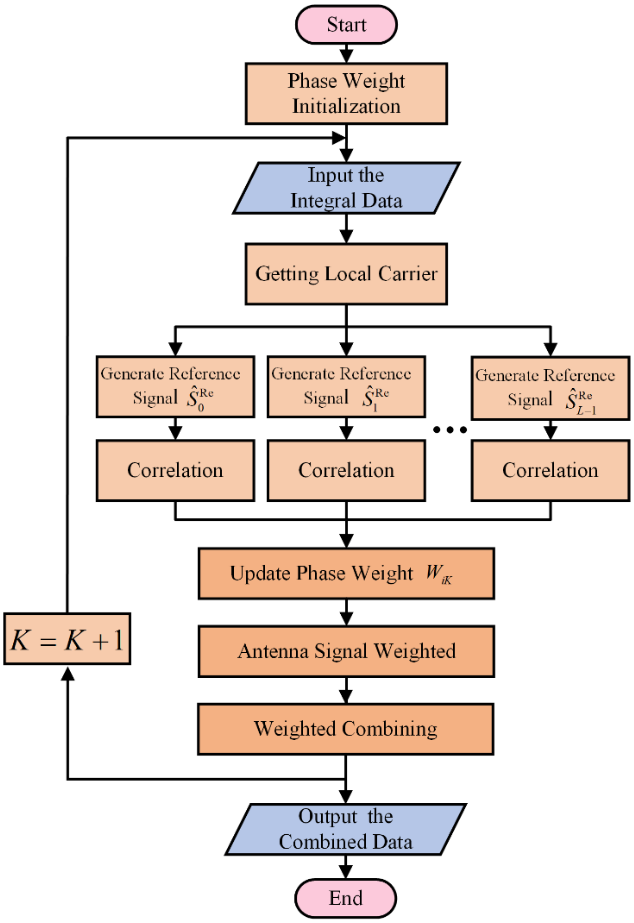

Start
Phase Weight Initialization
Input the Integral Data
Getting Local Carrier
Generate Reference Signal $\hat{S}_0^{Re}$
Generate Reference Signal $\hat{S}_1^{Re}$
Generate Reference Signal $\hat{S}_{L-1}^{Re}$
Correlation
Correlation
Correlation
Update Phase Weight W_{iK}
$K = K + 1$
Antenna Signal Weighted
Weighted Combining
Output the Combined Data
End

Fig. 3 Flowchart of the CPC algorithm.

$$\hat{W}_{iK} = \hat{w}_{iK} + \hat{n}^w_{iK}, \tag{5}$$

where $\hat{s}_i(t)$ and $\hat{n}_i^s$ represent the signal and noise components, respectively. $\hat{W}_{iK}$ represents the phase weight within an iteration. The number of correlated sampling points is the correlation averaging time, recorded as N, and the subscript K represents a time variable measured in the correlation time interval N units. Furthermore, $\hat{w}_{iK}$ represents the ideal weighting coefficient, and $\hat{n}^w_{iK}$ accounts for the estimation error of the weighting coefficient caused by the correlated noise.

Equation (2) and Eq. (3) indicate the existence of a time delay τ_i between the signals of different antenna array elements. This delay results in a phase offset $\omega_{IF}\tau_i$ among the carrier phases. The phase weight compensates for this offset such that the signal carrier phases of each array element are aligned. Therefore, weight can be expressed as follows:

$$\hat{w}_i = e^{j\omega_{IF}\tau_i} = e^{-j\Delta\theta_i} = e^{-j(\theta_i-\theta_L)}, \tag{6}$$

where θ_i represents the carrier phase of antenna i, and θ_L represents the local carrier phase and is the common phase of the carrier after the phase alignment is completed. $\Delta\theta_i = \theta_i - \theta_L$ represents the phase offset between the antenna signal and local carrier.

All antennas underwent correlation weighting, and the antenna signals were coherently combined using a combiner. The combined signal of the antenna array can be expressed as follows:

$$\hat{C}_K = \sum_{i=0}^{L-1} \hat{S}_{iK} \hat{W}_{iK}^{*}, \tag{7}$$

where $\hat{C}_K$ is the combined output signal, L represents the number of elements of the antenna array, $\hat{S}_{iK}$ represents the satellite navigation signal with N sampling intervals as the time unit, the subscript K represents the signal at intervals of the *K-th* unit, $(\cdot)^*$ represents the complex conjugate of a complex signal.

As illustrated in Fig. 2 and proposed by Rogstad [19], the weight of the $K+1$-*th* unit, denoted as $\hat{W}_{iK+1}$, is updated and recursively obtained from the weight of the previous step $\hat{W}_{iK}$. The weight is calculated by correlating the reference signal $\hat{S}_{iK}^{Re}$ with $\hat{S}_{ik}\hat{W}_{iK}^{*}$ over a length of N, as follows:

$$\hat{W}_{iK+1} = \hat{W}_{iK} \left\{ \frac{1}{N} \sum_{k=KN}^{(K+1)N-1} \left[\hat{S}_{ik} \hat{W}_{iK}^{*} \hat{S}_{iK}^{Re} \right] \right\}, \tag{8}$$

$$\hat{S}_{iK}^{Re} = \hat{C}_k - \hat{S}_{ik} \hat{W}_{iK}^{*} + \hat{S}_{k}^{L}, \tag{9}$$

where $\hat{S}_{ik}$ represents the satellite navigation signal with the signal-sampling interval as the unit interval. The satellite navigation signal is an IF signal output by the RFFE, in which parameters such as carrier phase and carrier frequency are unknown, and are subsequently estimated by the PLL loop. $\hat{S}_{k}^{L}$ represents the local carrier signal.

Equation (7) shows that $\hat{S}_i$, L, and $\hat{W}_{iK}$ influence the combined signal. Equation (8) indicates that $\hat{W}_{iK}$ is related to N and $\hat{S}_{iK}^{Re}$. Thus, an analysis of the relationships between $\hat{C}_K$, $\hat{S}_{ik}$, L, and N is necessary.

4. Performance Analysis of Multi-Antenna Signal Combining

These principles reveal that the antenna signal CNR, number of antennas, and correlation time interval affect the combined-signal gain. The following presents a theoretical analysis of the CNR and carrier phase of the combined signal of the CPC algorithm compared with those of the conventional algorithm.

4.1 Analysis of the Combined-Signal CNR

The CPC algorithm primarily estimates the phase offset among the signals from various antennas in a distributed antenna array system owing to the difference in the transmission path. Compensating the offset through phase weights ensures phase alignment between all antenna signals, improving the correlation between the antenna signals and coherent additions to enhance the CNR of the combined signal. Because the capability of the algorithm to improve the CNR of the combined signal is a crucial evaluation index, this subsection presents an analysis focusing on the CNR of the combined signal.

To analyze the CNR of the combined signal, if no error is observed in the weight estimation of the signal combination, the optimal combination performance can be achieved. Equation (7) can be expressed as follows:

$$\hat{C}_K = \sum_{i=0}^{L-1} \left[\hat{s}_{iK} e^{-j\Delta\theta_i} + \hat{n}_{iK}^{s} e^{-j\Delta\theta_i} \right]. \tag{10}$$

In practice, the weight inevitably contains errors owing to the influence of noise, reducing the output CNR of the combined signal. Assumptions: (1) Each antenna has the same caliber and performance, (2) the received signal power remains stable, (3) Each antenna signal is aligned (affected by noise; the weights may not be optimal) and (4) the weights of each antenna signal are mutually uncorrelated. From Eq. (5), Eq. (10) can be divided into its signal and noise components. The average of the combined-signal power is P_{CS}, expressed as follows:

$$\begin{aligned} P_{CS} &= E\left| \sum_{i=0}^{L-1} \left[\hat{s}_{iK} e^{-j\Delta\theta_i} + \hat{s}_{iK} \hat{n}_{iK}^{w\,*} \right] \right|^2 \\ &= L^2 \left|\hat{s}_0\right|^2 + L\left|\hat{s}_0\right|^2 \left|\hat{n}^w\right|^2, \end{aligned} \tag{11}$$

where $|\hat{s}_0|^2$ and $|\hat{n}^w|^2$ represent the average power of the antenna signal and weight noise components, respectively. According to assumptions (1) and (2), the average power of the *K-th* unit can be replaced by the first moment of the *K-th* unit, $\hat{s}_{iK} = \hat{s}_0$. The average power of the combined noise component P_{CN} is expressed as

$$\begin{aligned} P_{CN} &= E\left| \sum_{i=0}^{L-1} \left[\hat{n}_{iK}^{s} e^{-j\Delta\theta_i} + \hat{n}_{iK}^{s} \hat{n}_{iK}^{w\,*} \right] \right|^2 \\ &= L\left|\hat{n}_0^s\right|^2 + L\left|\hat{n}_0^s\right|^2 \left|\hat{n}^w\right|^2, \end{aligned} \tag{12}$$

where $\left|\hat{n}_0^s\right|^2$ is the average power of the noise component of the antenna signal, $\hat{n}_{iK}^{s} = \hat{n}_0^s$. Therefore, the SNR of the combined signal can be expressed as follows:

$$\rho_c = \frac{P_{CS}}{P_{CN}} = \frac{L^2 \left|\hat{s}_0\right|^2 + L\left|\hat{s}_0\right|^2 \left|\hat{n}^w\right|^2}{L\left|\hat{n}_0^s\right|^2 + L\left|\hat{n}_0^s\right|^2 \left|\hat{n}^w\right|^2}. \tag{13}$$

Equation (13) can be simplified as follows:

$$\rho_c = \frac{L\rho_s + \rho_s \left|\hat{n}^w\right|^2}{1 + \left|\hat{n}^w\right|^2}, \tag{14}$$

where $\rho_s = |\hat{s}_0|^2 / \left|\hat{n}_0^s\right|^2$ represents the antenna signal SNR. The optimal performance of the algorithm is analyzed assuming that the antenna array has already undergone phase alignment. Under ideal conditions, represented by $|\hat{n}^w|^2 \rightarrow 0$, when the phase weight completely compensates for the phase offset, and the IF signals are optimally combined, the combined-signal SNR can essentially achieve the theoretical performance, $\rho_c = L\rho_s$. However, the combined performance cannot reach the theoretical gain owing to the influence of weight noise. Therefore, the SNR loss factor caused by weight estimation errors can be defined as follows:

$$\gamma_\rho = \frac{1 + 1/L\,|\hat{n}^w|^2}{1 + |\hat{n}^w|^2}. \tag{15}$$

Rogstad [19] proposed the concept of weight $\hat{W}_{iK}$ and derived the SNR formula. The weight SNR is used to better analyze the phase-offset estimation error caused by noise in the algorithm. Therefore, in this study, the weight SNR ρ_w of the CPC algorithm is obtained as follows:

$$\rho_w \approx \frac{N\rho_s\rho_s - (1+\rho_s)\,L\mu}{\rho_s + (1+\rho_s)\,L\mu}, \tag{16}$$

where $\mu = R^2\,|\hat{s}_0|^4 \approx \left(\frac{1}{L}\right)^2$, where R is the normalization coefficient to prevent the phase weight amplitude from becoming unstable due to continuous accumulation. In the SUMPLE algorithm, the reference signal in each signal correlation operation is a combination of $L-1$ antenna signals. In the CPC algorithm, the reference signal also includes the local carrier signal. The local carrier signal is generated by the PLL and fed back to the signal synthesis algorithm. Therefore, the carrier signal frequency is equal to the antenna signal, but the phase is different from each antenna signal. That is, the local carrier signal and the antenna signal contain a same-frequency sinusoidal signal. Thus, in Eq. (16), the formula uses L rather than $L-1$. Unlike the SUMPLE algorithm aligns the antenna signal phases to an uncertain common phase, the CPC algorithm aligns the carrier phases of all antenna signals with the local carrier phases, thereby aligning the signal phases of each array element.

