Abstract—In this paper, a training design and channel estimation scheme is considered for uplink cloud radio access networks (C-RANs) consisting of multiple user equipments (UEs), remote radio heads (RRHs), and a centralized baseband unit (BBU) pool. Since most signal processing functions in C-RANs are moved from RRHs to the BBU pool, the individual channels over the links between UEs and RRHs and the links between RRHs and the BBU pool cannot be estimated directly. To address this issue, segment training based individual channel estimation for C-RANs is proposed in this paper, in which channel state information acquisition is performed through two consecutive segments. By using the Kalman filter, the sequential minimum mean-square-error (SMMSE) estimator is developed to efficiently estimate the individual channel states through prior knowledge of long-term channel correlation statistics and the latest radio channel state. A training structure design subject to a power constraint is obtained by minimizing the mean-square-error (MSE) of the SMMSE estimator. Since the MSE is insufficient to fully evaluate the overall performance of C-RANs, the uplink ergodic capacity is derived to exploit the impact of channel estimation on the data transmission by taking the estimation errors into consideration, and the tradeoff between the lengths of two segment training sequences is optimized by maximizing the corresponding spectral efficiency. Furthermore, the Cramér-Rao bound is used to evaluate the proposed SMMSE estimator’s performance. Simulation results show that the SMMSE estimator and the corresponding training design can effectively decrease MSE and significantly increase the quality and efficiency of data transmission in C-RANs.

Index Terms—Cloud radio access networks, channel estimation, training design, Cramér-Rao bound.

I. INTRODUCTION

Nowadays mobile Internet is undergoing unprecedented growth with the ubiquity of smart phones and the support of mobile broadband services provided by fourth generation (4G) networks. To address this situation, cloud radio access networks (C-RANs) are promising solutions as they exploit the benefits of incorporating cloud computing technology into wireless networks [1]. In current cellular networks, base stations (BSs) serve only their own users instead of sharing their processing capacity in a coordinated way. Compared to traditional cellular networks, C-RANs concentrate the processing functions into the cloud server to achieve global resource scheduling and management in an energy-efficient way [2]. Meanwhile, the radio access functions are handed over to the distributed remote radio heads (RRHs) to provide ubiquitous coverage [3]. To enable reception at the cloud server, baseband signals are transferred between RRHs and the baseband unit (BBU) pool by use of low-latency fronthaul links [4].

Motivated by the abundant advantages of C-RANs, considerable research effort has been dedicated to understanding the relevant technologies feasible for C-RANs. The constrained fronthaul links connecting RRHs with the BBU pool present the main hurdle to the rollout of C-RANs, and thus compression techniques have been extensively studied in [5]. Additionally, various other aspects of C-RANs such as large-scale precoding design [6] and resource allocation optimization [7] have also been considered. However, the acquisition of instantaneous channel state information (CSI) remains a challenging issue and this is the subject of the current work.

To achieve the significant performance gains of C-RANs, most previous work has assumed that the instantaneous CSI for all radio access links and fronthaul links is perfectly known at the BBU pool [8]. As a matter of fact, the knowledge of instantaneous CSI is often non-ideal due to the time-varying nature of the radio channel and the constrained overhead in real communication systems [9]. However, instantaneous CSI acquisition is necessary for the BBU pool to mitigate inter-RRH interference, design compression for the constrained fronthaul links, and perform large-scale precoding and decoding. Consequently, the channel estimator design in order to obtain accurate CSI is vital for improving the performance of C-RANs [10].

In C-RANs, there are two communication links: one is the radio access link (ACL) for communication between the user equipment (UE) and the RRH, and the other is the fronthaul link connecting the RRH to the BBU pool. To save costs and
make the deployment of RRHs more flexible, the fronthaul is often wireless, and thus the wireless fronthaul link (WFL) is a key component of C-RANs. Generally, knowledge of the composite CSI of ACLs and WFLs is not sufficient to achieve optimal system designs in certain scenarios in which the individual CSI is instead needed. For instance, the BBU pool needs to know the individual CSI of ACLs and WFLs to design beamforming schemes, similarly to relay networks [11]. Therefore, developing the corresponding channel estimation strategy to obtain the individual CSI of ACLs and WFLs is critical.

The traditional training design is not an efficient approach to acquire the individual CSI due to the high training overhead on the constrained WFLs and the significant performance degradation resulting from the CSI feedback. To tackle this problem, a superimposed training scheme is expected to be a candidate technique for channel estimation for C-RANs, in which the training overhead is greatly reduced by superimposing training sequences directly on data symbols [12]. However, the superimposed training suffers from severe estimation quality degradation since both training and data transmissions have to share the signal-to-noise ratio (SNR) originally allocated solely for data transmission.

To improve the data transmission rate, a segment training based individual channel estimation technique has been proposed in [13], in which two consecutive segments are used for the individual CSI acquisition of ACLs and WFLs. In [14], segment training based channel estimation in amplify-and-forward one-way relay networks has been fully exploited by developing the effective estimators and the corresponding training power allocation strategy. However, the objective function selected for optimal training design in terms of power allocation is not suitable because only the estimation error of the channel from source to relay is considered. In [15], the segment training scheme has been successfully applied in C-RANs, in which a superimposed-segment training scheme has been proposed to reduce the training overhead. Unfortunately, the data transmission quality is not high and the channel estimation accuracy is not efficient due to the fact that there are segments shared by both data and training transmissions in the proposed training scheme.

In this paper, segment training based channel estimation and the corresponding training design for C-RANs are examined, in which the instantaneous and individual CSI of composite links combining ACLs and WFLs is obtained at the centralized BBU pool with the signal processing overhead of RRHs greatly alleviated. The sequential minimum square-error (SMMSE) estimator is developed by making use of the Kalman filter to improve the channel estimation accuracy, and the corresponding training design subject to the power constraint is derived based on the mean-square-error (MSE) minimization criterion. In order to fully exploit how the channel estimation affects the overall performance of uplink C-RANs, a lower bound on the ergodic capacity for the joint maximum-ratio combining (MRC) and zero-forcing (ZF) detection is derived. By maximizing the lower bound and improve the corresponding spectral efficiency in C-RANs, the trade-off between training and data symbols is optimized. To evaluate the proposed estimator, the Cramér-Rao bound (CRB) under segment training scheme in C-RANs is derived as well. The main contributions of this paper are as follows:

- By exploiting the segment training design scheme for the SMMSE estimator, the CSI of ACLs originally acquired at RRHs are estimated at the BBU pool with a minimal amount of additional signal processing at RRHs.
- In order to be closer to the practical channel model, WFLs are assumed to be correlated in time and space domains. Meanwhile, the Kalman filter is applied to improve the performance of channel estimation for WFLs by tracking the channel variation over time according to prior knowledge of the temporal correlation coefficient.
- For both ACLs and WFLs, the training sequence structures are derived by minimizing the MSE of channel estimation with the power constraint. In particular, the tradeoff between the lengths of two segment training sequences is optimized by maximizing the spectral efficiency in the uplink transmission to improve the overall system performance.
- To evaluate the performance of the proposed SMMSE estimator, the CRBs of the individual CSI for ACLs and WFLs under the segment training design scheme are derived to provide a benchmark of channel estimation errors for the SMMSE estimator.

The rest of this paper is organized as follows. In Section II, the system model of C-RANs with the corresponding training design scheme and data transmission strategy will be presented. The sequential MMSE channel estimation strategy using the Kalman filter will be developed in Section III. According to the proposed SMMSE estimator, the problem of training design in terms of two segment lengths will be discussed in Section IV. Furthermore, the derivation of CRBs will be introduced in Section V. The numerical results will be presented in Section VI, followed by the conclusions in Section VII. For clarity of presentation, the derivations have been placed in an appendix.

**Notation:** Scalars are denoted by small letters; Vectors and matrices are denoted by boldface small and capital letters, respectively; the transpose, complex conjugate, Hermitian, and inverse of the matrix $A$ are denoted by $A^T$, $A^*$, $A^H$, and $A^{-1}$, respectively; $\text{tr}(A)$ is the trace of $A$ and $\text{diag}(A)$ denotes a diagonal matrix constructed from diagonal elements of $A$; $\text{vec}(A)$ is the vectorization of $A$; $[A]_{ij}$ represents the $(i, j)$-th element of $A$ and $I_M$ is an $M \times M$ identity matrix; $|a|$ denotes the 2-norm of the vector $a$; $\odot$ and $\otimes$ stand for the Hadamard product and Kronecker product, respectively; $\mathbb{E}\{\cdot\}$ denotes the statistical expectation; $[\cdot]_N$ represents modulo $N$. 

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2539126, IEEE Transactions on Signal Processing
II. SYSTEM MODEL

As shown in Fig. 1, the considered uplink C-RAN consists of $K$ UEs and $M$ RRHs, and all RRHs associate with the BBU pool via WFLs. It is assumed that all $M$ RRHs cooperate with one another. Each of UEs and RRHs is equipped with a single antenna, all RRHs are assumed to be perfectly synchronized. Synchronization is reasonable to assume, since there are a number of frequency offset solutions available. For instance, the protocol in [16] and post-facto timing synchronization [17] for distributed antenna and virtual multiple-input multiple-output (MIMO) systems, can be effective solutions to achieve synchronization among RRHs. The estimation of block-fading channels and the corresponding uplink data transmission are considered. The block-fading channel is assumed to have a coherence time of $L$. The first $T < L$ symbols of each fading block are dedicated for training with the remaining $L-T$ symbols for data communication. As indicated from Fig. 1, there is no direct link between UEs and the BBU pool, and the data transmission needs to go through two individual channel links. Generally, the cascaded CSIs of ACLs and WFLs would be adequate for subsequent signal processing in the BBU pool. However, the individual CSIs of ACLs and WFLs are essential in certain cases as the BBU pool needs to perform optimal system design, e.g., large-scale centralized beamforming design. Denote $G_{1}$ as the $M \times K$ channel matrix of the ACLs from $K$ UEs to $M$ distributed RRHs. The channel matrix $G_{1}$ is spatially uncorrelated, which can be represented as

$$G_{1} = \Gamma_{1} \otimes H_{1},$$

where $H_{1}$ is an $M \times K$ matrix of fast fading coefficients, and $\Gamma_{1}$ denotes an $M \times M$ matrix of large-scale fading coefficients, where $[\Gamma_{1}]_{i,j} = \sqrt{\kappa_{ij}}$ is the $(i,j)$-th element of $\Gamma_{1}$. The elements in $H_{1}$ are uncorrelated circularly symmetric complex Gaussian random variables with zero means and unit variances. $\kappa_{ij}$ is assumed to be time-invariant over many coherence time intervals and known a priori. This latter assumption is reasonable as the large-scale fading coefficients can be estimated due to the channel reciprocity in time division duplexing (TDD) systems or by using experimentally justified empirical models [18].

For WFLs, $N$ centralized antennas are deployed at the BBU pool for enhancing signal reception. Denote $G_{2}$ as the $N \times M$ channel matrix of the WFLs, which is given by

$$G_{2} = H_{2} \Gamma_{2}^{1/2},$$

where $H_{2}$ is the conventional $N \times M$ fast fading channel matrix, and $\Gamma_{2}^{1/2}$ is an $M \times M$ diagonal large-scale fading matrix with $\sqrt{\beta_{i}}$ as the $i$-th diagonal element. Similarly to $\kappa_{ij}$, $\beta_{i}, i = 1, \ldots, M$ is assumed to be time-invariant and known a priori. Unlike the ACLs, the large-scale fading channel matrix of the WFLs is assumed to be in a diagonal form.

$$H_{2} = R_{c}H_{w},$$

where $R = R_{c}R_{c}^{H}$ stands for the Cholesky decomposition, $R$ is an $N \times N$ receive correlation matrix known at the BBU pool, and $H_{w}$ is an $N \times M$ Gaussian random matrix with independent and identically distributed (i.i.d. zero mean and unit variance) entries. As the spatial correlation matrix $R$ is a Hermitian positive definite matrix, it can be decomposed as $R = U\Lambda^{1/2}U^{H}$, where $U$ is the eigenvector matrix and $\Lambda = \text{diag}(\lambda_{0}, \lambda_{0}, \ldots, \lambda_{0, N})$ is the eigenvalue matrix in descending order.