Figure 4 illustrates the relationship between ρ_s and ρ_c for various values of N. In this figure, $L = 6$, and the theoretical CNR increases by approximately 7.8 dB. Comparing the combined-signal curve for $N = 30$ ms with the antenna signal curve at a higher antenna signal CNR, the two curves were nearly parallel, indicating that the enhancement in the combined-signal CNR approached its limiting value. From Eq. (16), ρ_w decreased with the antenna signal CNR decreased. This decrease indicated the reduced credibility of the weights for the signal carrier phase and less ideal phase alignment effects. Resulting in the enhancement in the combined-signal CNR decreased as the antenna signal CNR. Notably, at an antenna signal CNR of 30 dB-Hz and $N = 20$ ms, the combined-signal CNR was lower than the antenna signal CNR. This phenomenon occurred because ρ_w became negative, indicating algorithm failure. In such cases, the signal carrier phases cannot converge, leading to significant losses in the combined-signal CNR. Comparing the three combined-signal curves, a smaller N led to a reduced combined-signal CNR. Therefore, the CNR of the combined signal was proportional to N.

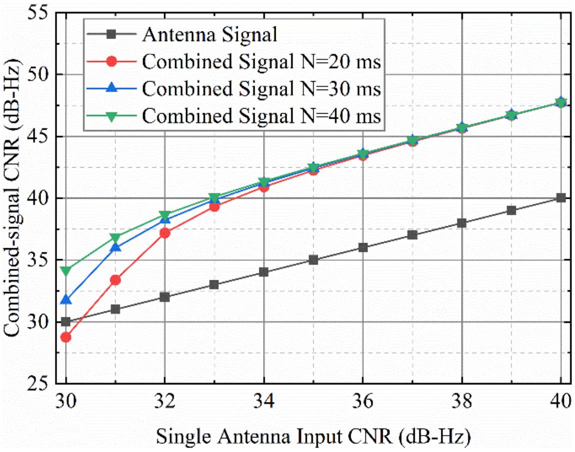


Fig. 4 Combining CNR for various correlation averaging intervals.

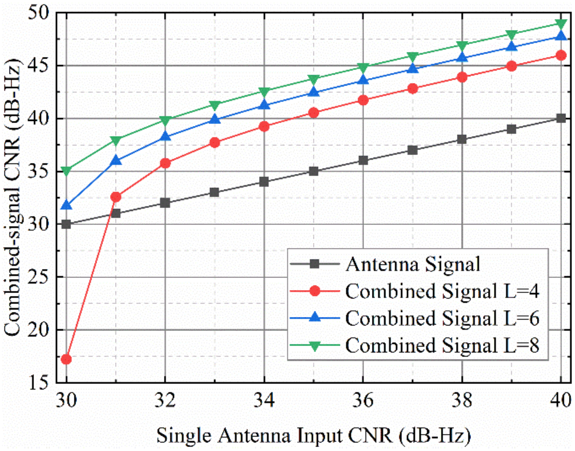


Fig. 5 Combining CNR for different numbers of antennas.

Figure 5 illustrates the relationship between ρ_s and ρ_c for different L, where $N = 30$ ms. The figure shows that an increase in L resulted in the enhancement of combined-signal CNR. The combined-signal CNR was directly proportional to L, consistent with the relationship expressed in Eq. (13). Similarly, the combination process incurred minimal losses at a higher antenna signal CNR, and the limit of the combined gain could be reached. At an antenna signal CNR of 30 dB-Hz and $L = 4$, the algorithm failed, resulting in a combined-signal CNR lower than the antenna signal CNR.

A simulation validation was conducted to assess the CNR enhancement performance of the CPC algorithm. The simulated signal was based on the GPS L1 C/A signal and characterized by the parameters presented in Table 1. Figure 6 compares the combined-signal CNR achieved by the CPC algorithm with that achieved by the conventional method. The CNR for the conventional algorithm was approximately 37.25 dB-Hz, whereas that for the CPC algorithm was approximately 38.20 dB-Hz. Compared with the single antenna signal CNR, the CPC algorithm has improved by 6.2 dB. Compared with the theoretical value, the combined CNR of the SUMPLE algorithm loses 2.52 dB,

Table 1 Simulation parameters.

Parameter	Value
Sampling frequency	24.552 MHz
Intermediate frequency	4.092 MHz
Noise bandwidth	2.046 MHz
CNR	32 dB-Hz
N	30 ms
L	6
Number of iterations	20

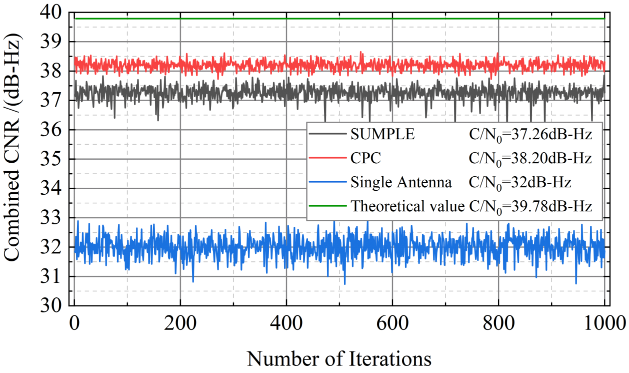


Fig. 6 Combined CNR for different algorithms.

the loss of the CPC algorithm is 1.58 dB, and the combined CNR is close to the theoretical value. Compared with the SUMPLE algorithm, the CNR loss of the CPC algorithm is reduced by about 1 dB. The combined-signal CNR was more stable because the reference signal $\hat{S}_{iK}^{Re}$ in the CPC algorithm contains the local carrier signal, which has high CNR and stable phase characteristics and enhances ρ_w. A higher ρ_w increased the combined-signal CNR and resulted in accurate phase-compensation values. Therefore, the phase alignment of the antenna array signals became stable, improving the stability of the CNR in the CPC algorithm.

4.2 Analysis of the Combined-Signal Carrier Phase

Equation (7) shows that the phase of the combined signal $\hat{C}_k$ is a function of time k. The phase is influenced by changes in $\hat{S}_{ik}$ and $\hat{W}_{iK}$. Assuming that the weight remains constant during the correlation time, the combined signal is integrated as

$$\begin{aligned}\hat{C}_{K+1} &= \frac{1}{ncor}\sum_{k=(K+1)N}^{(K+2)N-1}\left(\sum_{i=0}^{L-1}\hat{S}_{ik}\hat{W}_{iK+1}{}^{*}\right)\\ &= \hat{c}_{K+1} + \hat{n}_{K+1}^{c}, \qquad (17)\end{aligned}$$

where $\hat{c}_{K+1}$ and $\hat{n}_{K+1}^{c}$ correspond to the signal and noise components of the combined signal in $K+1$-*th* unit, respectively. The former is expressed as follows:

$$\hat{c}_{K+1} = \sum_{i=0}^{L-1}\left(\hat{s}_{iK}\hat{W}_{iK+1}^{*}\right). \qquad (18)$$

Substituting Eq. (8) into Eq. (18) and simplifying it yields the following expression:

$$\begin{aligned}\hat{c}_{K+1} &= \left|\hat{W}_{iK}\right|^{2}\left|\hat{s}_{0}\right|^{2}\hat{c}_{K} + \left|\hat{W}_{iK}\right|^{2}\left|\hat{s}_{0}\right|^{2}\hat{S}_{K}^{L}(L-1)\\ &\quad + \sum_{i=0}^{L-1}\hat{s}_{iK}\hat{n}_{iK+1}^{w}{}^{*}. \qquad (19)\end{aligned}$$

Equation (19) shows that $\sum_{i=0}^{L-1}\hat{s}_{iK}\hat{n}_{iK+1}^{w}{}^{*}$ and $\hat{S}_{K}^{L}$ influence the phase center of the combined signal. When using the conventional algorithm, during each signal correlation process, the reference signal is the weighted sum of all other antenna signals whose phase center is not fixed but varies owing to the influence of weight noise, causing the phase center of the combined signal to vary with time. Moreover, when the initial phase of each antenna signal is dispersed, the phase center may gradually deviate toward 0 or pi during the iteration process. This phenomenon can lead to convergence failure of conventional algorithms.

Rogstad [19] derived the combined-signal phase-compensation formula for the algorithm convergence problem as follows: $\Delta\hat{W}_{iK} = \sum_{i=0}^{L-1}\left(\hat{W}_{iK}\hat{w}_{iK-1}^{*}\right)$. However, $\hat{w}_{iK-1}$ in the equation is unknown; therefore, $\hat{W}_{iK}$ is used rather than $\hat{w}_{iK-1}$. This method is feasible when the SNR of the input signal is high because the weight noise $\hat{n}_{iK}^{w}$ is small, and $\hat{W}_{iK}$ can be approximately equal to $\hat{w}_{iK-1}$. However, in general, the existence of estimation errors adversely affects the compensation performance, causing the method to fail. By contrast, in the CPC algorithm, the phase of the local carrier signal remains constant and is added to the reference signal as a fixed-phase component. During each iteration, the phase of the antenna signals approaches this fixed component and eventually converges. This phenomenon effectively resolves the convergence problem in conventional algorithms.

Figure 7 compares the phase offset and correction versus iteration of the combined output for various algorithms, where the simulated signal is based on the GPS L1 C/A signal and characterized by the parameters presented in Table 1. The red line represents the phase of the combined signal using the SUMPLE algorithm, whereas the blue line represents the phase of the combined signal after using the SUMPLE algorithm and correcting the phase. The comparison revealed that although phase compensation reduced the combined phase error, it did not yield a significant improvement. The black line represents the phase of the combined signal using the CPC algorithm, showing a noticeable reduction in the phase error compared with the other two cases, with errors within 5°. Compared with that of the SUMPLE algorithm after phase compensation, the phase error variance of the CPC algorithm was reduced by 80%, and the stability was significantly improved.

Figure 8 shows the convergence performance of different algorithms, where the signal parameters are as shown in Table 1. From the figure, within 20 iterations of the CPC algorithm, the CNR of the combined output signal tends to be stable, and the algorithm completes convergence. The

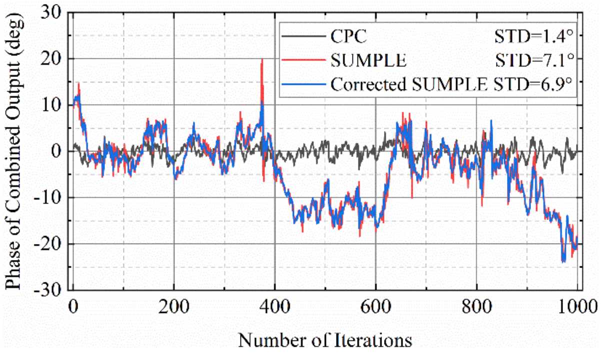


Fig. 7 Phase offset and correction versus iteration of the combined output for various algorithms.