By taking the temporal correlation into account, $H_{2}$ is also correlated among the fading blocks. Denote $h_{2,i}$ as the vectorization of $H_{2}$ in the $i$-th fading block $(i = 0, 1, \ldots)$, where vec $\{H_{2}\} = (I \otimes R_{c})h_{w}$, and $h_{w}$ is the vectorization of $H_{w}$. To clarify the state-space model of $h_{2,i}$, the Gauss-Markov model [19] is used, which is a convenient way of modeling the time evolution of $\{h_{2,i}\}$. Then, following the Gauss-Markov model, $h_{2,i}$ can be written in a state-space form [20] as

$$h_{2,0} = (I \otimes R_{c})v_{0},$$

$$h_{2,i} = \eta h_{2,i-1} + \sqrt{1 - \eta^2}(I \otimes R_{c})v_{i}, i \geq 1,$$
where $0 \leq \eta \leq 1$ is a temporal correlation coefficient, $v_i \sim \mathcal{CN}(0, I_{Mx})$ is an innovation process, and $h_{i,0}$ is independent of $v_i$ for all $i \geq 1$. A detailed derivation of the Gaussian-Markov channel model can be found in the appendix of [21].

### A. Segment Training Transmission

In order to acquire the individual CSI ($H_{r1}$ and $H_{r2}$) of ACLs and WFLs, respectively, a two-phase segment training design scheme, as illustrated in Fig. 2, is presented. Each phase takes up an individual time slot. During phase I, UEs transmit the training sequence $\phi_1$ and data symbols $s$ to RRHs. $\phi_1$ and $s$ are the corresponding received signals of $\phi_1$ and $s$ at RRHs, respectively. In phase II, RRHs append another training sequence $\phi_2$ in front of $\phi_1$, and then forward the re-organized signals to the BBU pool. $\phi_1$ and $\phi_2$ are assumed to occupy $\tau_1$ and $\tau_2$ time slots, respectively. The sum of $\tau_1$ and $\tau_2$ is equal to $T$. The received training signal at RRHs during phase I is given by

$$Y_{r1} = G_{r1} \Psi + N_{r1}, \quad (6)$$

where $\Psi$ is a $K \times \tau_1$ transmitted training matrix stacked by $\phi_1$, and the elements in $N_{r1}$ are additive white Gaussian noise (AWGN) processes with zero means and variances $\sigma^2_n$. By vectorizing the received training signal, we have

$$y_{r1} = \text{vec}(Y_{r1}), \quad \Phi_{r1} = \text{vec}(H_{r1}), \quad \text{and} \quad n_{r1} = \text{vec}(N_{r1}).$$

During phase II, each RRH prepends its own $M \times \tau_2$ training matrix $\Phi$ (stacked by $\phi_2$) to the received training signal in a time-division multiple access model (TDMA) fashion. The overall training transmitted by RRHs can be expressed as

$$\Psi_1 = [\Phi, C Y_{r1}], \quad (8)$$

where $\Psi_1$ is an $M \times T$ matrix, and $C$ is an $M \times M$ transformation matrix known at RRHs.

The application of $C$ can save RRHs from the complicated channel estimation and leave all signal processing functions to the more powerful BBU pool. As a consequence, RRHs need only to forward the combinational training signals to the BBU pool. The received training signal at the BBU pool can be partition into two parts:

$$Y_r = F G_{r2} \Phi + F N_r, \quad (9)$$

$$Y_d = G_{r2} C G_{r1} \Psi + G_{r2} C N_{r1} + N_{r2}, \quad (10)$$

Denote $Y_{r,i}$ as the received training signal $Y_r$ in the $i$-th fading block. Using the vectorization operator, $Y_{r,i}$ can be rewritten as

$$y_i = (\Phi_i^T \Gamma_i^2) \otimes h_{r,i} + (I \otimes F) n_{r,i}, \quad i \geq 0, \quad (11)$$

where $y_i = \text{vec}(Y_{r,i})$, $h_{r,i}$ and $n_{r,i}$ are the corresponding vectorization of $H_{r2}$ and $N_r$ in the $i$-th fading block, respectively.

### B. Data Transmission

Similar to the training transmission above, the uplink data transmission in C-RANs can also be divided into two phases. In phase I, UEs transmit the $K \times 1$ signal vector $s$ to RRHs, and $s$ satisfies the power constraint $\mathcal{E}\{s^H s\} = p_s I_K$, where $p_s$ is the transmit power bound at each UE. Then, the corresponding received signal at RRHs can be written as

$$r_1 = G_{r1} s + n_1, \quad (12)$$

where $n_1 \sim \mathcal{CN}(0, \sigma^2_n I_M)$ denotes the additive noise vector at RRHs. It is assumed that the power at RRHs is constrained, and the corresponding power control factor $\varrho$ can be derived as

$$\varrho = P_s / \text{tr} \{p_s G_{r1} G_{r1}^H + \sigma^2_n I_M\}, \quad (13)$$

where $P_s$ is the maximal allowed power at RRHs.

In phase II, RRHs forward $r_1$ to the BBU pool, and the received signal at the BBU pool can be written as

$$r_2 = \sqrt{\varrho} G_{r2} G_{r1} s + \sqrt{\varrho} G_{r2} n_1 + n_2, \quad (14)$$

where $n_2$ is an $N \times 1$ AWGN vector at RRHs with variance $\sigma^2_n$.

### III. SMMSE Channel Estimation

In this section, the SMMSE channel estimation algorithm is developed along with the corresponding training design for $H_{r1}$ and $H_{r2}$. The individual channel estimation of ACLs and WFLs is derived relying on the Kalman filter and the eigenvalue decomposition (EVD). According to the MSE minimization criterion, the optimal training of the SMMSE estimator is proposed as well.

#### A. Training Design for $H_{r2}$

In most previous works with respect to the MMSE estimator, the estimation of $h_{r2,i}$ is based only on the current received training signal $y_i$ with the previously received training signals $\{y_k\}_{k=0}^{i-1}$ discarded. As (4) denotes the channel evolution, the estimation of $h_{r2,i}$ is similar to the state prediction in dynamical systems. Relying on the Kalman filter, an accurate estimate is provided by using the entire received training signals $\{y_k\}_{k=0}^{i}$ and the channel statics $n$. Denote $h_{r2,i_1} \{i_2\}$ as the estimated value of $h_{r2,i}$ given $\{y_k\}_{k=0}^{i_2}$ for $i_1 \geq i_2$. Let $\tilde{\Phi}_i = (\Phi_i^T \Gamma_i^2) \otimes F$. Then, the
Algorithm 1 Sequential MMSE estimation

1: **Initialization:**
   The initial estimate: \( \hat{h}_{r,2,0|-1} = 0 \),
   The initial covariance matrix: \( \mathbf{R}_{0|-1} = \mathbb{E}\{h_{r,2,0}h_{r,2,0}^H\} = I_{M} \otimes \mathbf{R} \).
2: **Prediction:** \( \hat{h}_{r,2,i|i-1} = \eta \hat{h}_{r,2,i|-1} \),
3: **Minimum prediction covariance matrix:**
   \( \mathbf{R}_{i|i-1} = \eta^2 \mathbf{R}_{i-1|i-1} + (1 - \eta^2)(I_{M} \otimes \mathbf{R}) \).
4: **Optimized estimation of the current state:**
   The Kalman gain matrix:
   \( \mathbf{K}_i = \mathbf{R}_{i|i-1} \mathbf{P}_i \left( \sigma_i^2 \mathbf{I}_{N_{t_2}} + \mathbf{P}_i \mathbf{R}_{i|i-1} \mathbf{P}_i^T \right)^{-1} \),
   The optimized estimation:
   \( \hat{h}_{r,2,i|i} = \hat{h}_{r,2,i|i-1} + \mathbf{K}_i \left( y_i - \Phi_i \hat{h}_{r,2,i|i-1} \right) \).
5: **Minimum covariance matrix:**
   \( \mathbf{R}_{i|i} = \left( I_{MN} - \mathbf{K}_i \Phi_i \right) \mathbf{R}_{i|i-1} \).

SMME estimation based on the Kalman filter is shown in Algorithm 1.

To obtain the optimal training signal for the channel estimate of \( h_{r,2} \), the optimization of \( \Phi_i \) in the \( i \)-th fading block is formulated as
\[
\Phi_i = \min_{\Phi_i} \mathcal{M}_i \\
\text{s.t. } \text{tr} \left( \mathbf{X}_i \right) \leq \tau_2 \mathbf{P}_r, \\
\mathbf{X}_i \succeq 0,
\]
where \( \mathbf{X}_i = \Phi_i \Phi_i^T \succeq 0 \) is the positive semi-definite constraint, and \( \mathcal{M}_i = \text{tr} \left( \mathbf{R}_{i|i} \right) \) is the MSE in the \( i \)-th fading block.

**Lemma 1:** The optimal training signal \( \Phi_i \) obtained from minimizing the MSE in the \( i \)-th fading block is orthogonal with the maximum allowable transmit power.

**Proof:** See Appendix A. \( \blacksquare \)

If the optimal \( \Phi_i \) is applied, the MSE of the \( i \)-th fading block (\( i \geq 1 \)) can be generalized as
\[
\mathcal{M}_i = \text{tr} \left\{ \mathbf{R}_{i|i-1} - \left( \sigma_i^2 \mathbf{I}_{N_{t_2}} + \mathbf{P}_i \mathbf{R}_{i|i-1} \mathbf{P}_i^T \right) \left( \mathbf{I} + \mathbf{P}_i \mathbf{R}_{i|i-1} \mathbf{P}_i^T \right)^{-1} \right\} \\
= \sum_{j=1}^{N} \lambda_{i,j} \left( 1 - \sum_{p=0}^{i-1} \sum_{k=1}^{\tau_2} \eta^2 (j-p) \tau_2 p \beta_{p} (j-1) \lambda_{p,(j-1),k}^2 \right),
\]
where \( \lambda_{i,j} \) is the \( j \)-th dominant eigenvalue of \( \mathbf{R}_{i|i-1} \).

**Corollary 2:** If \( \tau_2, \mathbf{P}_r \) and the correlation matrix \( \mathbf{R} \) are fixed, then
\[
\mathcal{M}_i < \mathcal{M}_{i+1}, i \geq 0.
\]

The MSE in (16) shows that the estimation accuracy can be improved with the longer training sequences. While, allocating more symbols for the training degrades the data transmission rate as the channel coherence time \( T \) is fixed.

It is important to make the tradeoff between the lengths of training and data symbols to improve the overall system performance. Corollary 2 demonstrates that the performance of the SMME estimation becomes better as the fading block \( i \) increases. Therefore, the BBU pool can rely on the Kalman filter to achieve an accurate estimate.

Through the Kalman filter, the estimate of \( \hat{h}_{r,2,i} \) is obtained using the previously received training sequences \( \{y_k\}_{k=0}^{\tau_1} \).