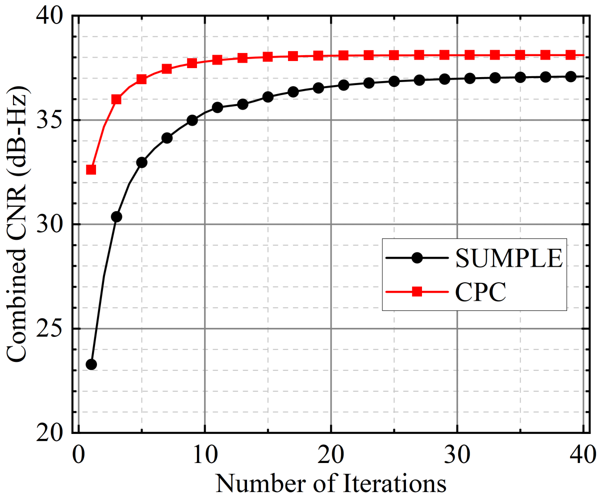


Fig. 8 Convergence performance for various algorithms.

SUMPLE algorithm converges within approximately 30 iterations. Moreover, the CNR of the combined signal of the CPC algorithm is higher than that of the SUMPLE algorithm, which is consistent with the results in Fig. 6.

5. Simulation Result and Performance Analysis

Simulation experiments were performed to validate the efficacy of the CPC algorithm in terms of the signal gain and phase alignment. The simulated signal was based on the GPS L1 C/A signal and characterized by the parameters presented in Table 1. In the experiment, the CNR of the navigation signal changed in the static environment of the receiver.

Figure 9 compares the theoretical combined-signal CNR deduced using Eq. (13) with the simulated combined-signal CNR. The simulated signal was enhanced by nearly 7 dB compared with the single-antenna signal when $\rho_s >$ 33 dB-Hz, consistent with the theoretical prediction. Comparing the theoretical and simulated curves, these results were consistent when $\rho_s >$ 33 dB-Hz. Below 33 dB-Hz, the simulated signal curve deviated from the theoretical curve, progressively increasing the discrepancy. This is because the antenna signal CNR decreases, the weight noise gradually increases, and the phase weight becomes more difficult to achieve the assumption of the optimal weight. The simulated signal CNR value was 1.6 dB lower than the theoretical value when ρ_s = 30 dB-Hz, which was improved by approximately 2.5 dB compared with the CNR value of a single antenna.

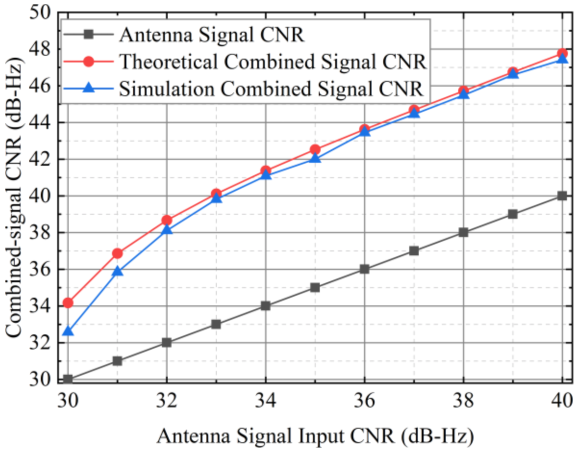


Fig. 9 Difference of combining CNR between simulation and theory.

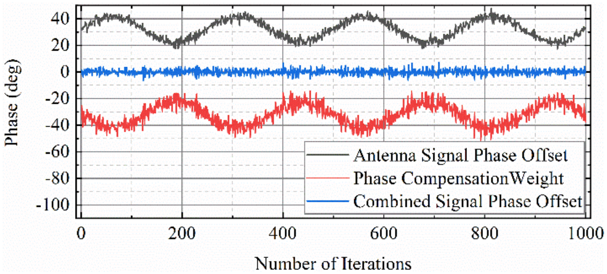


Fig. 10 CPC phase wander and correction versus iteration.

Figure 10 shows the simulation verification of the phase convergence of the CPC algorithm, in which a phase shift is added to the antenna signal carrier phase. The black curve represents the phase offset between the antenna and local carrier signals. The red curve represents the phase-compensation values applied by the CPC algorithm to the antenna signals. The compensatory phase exhibited an inverse relationship with the black curve, indicating that the CPC algorithm could effectively perform phase compensation on the antenna signal to align it with the local carrier signal. The blue curve represents the phase offset between the combined and local carrier signals after the phase alignment using the CPC algorithm and the blue curve value is within 5°, consistent with Fig. 7.

As shown in Fig. 11, six GPS baseband signals at different locations were simultaneously generated through GPS-SDR-SIM software to simulate the distributed antenna array receiving signals. The signal parameters were consistent with those listed in Table 1. The blue curve represents changes in the antenna signal CNR. The initial CNR of all antenna signals was 40 dB-Hz, which gradually decreased to 30 dB-Hz and reverted to 40 dB-Hz. Each signal

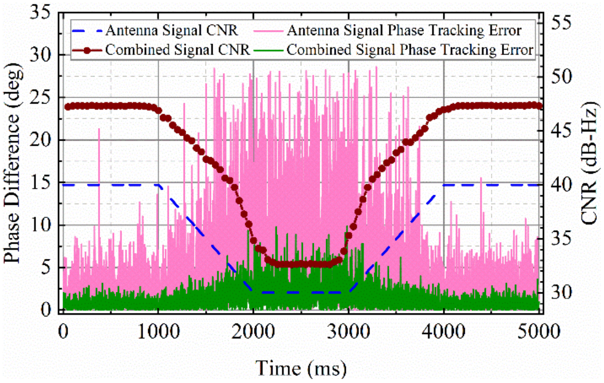


Fig. 11 Tracking performance between single and combined signals for various antenna signal CNRs.

segment persisted for 1000 ms. The brown curve represents the CNR of the combined signal calculated using the narrowband and wideband power ratios. When the antenna signal CNR was 40 dB-Hz, an enhancement of approximately 7.5 dB was observed, close to the theoretical value. When the CNR dropped below 33 dB-Hz, the enhancement value of the combined-signal CNR rapidly diminished, consistent with Fig. 9.

After 1000 ms, the CNR of the antenna signal decreased. At this point, the absolute value of the phase-tracking error in the PLL of the antenna signal increased rapidly. When the CNR decreased below 35 dB-Hz, the phase-tracking error was approximately 27°, indicating that the tracking loop was essentially out of lock. Compared with the antenna signal, the combined signal had more minor phase-tracking errors, and the performance improvement was more evident at a low CNR. When the antenna signal CNR was 30 dB-Hz, the phase-tracking error of the combined signal was maintained within 10°, and the phase variance was 2.3. By contrast, the phase variance of the antenna signal was 8.32, and the phase variance of the combined signal was approximately 72% lower than that of the antenna signal.

Figure 12 shows the in-phase prompt correlation values of the correlator output for antenna signal and combined signal carrier tracking loop. GPS navigation message information is obtained by binarizing the I_p value into 1 and −1 and demodulating it into a data bit stream. Therefore, the accuracy of the demodulation of the correlated value I_p affects the accuracy of the positioning result [35]. From the blue box in the picture, the decoding result of the antenna signal in-phase real-time correlation value I_p at 1340–1360 ms is wrong, which will lead to errors in subsequent positioning results. The decoding result of the combined signal in the figure is consistent with the real navigation message, and there is no decoding error.

6. Conclusion

This study investigated a multi-antenna signal-combining method in weak-signal environments. A CPC algorithm was proposed to effectively estimate phase offset and enhance signal gain. Theoretical derivation and analysis of the CPC algorithm demonstrated the relationship of the phase weight to the CNR of the combined signal. A higher CNR of the antenna signal, longer correlation time, and more array antennas indicated a smaller weight noise and superior algorithm performance. The results of the analysis and verification using a simulation platform demonstrated that the experiment results were consistent with those of the theoretical analysis. When using a distributed antenna array with six elements, the CPC algorithm could enhance the combined-signal CNR by 6 dB at 32 dB-Hz, exhibiting a 1 dB improvement compared with that of the SUMPLE algorithm. Moreover, the CPC algorithm could effectively estimate the phase offset between the signals of each array element, and the carrier phase error of the combined signal was maintained within 5°. Furthermore, tracking the combined signal through a PLL revealed that the phase-tracking error remained within 10° at 30 dB-Hz, representing a reduction of approximately 72% in the error compared with the antenna signal.

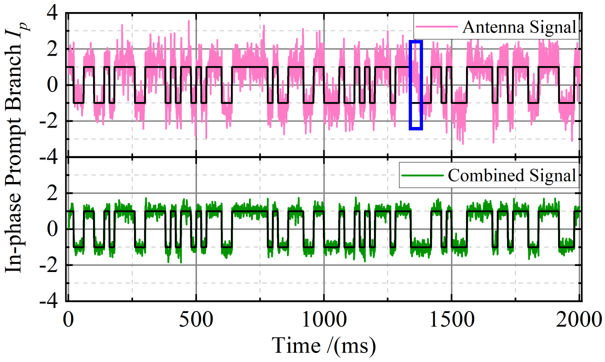


Fig. 12 PLL correlator output value I_p (navigation message).

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2021YFB3900703, and the National Science Foundation of China under Grant 41974038.

References

[1] E.D. Kaplan and C. Hegarty, eds., Understanding GPS: Principles and Applications, 2nd ed., Artech House, Boston, 2006.

[2] F.S.T.V. Diggelen, A-GPS: Assisted GPS, GNSS, and SBAS, Artech House, 2009.

[3] M.K. Bek, E.M. Shaheen, and S.A. Elgamel, "Analysis of the global position system acquisition process in the presence of interference," IET Radar, Sonar & Navigation, vol.10, no.5, pp.850–861, June 2016.

[4] H. Qin, X. Xue, and Q. Yang, "GNSS multipath estimation and mitigation based on particle filter," IET Radar, Sonar & Navigation, vol.13, no.9, pp.1588–1596, Sept. 2019.

[5] G. Gao, "INS-assisted high sensitivity GPS receivers for degraded signal navigation," vol.68, no.04, Library and Archives Canada = Bibliothèque et Archives Canada, Ottawa, 2007.

[6] P. Fan, X. Cui, S. Zhao, G. Liu, and M. Lu, "A two-step stochastic

hybrid estimation for GNSS carrier phase tracking in urban environments," IEEE Trans. Instrum. Meas., vol.70, pp.1–18, 2021.

[7] F. Xie, J. Liu, R. Li, and Y. Hang, "Adaptive robust ultra-tightly coupled global navigation satellite system/inertial navigation system based on global positioning system/BeiDou vector tracking loops," IET Radar, Sonar & Navigation, vol.8, no.7, pp.815–827, Aug. 2014.

[8] P.O. Gaggero and D. Borio, "Ultra-stable Oscillators: Limits of GNSS coherent integration," Proc. 21st International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2008), pp.565–575, Sept. 2008.

[9] X. Zhang, B. Guo, F. Guo, and C. Du, "Influence of clock jump on the velocity and acceleration estimation with a single GPS receiver based on carrier-phase-derived Doppler," GPS Solut, vol.17, no.4, pp.549–559, Oct. 2013.

[10] X. Feng, T. Zhang, X. Niu, T. Pany, and J. Liu, "Improving GNSS carrier phase tracking using a long coherent integration architecture," GPS Solut, vol.27, no.1, p.37, Dec. 2022.