With prior knowledge of the received training sequences \( \{y_k\}_{k=0}^{\tau_1} \), a more accurate estimate can be achieved using the forward-backward (FB) Kalman filter [22]. The FB Kalman filter consists of forward run and backward run. The forward run is the same as the standard Kalman filter. The backward run is executed after receiving all \( \tau_1 \) symbols. Here, a new variable \( \tau_{r,1} + \tau_1 = 0 \) is defined to better describe the backward run. By a backtrack calculation of \( \tau_{r,1} \) for \( i = \tau_1, \tau_1 - 1, \ldots, 0 \), the estimate of \( h_{r,2,i|i} \) can be calculated as
\[
\tau_{r,1} = \eta \left( I - \eta \Phi_i \mathbf{P}_i^T \mathbf{K}_i \right) \mathbf{P}_{i+1|T} + \Phi_i \mathbf{R}_{e,i}^{-1} \left( y_i - \Phi_i \hat{h}_{r,2,i|i-1} \right),
\]
(18)

where \( \mathbf{R}_{e,i} = \left( \sigma_i^2 \mathbf{I}_{N_{t_2}} + \Phi_i \mathbf{P}_i \Phi_i^T \right) \). In this way, a more accurate estimate of \( h_{r,2,i} \) is obtained, although considerable latency and storage are required. Compared to the FB Kalman filter, the standard Kalman filter can estimate the channel within one OFDM symbol, i.e., with no latency and also give an accepting estimation accuracy. Thus, the standard Kalman filter is chosen in our SMMSE estimator.

**B. Training Design for** \( \mathbf{H}_{r,2} \)

Given the channel estimate from phase I, \( \mathbf{G}_{r,2} \) is assumed to be known at the BBU pool and the estimate of \( \mathbf{G}_{r,2} \) is near perfect. Therefore, the estimation error of \( \mathbf{G}_{r,2} \) is neglected and has little impact on the estimation of \( \mathbf{G}_{r,1} \). The received signals during phase II can be vectorized as
\[
y_d = \text{vec}(\mathbf{Y}_d) = (\mathbf{P}^T \otimes \mathbf{G}_{r,2} \mathbf{C}) \mathbf{D}^{\frac{1}{2}} \mathbf{h}_{r,1} \\
+ (\mathbf{I} \otimes \mathbf{G}_{r,2} \mathbf{C}) \mathbf{n}_{r,1} + \mathbf{n}_{r,2}.
\]
(20)

For the fixed \( \mathbf{Y} \) and \( \mathbf{C} \), the covariance matrix of estimation error for \( \mathbf{h}_{r,1} \) is calculated as [23]
\[
\mathbf{R}_{\delta h_{r,1}} = \mathbf{I} - \mathbf{D}^{\frac{1}{2}} \tilde{\mathbf{Y}} \mathbf{P} \mathbf{Y}^T \mathbf{P} \mathbf{Y}^T \mathbf{D}^{-\frac{1}{2}} \\
+ \sigma^2 \mathbf{I}_{\tau_1} \otimes \mathbf{G}_{r,2} \mathbf{C} \mathbf{C}^H \mathbf{G}_{r,2}^{-1} \mathbf{P} \mathbf{Y}^T \mathbf{D}^{-\frac{1}{2}},
\]
(21)
where \( \tilde{\mathbf{Y}} = \mathbf{P}^T \otimes \mathbf{G}_{r,2} \mathbf{C} \).

The optimal structure of \( \mathbf{Y} \) should be obtained from the minimization of the covariance \( \mathbf{R}_{\delta h_{r,1}} \) with respect to \( \mathbf{Y} \) and \( \mathbf{C} \) subject to the power constraints both at UEs and RRHs, which yields
\[
\min_{\mathbf{Y}, \mathbf{C}} \mathcal{M} = \text{tr} \left( \mathbf{R}_{\delta h_{r,1}} \right) \\
\text{s.t. } \text{tr} \left( \mathbf{Y} \mathbf{Y}^T \right) \leq \tau_1 \mathbf{P}_s, \\
\mathbb{E} \left\{ \text{tr} \left( \mathbf{y}_s \mathbf{y}_s^H \right) \right\} \leq \tau_1 \mathbf{P}_r,
\]
(22)
where \( y_s = \text{vec}(C Y_r) \) and \( P_s = K p_s \) is the power bound at UEs.

The problem above is non-convex and the Karush-Kuhn-Tucker (KKT) conditions are not effective. In order to solve the problem, we resort to the EVD with the training and transformation matrix decomposed into unitary components and diagonal components, respectively. The EVDs of \( \Psi^T \Psi^* \) and \( G_{r2} C C^H G_{r2}^H \) are given by

\[
\Psi^T \Psi^* = U_1 \Lambda_1 U_1^H, \quad (23)
\]
\[
G_{r2} C C^H G_{r2}^H = U_2 \Lambda_2 U_2^H, \quad (24)
\]

where \( U_1 \) and \( U_2 \) are unitary eigenvector matrices, and \( \Lambda_1 \) and \( \Lambda_2 \) are diagonal eigenvalue matrices with elements in descending order. Then, the following equations can be obtained:

\[
\Psi^T = U_1 \Lambda_1^\frac{1}{2} Q_1, \quad (25)
\]
\[
G_{r2} C = U_2 \Lambda_2^\frac{1}{2} Q_2, \quad (26)
\]

where \( Q_1 \) and \( Q_2 \) are unitary matrices. The singular value decomposition (SVD) of \( G_{r2} \) is \( U_g \Lambda_g Q_g^H \) with the singular values in descending order, where \( U_g \) and \( Q_g \) are unitary singular vector matrices.

**Lemma 3**: The optimal solution to unitary matrices subject to the training design problem in (22) is that \( Q_1 = I \), \( Q_2 = I \), \( U_2 = U_g^H \) and \( U_1 \) is an arbitrary unitary.

**Proof**: See Appendix B.

Denote \( \lambda_1(i) \), \( \lambda_2(i) \) and \( \lambda_g(i) \) as the \( i \)-th diagonal elements of \( \Lambda_1 \), \( \Lambda_2 \), \( \Lambda_g \), respectively. The rank of \( G_{r2} \) is assumed to be full rank. The training design problem above can be further expressed as

\[
\begin{align*}
\max_{\lambda_1(i) \geq 0, \lambda_2(j) \geq 0} & \sum_{i=1}^{\tau_1} \sum_{j=1}^{N} \frac{\kappa_{ij} \lambda_1(i) \lambda_2(j)}{\sigma_n^2 + \sigma_{\nu}^2 \lambda_2(j) + \kappa_{ij} \lambda_1(i) \lambda_2(j)}, \\
st. & \sum_{i=1}^{\tau_1} \lambda_1(i) \leq \tau_1 P_s, \\
& \sum_{i=1}^{\tau_1} \sum_{j=1}^{N} \frac{\kappa_{ij} \lambda_1(i) \lambda_2(j)}{\lambda_g(j)^2} + \tau_1 \sigma_n^2 \sum_{j=1}^{N} \frac{\lambda_2(j)}{\lambda_g(j)^2} \leq \tau_1 P_r.
\end{align*}
\]  

(27)

The optimization over \( \lambda_1 \) and \( \lambda_2 \) in (27) is a non-convex problem. If \( \lambda_1(i) \) is fixed for all \( i \)'s, the optimization over \( \lambda_2 \) is converted to a convex problem subject to the second power constraint in (27), and vice versa. Then, \( \lambda_1 \) and \( \lambda_2 \) are sequentially optimized by alternatively solving two sub-optimization problems until a locally optimal solution to (27) is obtained. Convergence is also guaranteed as the cost function in (27) is upper bounded and increases with each iteration until a local optimum is found.

The solution to two sub-optimization problems can be given as follows. Firstly, \( \lambda_2(j) \) is optimized with \( \lambda_1(i) \) fixed. Let the fixed \( \lambda_1(i), i = 1, \ldots, \tau_1 \) satisfy \( \sum_{i=1}^{\tau_1} \lambda_1(i) = \tau_1 P_s \). Then, the training design problem is only subject to RRHs power constraint and can be solved by using the KKT conditions

\[
\begin{align*}
\sum_{i=1}^{\tau_1} \frac{\kappa_{ij} \lambda_1(i)}{\lambda_g(j)^2} = \nu \left[ \sum_{i=1}^{\tau_1} \kappa_{ij} \lambda_1(i) \lambda_g(j)^{-2} + \tau_1 \sigma_n^2 \lambda_g(j)^{-2} \right], \\
\sum_{i=1}^{\tau_1} \sum_{j=1}^{N} \kappa_{ij} \lambda_1(i) \lambda_g(j)^{-2} \lambda_2(j) + \tau_1 \sigma_n^2 \sum_{j=1}^{N} \frac{\lambda_2(j)}{\lambda_g(j)^2} = \tau_1 P_r,
\end{align*}
\]  

(28)

where \( \zeta = 1 + \lambda_2(j) + \kappa_{ij} \lambda_1(i) \lambda_g(j)/\sigma_n^2, \nu > 0, \) and \( \lambda_2(j) \) is either zero or a positive value. For any given \( \nu > 0, \lambda_2(j) \) is either equal to a positive solution to the first equation above or is set to zero. The optimal \( \nu \) can also be found from the second equation by substituting the \( \lambda_2(j) \) obtained above.

In the next, the optimization of \( \lambda_1(i) \) with the fixed \( \lambda_2(j), j = 1, 2, \ldots, N \) is discussed. The Lagrangian function is given by

\[
\sum_{j=1}^{N} \kappa_{ij} \lambda_2(j) (1 + \lambda_2(j)) / \lambda_g(j)^2 = \nu_1 + \nu_2 \sum_{j=1}^{N} \kappa_{ij} \lambda_2(j) / \lambda_g(j)^2,
\]  

(30)

where \( \nu_1 \geq 0 \) and \( \nu_2 \geq 0 \) are the Lagrange multipliers for the first and second constraints, respectively. For any given pair of \( \nu_1 \) and \( \nu_2 \), \( \lambda_1(i) \) is obtained by solving the equation above or simply equal to zero. To search the optimal pair of \( \nu_1 \) and \( \nu_2 \), one can be determined by a bisection search with the other one fixed when the two power constraints are active, namely, \( \nu_1 > 0 \) and \( \nu_2 > 0 \). In other case, \( \nu_1 = 0 \) or \( \nu_2 = 0 \) is possible, which should be examined firstly.

It can be seen that the optimal design of \( \Psi \) and \( C \) depends on the uplink channel \( G_{r2} \) due to the term \( G_{r2} C N_{r1} \) in (10). To avoid a design that depends on the instantaneous channel state, the expectation of \( R_{h_{\nu_1}} \) over \( G_{r2} \) can be used. Then, \( H_{r1} \) can be designed by maximizing the expectation of the second term on the RHS of (21):

\[
\hat{M}_1 = \mathbb{E} \{ \text{tr} \left( \tilde{\Psi} D \tilde{\Psi}^H \right) (\sigma_n^2 I_{N_{r_1}} + \tilde{\Psi} D \tilde{\Psi}^H) + \sigma_n^2 I_{\tau_1} \otimes (G_{r2} C C^H G_{r2}^H) \}.
\]  

(31)

Here, we assume that \( D = I \) and \( F_2 = I \) for the sake of tractability. Then, \( \hat{M}_1 \) can be converted to

\[
\begin{align*}
\hat{M}_1 &= \mathbb{E} \{ \text{tr} \left( (A_1 \otimes G_{r2} C C^H G_{r2}^H) (\sigma_n^2 I_{N_{r_1}} + (A_1 + \sigma_n^2 I_{\tau_1}) \otimes (G_{r2} C C^H G_{r2}^H) \} \\
&= \mathbb{E} \{ \text{tr} (\sigma_n^2 A_1^{-1} \otimes (G_{r2} C C^H G_{r2}^H)^{-1}) + I + \sigma_n^2 A_1^{-1} \otimes I \}^{-1}. \}
\end{align*}
\]  

(32)

It would be rather difficult to directly calculate the matrix inversion. Here, the standard Taylor series expansion is used to approximate the matrix inversion as [24]

\[
X^{-1} = \alpha \sum_{l=0}^{\infty} (I - \alpha X)^l \approx \alpha \sum_{l=0}^{L} (I - \alpha X)^l,
\]  