[11] J.B.-Y. Tsui, Fundamentals of Global Positioning System Receivers: A Software Approach, John Wiley & Sons, Hoboken, NJ, USA, 2005.

[12] R.V. Kurynin, "A method of evaluating the g-sensitivity of quartz oscillators in GNSS receivers," 2019 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Russia, pp.1–5, July 2019.

[13] D.H. Rogstad, A. Mileant, and T.T. Pham, Antenna Arraying Techniques in the Deep Space Network, John Wiley & Sons, 2005.

[14] S. Blandino, F. Kaltenberger, and M. Feilen, "Wireless channel simulator testbed for airborne receivers," 2015 IEEE Globecom Workshops (GC Wkshps), San Diego, CA, USA, pp.1–6, Dec. 2015.

[15] J. Huang, Y. Su, W. Liu, and F. Wang, "Adaptive modulation and coding techniques for global navigation satellite system intersatellite communication based on the channel condition," IET Communications, vol.10, no.16, pp.2091–2095, Nov. 2016.

[16] M. Sadeghi, F. Behnia, and R. Amiri, "Maritime target localization from bistatic range measurements in space-based passive radar," IEEE Trans. Instrum. Meas., vol.70, pp.1–8, 2021.

[17] M. Rashid and J.A. Nanzer, "Frequency and phase synchronization in distributed antenna arrays based on consensus averaging and Kalman filtering," IEEE Trans. Wireless Commun., vol.22, no.4, pp.2789–2803, April 2023.

[18] D. Divsalar, "Symbol stream combining versus baseband combining for telemetry arraying," The Telecommun. and Data Acquisition Rept., pp.13–28, Aug. 1983.

[19] D.H. Rogstad, "The SUMPLE algorithm for aligning arrays of receiving radio antennas: Coherence achieved with less hardware and lower combining loss," Interplanetary Network Progress Report, vol.42–162, pp.1–29, Aug. 2005.

[20] K.-M. Cheung, "Eigen theory for optimal signal combining: A unified approach," Telecommunications and Data Acquisition Progress Report, vol.126, pp.1–9, April 1996.

[21] V.A. Vilnrotter and E.R. Rodemich, "A real-time signal combining system for Ka-band feed arrays using maximum-likelihood weight estimates," The Telecommunications and Data Acquisition Report, Feb. 1990.

[22] J. Xu, S. Ni, and Y. Yang, "Analysis of distributed synthesis algorithm based on simple and sumple algorithm," 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, pp.1327–1332, June 2021.

[23] C. Chen, H. Yu, C. Shen, C. Tan, A. Zhang, and Z. Tang, "A combining algorithm for signal arraying using MMSE estimator," 2009 Fourth International Conference on Communications and Networking in China, Xian, China, pp.1–4, Aug. 2009.

[24] L. Wang and D. Wang, "A gain factor controlled SUMPLE algorithm and system," 2017 IEEE 12th International Conference on ASIC (ASICON), Guiyang, pp.183–186, Oct. 2017.

[25] Y. Di, S. Weiyi, L. Peijie, S. Ke, and L. Xiaoyu, "Parallel implementation of SUMPLE algorithm in large-scale antenna array," Communications, Signal Processing, and Systems, Singapore, pp.433–439, 2020.

[26] S. Ni and J. Xu, "A distributed receiving synthesis algorithm based on resampling," Telecommunication Engineering, vol.63, no.5, pp.676–680, 2023.

[27] P. Siu, C. Apker, J. McMillen, and S. Sorber, "G-STAR GPS antijam technology," Proc. 2006 National Technical Meeting of The Institute of Navigation, pp.1057–1063, Jan. 2006.

[28] D. Rowe, J. Weger, and J. Walker, "Integrated GPS anti-jam systems," Proc. 18th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2005), pp.1–7, Sept. 2005.

[29] S. Backen, D.M. Akos, and M.L. Nordenvaad, "Post-processing dynamic GNSS antenna array calibration and deterministic beamforming," Proc. 21st International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2008), pp.2806–2814, Sept. 2008.

[30] N. Vukmirović, M. Erić, M. Janjić, and P.M. Djurić, "Direct wideband coherent localization by distributed antenna arrays," Sensors, vol.19, no.20, p.4582, Oct. 2019.

[31] X. LI, P. LI, X. Liu, and Z. He "Signal combination of distributed antenna array using power inversion criteria," Telecommunication Engineering, vol.60, no.12, p.1432, 2020.

[32] R.T. Compton, "The power-inversion adaptive array: Concept and performance," IEEE Trans. Aerosp. Electron. Syst., vol.AES-15, no.6, pp.803–814, Nov. 1979.

[33] L. Lu and M. Gao, "A satellite calibration method for the baseline coordinate and phase difference of distributed radar array," Journal of Electronics & Information Technology, vol.41, no.12, pp.2896–2902, 2019, doi: 10.11999/JEIT181152.

[34] H. Tong, "Joint processing for satellite navigation signals," Ph.D, National University of Defense Technology, 2014.

[35] K. Borre, D.M. Akos, N. Bertelsen, P. Rinder, and S.H. Jensen, A Software-Defined GPS and Galileo Receiver, Birkhäuser, Boston, MA, 2007.

Wenfei Guo is currently a professor in GNSS Research Center, Wuhan University (WHU). He received his PhD degree in communication and information system from Wuhan University, Wuhan, China, in 2011. His research interests include GNSS receivers and related signal processing technologies, including high-precision timing receivers, anti-jamming receivers, and LEO navigation, etc.

Jun Zhang was born in 1999. He received the B.E. degree in electronic information school from Wuhan University, Wuhan, China, in 2017. He is currently pursuing the M.S. degree with the GNSS Center, Wuhan University, Wuhan, China. His current research interests include GNSS baseband signal processing, GPS acquisition and tracking algorithm.

Chi Guo received the Ph.D. degree in computer science from Wuhan University, Wuhan, Hubei, China, in 2010. He is currently a professor with the National Satellite Positioning System Engineering Technology Research Center, Wuhan University. His current research interests include BeiDou application, unmanned system navigation, and location-based services (LBS).

Weijun Feng was born in 2000. He received the B.E. degree in communication engineer school from Hangzhou Dianzi University, Hangzhou, China, in 2018. He is currently pursuing the M.S. degree with the GNSS Center, Wuhan University, Wuhan, China. His current research interests include Iridium satellite signal baseband signal processing and position method.

PAPER

UAV-BS Operation Plan Using Reinforcement Learning for Unified Communication and Positioning in GPS-Denied Environment

Gebreselassie HAILE[†] ***and*** **Jaesung LIM**[††a)], ***Nonmembers***

SUMMARY An unmanned aerial vehicle (UAV) can be used for wireless communication and localization, among many other things. When terrestrial networks are either damaged or non-existent, and the area is GPS-denied, the UAV can be quickly deployed to provide communication and localization services to ground terminals in a specific target area. In this study, we propose an UAV operation model for unified communication and localization using reinforcement learning (UCL-RL) in a suburban environment which has no cellular communication and GPS connectivity. First, the UAV flies to the target area, moves in a circular fashion with a constant turning radius and sends navigation signals from different positions to the ground terminals. This provides a dynamic environment that includes the turning radius, the navigation signal transmission points, and the height of the unmanned aerial vehicle as well as the location of the ground terminals. The proposed model applies a reinforcement learning algorithm where the UAV continuously interacts with the environment and learns the optimal height that provides the best communication and localization services to the ground terminals. To evaluate the terminal position accuracy, position dilution of precision (PDOP) is measured, whereas the maximum allowable path loss (MAPL) is measured to evaluate the communication service. The simulation result shows that the proposed model improves the localization of the ground terminals while guaranteeing the communication service.

key words: *unmanned aerial vehicle, communication, localization, reinforcement learning, PDOP*

1. Introduction

An unmanned aerial vehicle (UAV), also known as a drone, or an airborne relay, is an aircraft controlled by a computer system through a radio communication link. UAVs have become the center of research in the industry because of their paramount importance for military and civilian applications. In the civilian application, UAVs are highly demanded for public safety and rescue operations when natural and/or man-made disasters occur. In such cases, UAVs can be quickly deployed to serve as base station in the sky (UAV-BS) and provide communication as well as localization services [1], [2].

UAV has size, weight, and power (SWaP) limitations. Therefore, it is crucial to optimize the transmission power and bandwidth of UAV-BS communications. Various research issues and challenges regarding efficient UAV operation in wireless networks were introduced in [3], [4].

The use of UAV-BS for communication service has been studied in [5]–[10]. In [5], the authors proposed an analytical approach to optimize the altitude of low area platforms (LAPs) which can deliver essential wireless communication for public safety agencies in remote areas or during the aftermath of natural disasters. The main goal of this research work is to provide maximum radio coverage on the ground. In [6], the authors proposed an energy efficient placement of a drone base station for minimum required transmit power. They formulated the problem in a way such that it minimizes the average transmit power of the UAV-BS that serves a set of ground users. The authors in [7] proposed 3-D placement of a directional-antenna equipped UAV-BS aiming to maximize the number of flying/hovering UAV-UEs under its coverage area.

In [8], the authors applied deep reinforcement learning to make drones behave autonomously inside a suburban neighborhood which has plenty of obstacles such as trees, cables, parked cars, houses, and other moving drones. The UAV learns about the environment to avoid these stationary and moving obstacles as it navigates through the neighborhood showing how it can be used to provide communication services safely. The authors in [9] proposed a Q-learning based UAV deployment algorithm in which the UAV makes its own decision for attaining an optimal 3-D position by learning from trial and mistake for maximizing the sum mean opinion score of ground users. In [10], the authors studied how to maximize the overall data rate through an intelligent deployment of an UAV-BS in the downlink of a cellular system. They apply a reinforcement learning algorithm to avoid collision between multiple UAVs and optimize the UAV-BS positions that provide maximum sum data rate of multiple user equipment.

The use of UAV-BS for localization service in non-GPS environments has been studied in [11]–[15]. In [11], the authors proposed the use of a single UAV to localize terminals in battlefield environments as the use of global navigation satellite system (GNSS) such as GPS is prone to jamming and has weak signal reception capability. They analyzed the localization service by varying the number of received navigation signals, and the velocity of the UAV. In [12], the authors proposed a Doppler shift-based user position detection system using UAV. They measured the statistical and quantitative performance of the positioning errors of a single ground user as the UAV moves in sinusoidal curve. The ground user sends continuous signal with a fixed frequency, the UAV receives it, and relays it to the terrestrial

Manuscript received November 2, 2023.
Manuscript revised March 9, 2024.
Manuscript publicized May 6, 2024.

[†]Dept. of AI Convergence Network, Ajou University, Suwon, South Korea.
[††]Dept. of Military Digital Convergence, Ajou University, Suwon, South Korea.
a) E-mail: jaslim@ajou.ac.kr (Corresponding author)

DOI: 10.23919/transcom.2023EBP3174

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

control station where the position computation takes place. In [13], the authors proposed the use of multiple UAV-BSs and additional ground references to locate a ground user. In [14], the authors proposed UAV-assisted localization of wireless devices that are in network outage and have run out of power. To localize the inactive devices, the authors used wireless power transfer (WPT) based wireless charging in which a small amount of power is transferred by the UAV to enable the target device to broadcast a beacon. The beacon from the devices contains the information about the neighboring nodes and their signal strength. The paper in [15] proposed the use of UAV-BSs for localization of a connected autonomous vehicle (CAV). They applied reinforcement learning algorithm to find the best spatial configuration of the UAV-BSs to localize the CAV in an unknown environment.