(33)
where $X$ is a positive-definite Hermitian matrix and $\alpha$ is a scaling factor satisfying $0 < \alpha < 2/\lambda_{\max}(X)$. $\lambda_{\max}(X)$ is the largest eigenvalue of $X$. Using this approximation, $\mathcal{M}_1$ can be expressed as

$$\mathcal{M}_1 = \mathbb{E}\{\text{tr}(\sum_{t=0}^{L} \alpha(I - \alpha S)^t)\},$$  \hspace{1cm} (34)$$

where $S = \sigma_n^2 A_1^{-1} \otimes (G_{r2} C C^H G_{r2}^H)^{-1} + (I + \sigma_n^2 A_1^{-1}) \otimes I$. Here, the second-order approximation can be chosen by setting $L$ to be 1. After re-organization, the maximization of $\mathcal{M}_1$ is approximately equivalent to minimizing

$$\mathcal{M}_2 = \mathbb{E}\{\text{tr}(A_1^{-1} \otimes S_1)\},$$  \hspace{1cm} (35)$$

where $S_1 = I + (G_{r2} C C^H G_{r2}^H)^{-1}$. According to the properties of a central Wishart matrix [25], the expectation of $S_1$ is $\text{tr}\left(\left(CC^H\right)^{-1}\right) \frac{N}{M-1} I$. The optimization problem in (22) can be approximated as

$$\min_{\Lambda_1, C} \mathcal{M}_2 = \text{tr}(A_1^{-1} \otimes S_1) $$

subject to

$$\text{tr}(A_1) \leq \tau_1 P_s,$$

$$\text{tr}\left\{(A_1 + \sigma_n^2 I) \otimes CC^H\right\} \leq \tau_1 P_r.$$  \hspace{1cm} (36)$$

Similarly to the way of optimizing (27), $C$ can be optimized by fixing the matrix $A_1$. Then, (36) can be directly converted to the minimization of $\text{tr}\left\{(CC^H)^{-1}\right\}$ subject to the power constraint $\text{tr}\left\{(CC^H)\right\} \leq \frac{P_r}{P_s \sigma_n^2}$. The formula $\text{tr}(J^{-1}) \geq \sum_{i=0}^{M} (J_{ii})^{-1}$ holds, in which $J$ is an arbitrary $M \times M$ positive definite matrix and equality holds if and only if $J$ is a diagonal matrix [23]. Based on this inequality, the minimum of $\text{tr}\left\{(CC^H)^{-1}\right\}$ can be achieved if $CC^H$ has the diagonal structure. $A_1$ can also be optimized by fixing the matrix $C$. As $CC^H$ and $A_1$ both have the diagonal structure, it can concluded that the suboptimal design of $\Psi$ and $C$ is orthogonal. This training design is independent of instantaneous channel values, thus, it can be a practical choice for channel estimation in C-RANs.

IV. TRAINING LENGTH DESIGN

The MSE minimization is an effective criterion to evaluate the performance of the channel estimator and the training design, while it can’t fully depict the overall performance of C-RANs. Moreover, the effect of channel estimation to the data detection is significant due to the inverse relationship between the channel estimation error and SNR [26]. In this section, the ergodic capacity of the uplink data transmission with the joint MRC-ZF detection in C-RANs is derived, which is also adopted as the objective for the optimal training design. Considering the channel estimation errors, the channel matrices $G_{r1}$ and $G_{r2}$ can be expressed as

$$\hat{G}_{r1} = G_{r1} + \Omega_1,$$

$$\hat{G}_{r2} = G_{r2} + \Omega_2,$$  \hspace{1cm} (37)$$

$$\hat{G}_{r1} = G_{r1} + \Omega_1,$$

$$\hat{G}_{r2} = G_{r2} + \Omega_2,$$  \hspace{1cm} (38)$$

where $G_{r1}$ and $G_{r2}$ are the estimated CSI of ACLs and WFLs, and $\Omega_1$ and $\Omega_2$ are the corresponding channel estimation errors, respectively.

Owing to the properties of SMMSE estimation, $\Omega_1$ and $\Omega_2$ are independent of $G_{r1}$ and $G_{r2}$, respectively. The optimal training derived for ACLs and WFLs in Section III is applied. As $A_1$ and $A_2$ are not obtainable in closed-form, it is assumed that the orthogonal training is used at UEs and RRHs for the estimation of ACLs. The large-scale fading coefficients of $G_{r1}$ from the $k$-th UE to RRHs are assumed to be equal to $\sqrt{\kappa_k}$. The estimation error matrix $\Omega_2$ is obtained from the first fading block and the spatial correlation of $G_{r2}$ is ignored as well. Then, the elements in the $i$-th column of $\Omega_2$ are random variables (RVs) with zero means and variances $\frac{\sigma^2_n}{\sqrt{\kappa_k} p_r / M}$, where $p_r = P_r / M$ is the power bound at each RRH.

For the estimation error matrix $\Omega_1$, the orthogonal source training $\Psi H = \tau_1 p_s I$ is applied with co-training matrix $C$ set as $\sqrt{\Omega_M}$. Since the projection of $Y_d$ onto $\Psi^H$ is sufficient to acquire the MMSE estimate of $G_{r1}$, we have

$$Y = \frac{1}{\sqrt{\tau_1 p_s}} Y_d \Psi^H = \sqrt{\tau_1 p_s} G_{r1} G_{r1}^H + \sqrt{\tau_1 p_s} G_{r2} V_1 + V_2,$$  \hspace{1cm} (39)$$

Then, the MMSE estimate of $G_{r1}$ can be derived as

$$\hat{G}_{r1} = G_{r1} + \frac{1}{\sqrt{\tau_1 p_s}} V_1 + \frac{1}{\sqrt{\tau_1 p_s}} G_{r1}^H \hat{V}_2.$$  \hspace{1cm} (40)$$