In all the above research works, the focus of the authors is either on the use of the UAV-BS for communication or localization. In this study, we propose a reinforcement learning based UAV-BS deployment scheme to provide both communication and localization services to terminals in a suburban environment without cellular communication and GPS connectivity. To do so, the UAV-BS is first deployed within the defined minimum and maximum heights. Then, the maximum allowable path loss (MAPL) of the edge terminal is computed using the air to ground (ATG) path loss model to analyze the communication service. For the localization service, the UAV-BS periodically sends navigation signals to the ground terminals. Then, each ground terminal calculates its own position using the time difference of arrival (TDOA) algorithm. We assume that a single UAV-BS moves in a circular fashion and sends navigation signals at N navigation signal transmission points (NSTP) which define the UAV positions. The NSTPs of the UAV-BS serve as reference (anchor) signals. The UAV-BS to terminal geometry impacts the position accuracy of the terminals. The metric that is normally used for measuring the terminal position accuracy is known as dilution of precision (DOP) which represents the degree to which the UAV-BS to terminal geometry dilutes the position accuracy of the terminals [16].

The main contributions of this work are:

- Defined localization and communication models using a single UAV-BS in a GPS and cellular communication denied suburban environment.
- Defined an optimization problem that integrates both the localization and communication services.
- Proposed a reinforcement learning based model to solve the optimization problem.

The rest of the paper is organized as follows. Section 2 provides the proposed system model. From the system model, the proposed unified positioning and communication schemes are described in detail. Section 3 presents the reinforcement learning based approach. Section 4 provides the simulation results, and finally, the conclusion and future work are provided in Sect. 5.

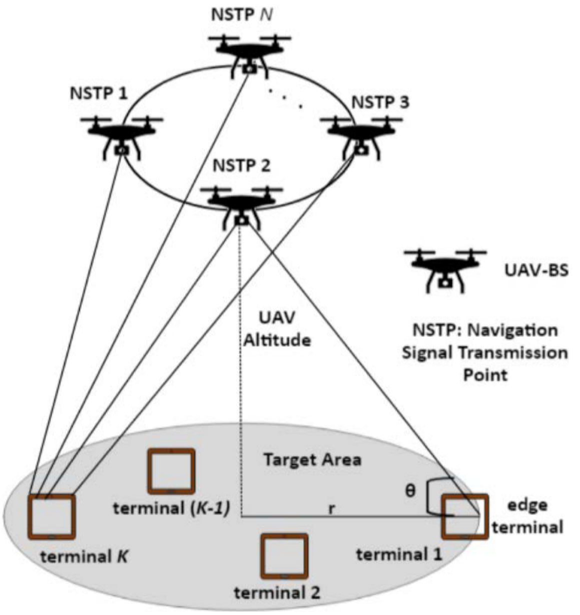


Fig. 1 Proposed system model.

2. System Model

Figure 1 shows the proposed system model. The UAV-BS is deployed to the target area where it moves in a circular route to provide the localization and communication services to the terminals. It's assumed that the UAV-BS is equipped with its own navigation equipment that provides accurate location information at any navigation signal transmission point, NSTP 1 to NSTP N as shown in Fig. 1. Also, it is assumed that the UAV platform consists of fixed-wing aircraft which can turn during flight in the sky and send downlink navigation signals to the terminals periodically. The terminals are assumed to be static (no mobility). Each navigation signal transmission point can provide different coverage area ranges as the UAV-BS moves in a circular path in the air. In this research work, however, we considered a fixed target area where the terminals are located. Only the edge terminal is located at the edge of the target area. Hence, the target area coverage and the location of the edge terminal are fixed as shown in Fig. 1.

2.1 Positioning Scheme

Let t_n be the time when the UAV-BS transmits a navigation signal from the n-th NSTP and, τ_n be the time when a ground terminal receives it.

The pseudo-range between the n-th UAV-BS NSTP and the k-th ground terminal is computed as

$$\rho_k^n = c \times (\tau_n - t_n) + \varepsilon_n \tag{1}$$

where $k = \{1, 2, \ldots, K\}$ is a ground terminal index, $n = \{1, 2, \ldots, N\}$ is a UAV-BS navigation signal transmission point index, c is the speed of light and ε_n denotes the error that occurs during navigation signal transmission.

For any k-th ground terminal, the pseudorange difference between the n-th and the first UAV-BS navigation signal transmission points becomes:

$$\rho_k^n - \rho_k^1 = c \times ((\tau_n - \tau_1) - (t_n - t_1)) \tag{2}$$

where $\rho_k^1 = c \times (\tau_1 - t_1) + \varepsilon_1$ is the pseudo-range between the first UAV-BS navigation signal transmission point, $n = 1$, and the k-th ground terminal. The errors $\epsilon_1 = \epsilon_2 = \ldots = \epsilon_n$ are similar for the same environment, the suburban environment in our case. So, in the pseudo-range difference computations, these values cancel each other out.

In the 3-D Euclidean space orthogonal coordinate system, the pseudo-ranges ρ_k^1 and ρ_k^n are computed as follows:

$$\rho_k^1 = \left\|\mathrm{R}^1 - \mathrm{R}_k\right\| \tag{3}$$

$$\rho_k^1 = \sqrt{(x^1 - x_k)^2 + (y^1 - y_k)^2 + (z^1 - z_k)^2}$$

$$\rho_k^n = \|\mathrm{R}^n - \mathrm{R}_k\| \tag{4}$$

$$\rho_k^n = \sqrt{(x^n - x_k)^2 + (y^n - y_k)^2 + (z^n - z_k)^2}$$

where $||\cdot||$ represents the Euclidean norm vector, R^n is the position vector of the UAV-BS at t_n, and R_k is the position vector of the k-th terminal. Here, the UAV-BS location (x^n, y^n, z^n) is known at each navigation signal transmission time, t_n, whereas the position of the k-th terminal, $\mathrm{R}_k = (x_k, y_k, z_k)$, is unknown.

From Eq. (2), Eq. (3), and Eq. (4), the position of the k-th terminal, R_k, can be determined using the TDOA algorithm and the non-linear least squares method in the Levenberg-Marquardt algorithm [13].

From the pseudorange equation provided in Eq. (2), let's define matrices H and Z as follows:

$$H = \begin{bmatrix} \rho_k^2 & - & \rho_k^1 \\ \rho_k^3 & - & \rho_k^1 \\ \vdots & \vdots & \vdots \\ \rho_k^N & - & \rho_k^1 \end{bmatrix} \tag{5}$$

$$Z = \begin{bmatrix} -\dfrac{x^1 - x_k}{\rho_k^1} & -\dfrac{y^1 - y_k}{\rho_k^1} & -\dfrac{z^1 - z_k}{\rho_k^1} \\ -\dfrac{x^2 - x_k}{\rho_k^2} & -\dfrac{y^2 - y_k}{\rho_k^2} & -\dfrac{z^2 - z_k}{\rho_k^2} \\ \vdots & \vdots & \vdots \\ -\dfrac{x^N - x_k}{\rho_k^N} & -\dfrac{y^N - y_k}{\rho_k^N} & -\dfrac{z^N - z_k}{\rho_k^N} \end{bmatrix} \tag{6}$$

where matrix H is a column vector which consists of the pseudo-range differences between the first and the remaining $(N - 1)$ UAV-BS positions for the k-th terminal, and matrix Z is a set of unit vectors of the N UAV-BS positions for the k-th terminal.

From Eq. (5) and Eq. (6), the terminal position is computed as:

$$\mathrm{R}_k = \frac{1}{2}(H^T H)^{-1} H^T Z \tag{7}$$

where R_k refers to the position of the k-th terminal, H^T is the transpose of matrix H, and $(H^T H)^{-1}$ indicates the inverse of the matrix $(H^T H)$. At least four UAV-BS positions ($N \geq 4$) are required to calculate the positions of each terminal, $\mathrm{R}_k = (x_k, y_k, z_k)$, using Eq. (7).

Position DOP (PDOP) is a metric used to measure the accuracy of terminal positioning in global navigation systems, particularly in the context of global positioning systems (GPS) and other airborne-relay based navigation systems. It is the uncertainty of 3-D parameters (latitude, longitude, and height) and depends on the geometric arrangement of the navigation signal transmission points and the altitude of UAV-BS from perspective of the ground terminals. To compute the PDOP, let's define the geometric matrix, G:

$$G = \begin{bmatrix} -\dfrac{x^1 - x_k}{\rho_k^1} & -\dfrac{y^1 - y_k}{\rho_k^1} & -\dfrac{z^1 - z_k}{\rho_k^1} & 1 \\ -\dfrac{x^2 - x_k}{\rho_k^2} & -\dfrac{y^2 - y_k}{\rho_k^2} & -\dfrac{z^2 - z_k}{\rho_k^2} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ -\dfrac{x^N - x_k}{\rho_k^N} & -\dfrac{y^N - y_k}{\rho_k^N} & -\dfrac{z^N - z_k}{\rho_k^N} & 1 \end{bmatrix} \tag{8}$$

From the geometric matrix in Eq. (8), we define the covariance matrix $Q = (GTG)^{-1}$ which is a 4×4 matrix.

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} \\ Q_{41} & Q_{42} & Q_{43} & Q_{44} \end{bmatrix} \tag{9}$$

Then, the PDOP of each terminal is extracted from the covariance matrix, Q, as follows:

$$\mathrm{PDOP}_k = \sqrt{Q_{11} + Q_{22} + Q_{33}} \tag{10}$$

PDOP is a dimensionless number. A lower PDOP value indicates a more favorable geometric configuration, leading to higher position accuracy, while a higher PDOP value suggests less favorable geometry and potentially reduced accuracy.

Another metric used to evaluate the terminal position accuracy is root mean square error (RMSE). RMSE is the measure of the root of the mean of the squared errors between the predicted and true/actual terminal position values.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}} \left((\mathrm{x_k} - \mathrm{x\hat{}_k})^2 + (\mathrm{y_k} - \mathrm{y\hat{}_k})^2 + (\mathrm{z_k} - \mathrm{z\hat{}_k})^2\right)}{\mathrm{K}}} \tag{11}$$

where (x_k, y_k, z_k) is the true position and $(\mathrm{x\hat{}_k}, \mathrm{y\hat{}_k}, \mathrm{z\hat{}_k})$ is the estimated position of the k-th terminal.

2.2 Communication Scheme

From the system model provided in Fig. 1, the ATG model is used to evaluate the communication service of the terminals located in the target area. From the ATG model [17], the average path loss between the UAV-BS and the ground terminal is computed as:

$$\mathrm{PL} = \mathrm{PL}_{LoS} \times p(LoS, \theta) + \mathrm{PL}_{NLoS} \times p(NLoS, \theta) \qquad (12)$$

where PL_{LoS} is the line-of-sight (LOS) path loss, $p(LoS, \theta)$ is the LOS probability at elevation angle θ, PL_{NLoS} is the non-line-of-sight (NLOS) path loss, and $p(NLoS, \theta)$ is the NLOS probability at elevation angle θ.

Now, let's see how each of the parameters in the average path loss equation provided in Eq. (12) are computed. The elevation angle is defined by $\theta = \arctan\left(\frac{h}{r}\right)$ where h is the UAV height, and r is the horizontal distance between the center of the coverage area and the ground terminal.

The LOS and NLOS path loss parameters are given by:

$$\mathrm{PL}_{LoS} = \mathrm{FSPL} + \eta_{LoS} \qquad (13)$$

$$\mathrm{PL}_{NLoS} = \mathrm{FSPL} + \eta_{NLoS} \qquad (14)$$

where η_{LoS} and η_{NLoS} are the LOS and NLOS excessive path losses respectively. Their values are given in Table 1.

The free space path loss, FSPL, is given by:

$$FSPL = 20 \times log_{10}\left(\frac{4 \times \pi \times f \times d}{c}\right) \qquad (15)$$

where f is the operating frequency, $d = \sqrt{h^2 + r^2}$ is the distance between the UAV-BS and the ground terminal, and c is the speed of light.

The LOS probability is given by:

$$p(LoS, \theta) = \frac{1}{1 + \alpha \times \exp(-\beta \times (\theta - \alpha))} \qquad (16)$$

where α and β are environmental constants, whose values are shown in Table 1.

Equation (16) shows that the probability of having a line-of-sight connection between the UAV-BS and a ground terminal increases as the elevation angle increases. This decreases the mean path loss because the shadowing effect, which is the attenuation of the signal due to obstacles, decreases as the elevation angle increases. On the other hand, as the elevation angle increases, the distance between the UAV-BS and the ground terminal also increases which results in higher path loss. The NLOS probability at the given θ becomes:

$$p(NLoS, \theta) = 1 - p(LoS, \theta) \qquad (17)$$

Now, we have all the parameters to compute the average path loss value, PL, at each terminal using Eq. (12).

The minimum received power at each ground terminal depends on the transmitted power of the UAV-BS, and the maximum path loss. In [18], the authors proposed a path loss and height optimization (PLaHO) model where they defined the maximum path loss for a given UAV height as follows:

Table 1 Environment constants [1], [18].

Parameters	Suburban	Urban	Dense Urban
α	4.88	9.61	12.08
β	0.43	0.16	0.11
η_{LoS}	0.1	1	1.6
η_{NLoS}	21	20	23
a_0	0.1154	0.1150	0.1151
a_1	4.8008	4.5068	4.4511

$$h = \exp\left(a_0\left(\mathrm{PL}_{max} - u\right) - a_1\right) \qquad (18)$$

where a_0 and a_1 are environmental constants, whose values are given in Table 1, and $u = 20 \times log_{10}\left(f \times 10^{-9}\right)$ where f is the operating frequency.

According to Eq. (18), the maximum path loss for an UAV-BS placed at 1400 m in a suburban environment is 110.4 dB. So, the maximum path loss value of 110.4 dB would be used as the threshold path loss in this study.

2.3 Communication and Localization

By combining the communication and localization schemes, we define the following optimization problem.

$$\begin{aligned} &\text{min. } \mathrm{PDOP}_{ave} \\ &s.t. \ \mathrm{PL} \leq \mathrm{PL}_{\max} \\ &\mathrm{h}_{min} \leq \mathrm{h} \leq \mathrm{h}_{max} \end{aligned} \qquad (19)$$

where PDOP_{ave} is the average PDOP for the terminals in the target area, PL is the path loss of a terminal, and h is the UAV-BS height.

We are going to apply a reinforcement learning algorithm to solve Eq. (19) which will be described in the next sections in detail.

3. Communication and Localization Using RL

Reinforcement learning enables an agent to learn by continuously interacting with the eenvironment by trial and error using feedback from its own actions and experiences. In RL, the agent is not programmed what actions to take; instead, it learns the consequence of its actions. At each time step, the agent receives a state s_t from the state space and selects an action a_t from the set of possible actions in the action space. As a result of the action it takes, the agent gets a numerical reward r_{t+1} one time step later from the environment, and it finds itself in a new state s_{t+1} [19]. Figure 2 shows the agent-environment interaction in reinforcement learning algorithm.

Q-learning is a model-free RL algorithm which learns the value of an action in a particular state. Q-learning algorithms carry out an action multiple times and adjust the policy for optimal rewards based on the outcomes of the actions. Epsilon-Greedy action selection policy is applied

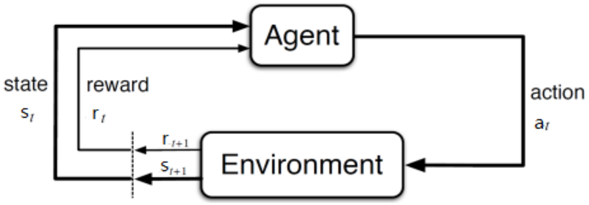


Fig. 2 Agent-Environment interaction in RL.

in the Q-learning algorithm where epsilon is a probability value that balances the exploration and exploitation of the action by the agent. Epsilon helps the agent to exploit the action with small probability of exploring.

In this paper, Q-learning algorithm is applied to solve the problem defined in Eq. (19). The goal is to apply Q-learning algorithm to acquire the minimum PDOP$_{ave}$ value under the path loss and height constraints for unified communication and localization services using UAV-BS. For any Q-learning algorithm, the environment, agent, state, action, and reward should be defined. In this paper, these parameters are defined as follows:

Environment – An environment represents the system an agent interacts with. In this paper, the environment is a GPS and cellular communication denied suburban environment.

Agent – An agent is the entity that interacts with the environment to achieve a specific task. In this paper, the agent is the UAV-BS. So, UAV-BS and agent can be used interchangeably.

State space – The UAV-BS height (altitude) forms the state space in our model. Originally, the state space is continuous as the UAV-BS can take any value as it moves within the defined minimum height, h_{min} and the maximum height, h_{max}. This continuous UAV-BS height is then discretized to give an integer number of states. By defining Δh as the change of height after each action, the total number of states is computed as:

$$N_{states} = \left\lfloor \frac{(\mathrm{h}_{max} - \mathrm{h}_{min})}{\Delta\mathrm{h}} + 1 \right\rfloor \tag{20}$$

where the floor function $\lfloor x \rfloor$ takes a real number x and gives the greatest integer less than or equal to x as an output.

In this paper, h_{min} and h_{max} are 400 m and 1400 m respectively, and Δh is 100 m. Applying these values to Eq. (20) provides $N_{states} = 11$ discrete number of states which define the state space. Figure 3 shows how each action the agent takes changes the state-space.

Action space: An action indicates what the agent (UAV-BS) does from the current state. An action space indicates the possible set of actions that the agent can take in the agent-environment interaction. In this paper, we have defined three types of actions the UAV-BS can take from the current state: An Upward Action, a Downward Action, and a Static Action, as illustrated in Fig. 3. Assuming that the current state of the agent is h, depending on the epsilon greedy policy, the agent takes one of the three possible actions. An Upward Action takes the agent to the (h + Δh) state, a Downward Action takes it to the (h − Δh) state, and a Static Action causes the agent to remain int its current state at h.

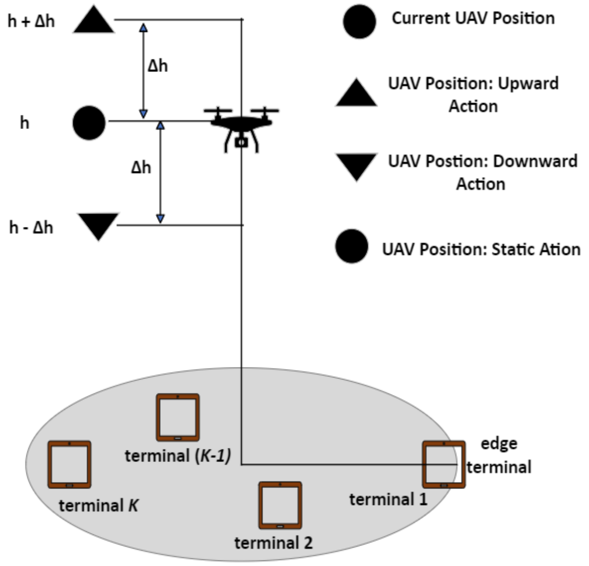


Fig. 3 Action space of the agent, UAV-BS.

Policy: Policy indicates how an agent chooses its actions. In Q-learning, an epsilon greedy strategy decides whether the agent should explore or exploit while interacting with the environment. The agent initially starts out by choosing a random action (exploration). As the episode progresses, the epsilon value provides a balance between exploration and exploitation. In exploitation, the agent chooses its action based on the highest Q-value from the Q-table for the given state.

Reward – A reward is a scalar value received after each action for transition to the new state. The average PDOP value is used as a reward in this paper. It measures the position accuracy of the terminals in the target area. The UAV-BS sends the navigation signals from the N UAV-BS positions of the current state, and each terminal computes the PDOP value as given in Eq. (10). When the number of terminals in the target area is more than one ($K > 1$), an average PDOP is used as the reward. The average PDOP is defined by:

$$\mathrm{PDOP}_{ave} = \frac{\sum_{k=1}^{K} \mathrm{PDOP}_k}{K} \tag{21}$$

The PDOP$_{ave}$ decides the reward in the Q-learning algorithm. Low average PDOP shows good terminal position accuracy, and large average PDOP shows bad terminal position accuracy.

The Q-learning algorithm uses a Q-table which contains the state and action pair known as Q-values. At each state, the agent computes the numerical reward r_{t+1}, based on the average PDOP as follows:

$$\mathrm{r}_{t+1} = \begin{cases} -1, & \textit{if } \mathrm{PDOP}_{ave}(\mathrm{s}_{t+1}) > \mathrm{PDOP}_{ave}(\mathrm{s}_t) \\ \ 0, & \textit{if } \mathrm{PDOP}_{ave}(\mathrm{s}_{t+1}) = \mathrm{PDOP}_{ave}(\mathrm{s}_t) \\ +1, & \textit{if } \mathrm{PDOP}_{ave}(\mathrm{s}_{t+1}) < \mathrm{PDOP}_{ave}(\mathrm{s}_t) \end{cases} \qquad (22)$$

where $\mathrm{PDOP}_{ave}(\mathrm{s}_{t+1})$ is the average PDOP at the next state, and $\mathrm{PDOP}_{ave}(\mathrm{s}_t)$ is the average PDOP at the current state.

Depending on the PDOP_{ave}, the UAV-BS decides r_{t+1} as shown in Eq. (22). If the average PDOP at the next state is greater than the average PDOP at the current state, the agent gets a negative reward. If the average PDOP at the next state is lower than the average PDOP at the current state, the agent receives a positive reward. If there is no change in the average PDOP values, the agent gets zero reward. The agent updates its Q-table and takes one of the 3 actions based on the epsilon greedy policy to move to the next state as shown in Fig. 3.

UAV-BS Q-table update – the UAV-BS has an action-value matrix which represents the value of being in a specific state s_t, while taking an action a_t. The UAV-BS updates the Q-value of the current state, $\mathrm{Q}_n(\mathrm{s}_t, \mathrm{a}_t)$, through the Q-learning function defined by:

$$\begin{aligned} \mathrm{Q}_n(\mathrm{s}_t, \mathrm{a}_t) \ =\ & \mathrm{Q}_o(\mathrm{s}_t, \mathrm{a}_t) + \\ & \mu \times \left(\mathrm{r}_{t+1} + \gamma \times \max_{\mathrm{a}} \mathrm{Q}(\mathrm{s}_{t+1}, \mathrm{a}) - \mathrm{Q}_o(\mathrm{s}_t, \mathrm{a}_t)\right) \end{aligned}$$

which can be simplified to:

$$\begin{aligned} \mathrm{Q}_n(\mathrm{s}_t, \mathrm{a}_t) \ =\ & (1-\mu) \times \mathrm{Q}_o(\mathrm{s}_t, \mathrm{a}_t) + \\ & \mu \times \left(\mathrm{r}_{t+1} + \gamma \times \max_{\mathrm{a}} \mathrm{Q}(\mathrm{s}_{t+1}, \mathrm{a})\right) \end{aligned} \qquad (23)$$

where $\mathrm{Q}_n(\mathrm{s}_t, \mathrm{a}_t)$ is the new Q-value of the current state, $\mathrm{Q}_o(\mathrm{s}_t, \mathrm{a}_t)$ is the old Q-value of the current state, μ is the learning rate, r_{t+1} is the reward defined in Eq. (22), γ is the discount factor, and $\max_{\mathrm{a}} \mathrm{Q}(\mathrm{s}_{t+1}, \mathrm{a})$ is the action that maximizes the Q-value of the next state. μ determines how much the agent adjusts its estimates based on new information obtained from the interactions with the environment. It's a value between 0 and 1. γ is a parameter that controls the importance of future rewards in the agent's decision-making process. It's a value between 0 and 1 and represents the extent to which the agent values future rewards compared to immediate rewards.

Stopping criteria: Initially, the UAV-BS is randomly located in one of the discrete states which correspond to the UAV-BS heights. Then, the interaction between the agent and the environment proceeds in sequences of steps until a stopping criterion is met. Stopping criteria decide when the agent should stop interacting with the environment. One way to define the stopping criteria is to let the agent continue until all the available states are visited. This is done to give the agent enough opportunity to interact with the environment and learn about it through exploitation and exploration following the e-greedy policy. Another way to outline the stopping criteria is to specify the number of steps in each episode. In this paper, we have defined the stopping criteria based on the number of steps. The agent runs for 200 steps and then stops. This is decided based on a repeated simulation observation where the reward does not improve when the number of steps exceeds 200.

DQN versus Q-learning: Deep Q-learning network (DQN) has become widespread in many of the RL-based research works recently. In this paper, however, Q-learning has been selected because the number of state and action spaces, (states = 11, and actions = 3), is very small and memory is not a problem. When the state and action spaces are large, using the Q-table is impractical because of memory limitations which affect the performance. In that case, DQN should be used as it addresses the memory limitation of Q-learning through *Replay Memory* technique where only limited number of state-action pairs are used instead of the whole state-action pairs.

4. Simulation Results

MATLAB is used to simulate the proposed UCL-RL model. We have developed a customized suburban environment that contains randomly generated terminals within a defined coverage area. The agent (UAV-BS) continuously interacts with the environment and learns about it using the reinforcement algorithm.

To the best of our knowledge, this is the first work that proposed the use of UAV for unified communication and localization services using reinforcement learning. To evaluate the performance of the proposed *UCL-RL* model, we have used two models for comparison. The first model is proposed in [11] which analyzed the use of single and dual UAV to localize terminals in battlefield environments. Since we are using one UAV-BS in this study, we have selected the single UAV-based localization (SUL) model of [11] for comparison. In the *SUL* model, the UAV-BS is placed at the middle of the UAV-BS height ranges which is 0.9 Km which is the average of the minimum (400 m) and maximum (1400 m) UAV-BS altitudes. The PDOP, RMSE, and PL metrics are then measured from this fixed altitude. As a second model, we have defined a *Basic* model where the UAV-BS is randomly placed within the minimum and maximum UAV-BS heights throughout all the simulation episodes. In the *Basic* model, the PDOP, RMSE, and PL metrics are computed from the random position the UAV-BS takes at each episode. For the proposed *UCL-RL* model, however, the UAV-BS learns the optimal UAV-BS height using the reinforcement learning algorithm through a continuous interaction with the environment.

Table 2 shows the simulation parameters. The simulation scenario consists of a 2.8284 km radius target area as shown in Fig. 4. Figure 4 illustrates one instance of the simulation scenario where the UAV-BS is placed at h = 1400. From this position, it moves in a circular path, generating navigation signals at each UAV position. Subsequently, the localization and communication services are measured. At another time step, the UAV-BS takes different UAV-BS heights according to the *UCL-RL* algorithm as depicted in

Table 2 Simulation parameters.

Parameters of the table	Values
UAV-BS speed	180 Km/h
UAV-BS turning radius	2000 m
UAV-BS height (Min, Max)	400 m, 1400 m
Δh	100 m
State space	11
Action space	3
Frequency	2 GHz
Learning rate	0.1
Discount factor	0.95
Number of terminals	1 and 20
Minimum received power	-80 dBm
UAV-BS transmitted power	15 W
Navigation signal transmission points	15

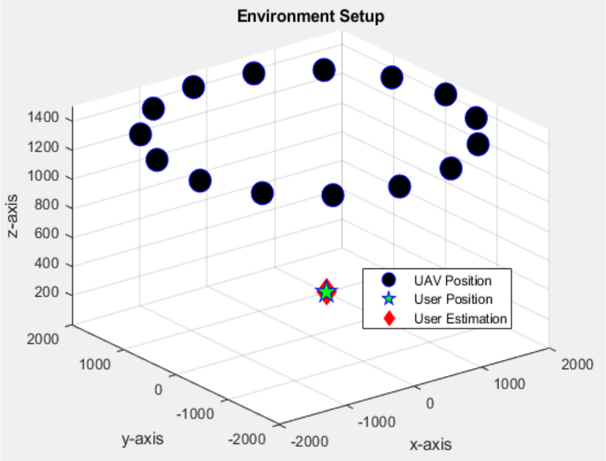


Fig. 4 Simulation scenario.

Fig. 3, to produce the required state, action, and reward. The agent continues interacting with the environment until it reaches a stopping condition.

There are two simulation scenarios: single user simulation scenario and multiple user simulation scenario. In the single user simulation scenario, one terminal ($K = 1$), which is an edge user, is used to evaluate the performance of the *SUL*, *Basic* and *UCL-RL* models. In the multiple user simulation scenario, 20 terminals ($K = 20$) are generated within the defined target area. Out of the 20 terminals, 19 terminals are randomly generated whereas 1 terminal is an edge user. The values for the learning rate $\mu = 0.1$ and the discount factor $\gamma = 0.95$ are selected because they are very common values in many of the Q-learning algorithm-based research works. The value for epsilon is initially 1 and decreases as the episode progresses to balance the exploitation and exploration strategies which are crucial to maximize the cumulative reward over time.

Figure 5 shows the PDOP simulation result for the *SUL*, *Basic* and *UCL-RL* models for the single user simulation scenario. Here, the dynamicity in the agent-environment interaction is the result of the change in altitude of the agent. The PDOP of the edge terminal is computed at each episode and serves as the reward. In Fig. 5, the PDOP for the *SUL* model doesn't vary throughout the episodes because it's measured from a fixed height. Initially, up until episode 13, the proposed *UCL-RL* model has worse PDOP value than the *SUL* and *Basic* models as it has not interacted with the environment and learned the best reward yet. As the agent-environment interaction proceeds (defined by the episodes), the proposed *UCL-RL* model has resulted in an improved PDOP value compared to the *SUL and Basic* models. After 24 episodes, the *UCL-RL* model has converged to the best reward, while the *SUL* model has fixed value, and the *Basic* model has random values in every episode.

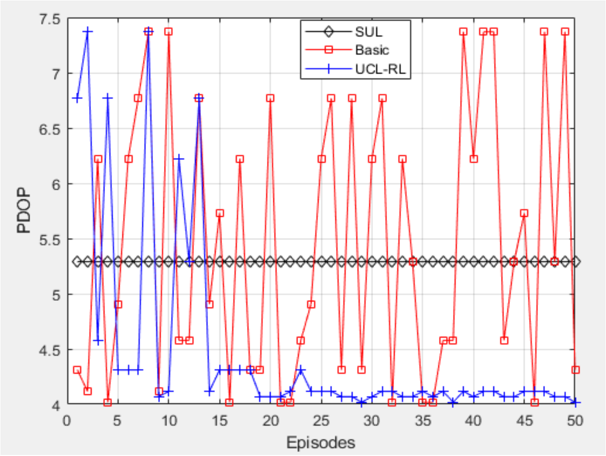


Fig. 5 PDOP comparison of SUL, Basic and UCL-RL models, K = 1.

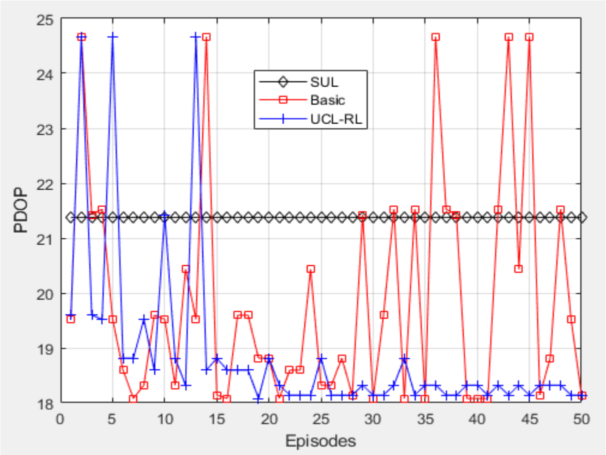


Fig. 6 PDOP$_{ave}$ comparison of SUL, Basic and UCL-RL models for K = 20.

Figure 6 shows the average PDOP simulation result for the *SUL, Basic* and *UCL-RL* models for the multiple user simulation scenario. The average PDOP for the terminals is computed at each episode and serves as the reward. Initially, like in the single user scenario, the average PDOP value for the *UCL-RL* model is worse than the PDOP value of the *SUL* and *Basic* models. As the episode progresses, however, the average PDOP value for the *UCL-RL* model has improved. Starting from episode 14, the *UCL-RL* model provides better average PDOP compared to the *SUL* and *Basic* models as shown in Fig. 6.

The PDOP range in the single-user simulation scenario

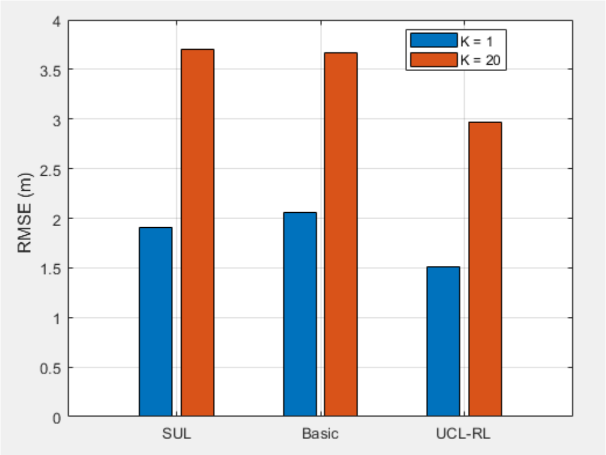


Fig. 7 RMSE comparison of SUL, Basic and UCL-RL models for K = 1 and K = 20 averaged over 50 episodes.

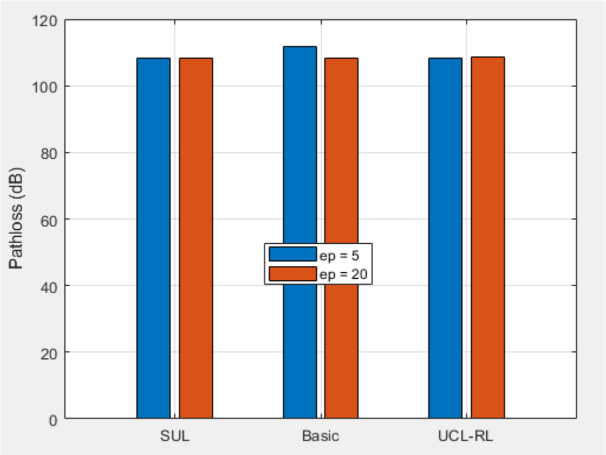


Fig. 8 MAPL comparison of SUL, Basic and UCL-RL models at episodes 5 and 20.

(PDOP = 4 to 7.5 in Fig. 5) is lower than the PDOP range in the multiple user simulation scenario (PDOP_{ave} = 18 to 25 in Fig. 6). This variation in PDOP value in the two simulation scenarios come from the different UAV-BS to terminal geometries. The UAV-BS to terminal geometry is an important factor that affects the PDOP value. It describes the geometry of the navigation signal transmission points (NSTP) of the UAV-BS from the ground terminals perspective. In the single-user simulation scenario, the geometry of the terminal depends on the NSTPs and height of the UAV-BS only. This provides better UAV-BS to terminal geometry which corresponds to the lower PDOP range. In the multiple user simulation scenario, however, there are many UAV-BS to terminal geometries, one for each terminal, which depend on the NSTPs and the height of the UAV-BS as well as the positions of the terminals. These multiple geometries result in a large average PDOP value at each episode. That is the reason why the PDOP range in the multiple-user simulation scenario is larger than the PDOP range in the single-user simulation scenario. In both simulation scenarios, the proposed *UCL-RL* model provides better PDOP value as the episode increases compared to the *SUL and Basic* models as illustrated in Fig. 5 and Fig. 6.

Another way to measure the accuracy of the terminal positioning is to apply root mean square error (RSME). Figure 7 shows the RMSE for the *SUL, Basic and UCL-RL* models for the two simulation scenarios averaged over 50 episodes. For $K = 1$, the RMSE values for the *SUL, Basic* and *UCL-RL* models are 1.91 m, 2.06 m and 1.51 m respectively. For $K = 20$, the RMSE values for the *SUL, Basic* and *UCL-RL* models are 3.70, 3.67 m and 2.97 m respectively. To compute the estimated positions of the terminals, we apply a non-linear least square method using the Levenberg-Marquardt algorithm [13]. The algorithm iteratively adjusts the estimated terminal positions, leading to reduced errors and minimized RMSE values. Consequently, despite the high PDOP values shown in Fig. 6, the RMSE values in Fig. 7 are small. The improvement is due to the effectiveness of the Levenberg-Marquardt algorithm in minimizing errors.

In both simulation scenarios, the *UCL-RL* model has provided smaller RMSE values compared to the *SUL* and *Basic* models. This shows the proposed *UCL-RL* model provides better terminal position accuracy as the agent learns the best parameters that minimize the positioning error of the terminals.

To evaluate the communication, the maximum allowable path loss (MAPL) metric is measured. The purpose of this evaluation is to show the path loss of the proposed *UCL-RL* model lies within the given path loss range defined by the maximum path loss, PL_{max}, which is the threshold path loss. According to the PLaHO model defined in [18] and given in Eq. (18), the MAPL for an UAV-BS to ground terminal communication in suburban environment is 110.4 dB which is the threshold path loss value. Figure 8 shows the path loss for the *SUL*, *Basic* and *UCL-RL* models at two episodes (ep = 5, and 20) for the single user simulation scenario ($K = 1$) by considering the MAPL for the suburban environment. The values of ep = 5 and ep = 20 are carefully selected to demonstrate learning properties of the agent. The lower episode, ep = 5, represents the learning process at the beginning of the learning. At this episode, the agent has had limited interaction with the environment. At ep = 20, the agent demonstrates substantial learning about the environment due to increased interaction. The characteristics of the other episodes are similar to these two episodes.

At ep = 5, the path loss values for the *SUL, Basic* and *UCL-RL* models are 108.15 dB, 111.94 dB and 108.15 dB respectively. At ep = 20, the path loss values are 108.15 dB, 108.12 dB, 108.74 dB for the *SUL, Basic* and *UCL-RL* models respectively. The proposed *UCL-RL* model has produced path loss value below the MAPL as the episode increases from 5 to 20 as shown in Fig. 8. There is a slight increase in the path loss value for the *UCL-RL* model when the episode increases from 5 to 20, but it is still less than the threshold path loss value (110.4 dB) which shows that the proposed

UCL-RL model maintains the communication service. This proves that the proposed *UCL-RL* model provides improved localization service while enabling communication service to the terminals in the target area when compared to the *SUL* and *Basic* Models.

5. Conclusion

This paper proposed the use of a single UAV-BS to provide unified communication and localization services in suburban environment with no cellular and GPS connectivity by applying reinforcement learning. The UAV-BS is flown to the target area and deployed within the minimum and maximum heights where it moves in a circular path to send navigation signals to the terminals in the target area. The combination of the UAV-BS turning radius, navigation signal transmission points, UAV-BS height, and the position of the ground terminals provides a dynamic environment. The UAV-BS interacts with the environment and learns the average PDOP value as a reward through the Q-learning algorithm. The path loss of an edge terminal is also measured to assess the communication service. Simulation results have shown that the proposed model provides improved terminal positioning accuracy while guaranteeing communication service.

In this work, the UAV-BS turning radius is constant. In our next work, we will design the problem by varying the turning radius and assess how it affects the localization and communication capabilities. In addition to that, we will expand the UAV-BS height range to increase the state space and then apply other reinforcement learning algorithms, like DQN, to evaluate the performance.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2007112).

References

[1] M. Erdelj, E. Natalizio, K.R. Chowdhury, and I.F. Akyildiz, “Help from the sky: Leveraging UAVs for disaster management,” IEEE Pervasive Comput., vol.16, no.1, pp.24–32, Jan.-March 2017.

[2] Y. Zeng, R. Zhang, and T.J. Lim, “Wireless communications with unmanned aerial vehicles: Opportunities and challenges,” IEEE Commun. Mag., vol.54, no.5, pp.36–42, May 2016.

[3] M. Mozaffari, W. Saad, M. Bennis, Y.H. Nam, and M. Debbah, “A tutorial on UAVs for wireless networks: Applications, challenges, and open problems,” IEEE Commun. Surveys Tuts., vol.21, no.3, pp.2334–2360, 2019.

[4] L. Gupta, R. Jain, and G. Vaszkun, “Survey of important issues in UAV communication networks,” IEEE Commun. Survey Tuts., vol.18, no.2, pp.1123–1152, 2016.

[5] A. Al-Hourani, S. Kandeepan, and S. Lardner, “Optimal LAP altitude for maximum coverage,” IEEE Wireless Commun. Lett., vol.3, no.6, pp.569–572, Dec. 2014.

[6] L. Wang, B. Hu, and S. Chen, “Energy efficient placement of a drone base station for minimum required transmit power,” IEEE Wireless Commun. Lett., vol.9, no.12, pp.2010–2014, Dec. 2020.

[7] N. Cherif, W. Jaafar, H. Yanikomeroglu, and A. Yongacoglu, “On the optimal 3D placement of a UAV base station for maximal coverage of UAV users,” IEEE Global Communications Conference, Taipei, Taiwan, pp.1–6, 2020.

[8] E. Çetin, C. Barrado, G. Muñoz, M. Macias, and E. Pastor, “Drone navigation and avoidance of obstacles through deep reinforcement learning,” 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), San Diego, CA, USA, pp.1–7, 2019.

[9] X. Liu, Y. Liu, and Y. Chen, “Reinforcement learning in multiple-UAV networks: Deployment and movement design,” IEEE Trans. Veh. Technol., vol.68, no.8, pp.8036–8049, Aug. 2019.

[10] S.P. Gopi and M. Magarini, “Reinforcement learning aided UAV base station location optimization for rate maximization,” Electronics, vol.10, no.23, p.2953, 2021.

[11] D.-H. Kim, K. Lee, M.-Y. Park, and J. Lim, “UAV-based localization scheme for battlefield environments,” MILCOM 2013 - 2013 IEEE Military Communications Conference, San Diego, CA, USA, pp.562–567, 2013.

[12] H. Ishikawa, Y. Horoikawa, and H. Shinonaga, “Maximum positioning error estimation method for detecting user positions with unmanned aerial vehicle based on Doppler shifts,” IEICE Trans. Commun., vol.E103-B, no.10, pp.1069–1077, Oct. 2020.

[13] K. Lee, H. Noh, and J. Lim, “Airborne relay-based regional positioning system,” Sensors, vol.15, no.6, pp.12682–12699, 2015.

[14] M. Atif, R. Ahmad, W. Ahmad, L. Zhao, and J.J.P.C. Rodrigues, “UAV-assisted wireless localization for search and rescue,” IEEE Syst. J., vol.15, no.3, pp.3261–3272, Sept. 2021.

[15] E. Testi, E. Favarelli, and A. Giorgetti, “Reinforcement learning for connected autonomous vehicle localization via UAVs,” 2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Trento, Italy, pp.13–17, 2020.

[16] R.B. Langley, “Dilution of precision,” GPS World, vol.10, no.5, pp.52–59, 1999.

[17] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, “Modeling air-to-ground path loss for low altitude platforms in urban environments,” 2014 IEEE global Communications Conference, Austin, TX, USA, pp.2898–2904, 2014.

[18] I. Mohammed, I.B. Collings, and S.V. Hanly, “A new connectivity model for unmanned aerial vehicle communications and flying height optimization,” Transactions on Emerging Telecommunications Technologies, vol.34, no.6, e4767, 2023.

[19] R.S. Sutton and A.G. Barto, Reinforcement Learning: An Introduction, 2nd ed., The MIT Press, 2018.

Gebreselassie Haile received B.S. in Electronics & Communications Engineering from Mekelle Institute of Technology, Ethiopia, in 2007 and M.S. in computer Engineering from Ajou University, Korea, in 2013. During 2007–2020, he was with Information Network Security Administration of Ethiopia where he was involved in Telecom & Network signal Analysis, Speech Compression, and Wireless Network Audit. Starting from March 2020, he is a Ph.D. student at Ajou University, Korea, in the department of Artificial Intelligence Convergence Network. His research interests are Resource Management in Wireless Networks, UAV for Communication & Localization, and Machine Learning for Wireless Networks.

Jaesung Lim received B.S. in electronic engineering from Ajou University, Korea, in 1983, and M.S. and Ph.D. in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), in 1985 and 1994, respectively. In 1985, he started as a researcher at Daewoo Telecommunication. In April 1988, he joined the institute of DigiCom, and was engaged in research and development of data modem, radar signal processing and packet data systems. From 1995 to 1997, he served as a senior engineer in the Central Research and Development Center of SK Telecom. Since March 1998, he has been with Ajou University where he is a professor of the department of military digital convergence teaching and doing research in the areas of wireless, mobile, and tactical communications and networks.