Assuming that the linear receiver is applied at the BBU pool, the received signal after using the linear detector is given by

$$r = \sqrt{\beta} W^H G_{r2} (G_{r1} s + n_1) + W^H n_2,$$  \hspace{1cm} (41)$$

where $W$ is a $K \times N$ linear detector matrix. The conventional MRC-ZF linear detector [27] is considered, and the explicit form of $W$ can be written as

$$W^H = G_{r1}^H \hat{G}_{r2}^\dagger.$$  \hspace{1cm} (42)$$

By substituting (31), (32) and (44) into (43) and omitting some relatively small terms, the received signal can be simplified as

$$r = \sqrt{\beta} Y s + \hat{n},$$  \hspace{1cm} (43)$$
where
\[ \mathbf{Y} = \mathbf{G}_{r1}^{H} \mathbf{G}_{r1}, \] (46)
and
\[ \hat{n} = \sqrt{\frac{\varrho}{\tau_{1}} \mathbf{V}_{1}^{H} \mathbf{G}_{r1} \mathbf{s}} + \sqrt{\frac{1}{\tau_{1} p_{s}} \mathbf{V}_{2}^{H} (\mathbf{G}_{r2})^{H} \mathbf{G}_{r1} \mathbf{s}} \]
\[ - \sqrt{\varrho} \mathbf{G}_{r1}^{H} \mathbf{G}_{r1}^{\dagger} \mathbf{n}_{1} + \sqrt{\varrho} \mathbf{G}_{r1}^{H} \mathbf{G}_{r2}^{\dagger} \mathbf{n}_{2}. \] (47)

For a fixed channel realization \( \mathbf{G}_{r1} \) and \( \mathbf{G}_{r2} \), the covariance matrix of the equivalent noise vector \( \hat{n} \) is given by
\[ \mathcal{E} \{ \hat{n} \hat{n}^{H} \} = \begin{bmatrix} \frac{\varrho \sigma_{n}^{2}}{\tau_{1}} \text{tr}(\mathbf{Y}) + \frac{\sigma_{n}^{2}}{\tau_{1}} \text{tr}(\mathbf{P}_{1}) & \varrho \sigma_{n}^{2} \mathbf{Y} \\ \varrho \sigma_{n}^{2} \mathbf{Y}^{\dagger} & \mathbf{P}_{2} \end{bmatrix}, \] (48)
where \( \mathbf{P}_{1} = \mathbf{G}_{r1}^{H} (\mathbf{G}_{r2}^{H} \mathbf{G}_{r2})^{\dagger} \mathbf{G}_{r1} \), \( \mathbf{P}_{2} = \mathbf{G}_{r1}^{H} (\mathbf{G}_{r2}^{H} \mathbf{G}_{r2})^{\dagger} \mathbf{G}_{r1} \), and \( \Theta \) is an \( M \times M \) diagonal matrix with the \( m \)-th diagonal element to be \( \frac{\sigma_{m}^{2} \beta_{m}}{\tau_{1}} \).

By modeling \( \hat{n} \) as Gaussian noise independent of \( \mathbf{s} \), the instantaneous SNR for the \( k \)-th UE at the BBU pool can be expressed by (49) shown at the top of this page. The ergodic achievable uplink rate of the \( k \)-th UE is given by
\[ R_{k} = \mathcal{E} \{ \log_{2}(1 + \gamma_{k}) \}. \] (50)

According to Jensen’s inequality and the convexity of (50), a lower bound on the ergodic achievable uplink rate of the \( k \)-th UE can be written as
\[ R_{k} \geq \tilde{R}_{k} = \log_{2} \left( 1 + \frac{1}{\mathcal{E}[1/\gamma_{k}]} \right). \] (51)

To simplify the derivation, a fixed power control factor \( \varrho \) is selected to maintain the average power of RRHs during the data transmission. Then, the power control factor for the MRC-ZF receiver is given by
\[ \varrho = p_{s} / \left( p_{s} \text{tr}(\mathbf{S}) + \sigma_{n}^{2} \right), \] (52)
where \( \mathbf{S} = \text{diag} \{ \kappa_{1}, \kappa_{2}, \ldots, \kappa_{K} \}. \)

As \( \mathbf{G}_{r2} \) and \( \hat{\mathbf{G}}_{r2} \) are independent of \( \mathbf{G}_{r1} \), the expectation in (51) is taken over \( \mathbf{G}_{r2} \) and \( \hat{\mathbf{G}}_{r2} \) firstly, and the corresponding result is presented in the following proposition.

**Proposition 4:** The lower bound on the uplink achievable ergodic rate is given by (53) shown at the top of the next page, where \( \hat{\mathbf{P}}_{1} = \frac{1}{\varrho \sigma_{n}^{2}} \mathbf{G}_{r1}^{H} \mathbf{G}_{r2}^{\dagger} \mathbf{G}_{r1} \) and \( \hat{\mathbf{P}}_{2} = \frac{1}{\varrho \sigma_{n}^{2}} \mathbf{G}_{r1}^{H} \mathbf{G}_{r2}^{\dagger} \mathbf{G}_{r1} \).

**Proof:** See Appendix C.

Let \( g_{k} \) be the \( k \)-th column of \( \mathbf{G}_{r1} \), and the following lemma is held to further simplify the lower bound on the uplink rate.

**Lemma 5:** In the presence of imperfect CSI and for \( M \geq 3 \), the achievable ergodic uplink rate of the \( k \)-th user for the MRC-ZF receiver is lower bounded by
\[ \tilde{R}_{k} = \log_{2} \left( 1 + \frac{\varrho p_{s} (M-1)(M-2) \kappa_{k}^{2}}{a_{1}(M-1)(M-2) \kappa_{k}^{2} + a_{2}(M-2) \kappa_{k} + a_{3}} \right), \] (54)
where
\[ a_{1} = \frac{\varrho p_{s} \sigma_{n}^{2}}{(N-M) \tau_{2} p_{s} \beta}, \]
\[ a_{2} = \frac{M \varrho p_{s} \sigma_{n}^{2}}{(N-M) \tau_{2} p_{s} \beta} \sum_{i=1, i \neq k}^{K} \kappa_{i} + (1 + \tau_{1}) \varrho \sigma_{n}^{2} \]
\[ + \frac{\sigma_{n}^{2}}{(N-M) \tau_{1} \beta} + \frac{\sigma_{n}^{2} \beta_{1}}{(N-M) \tau_{2} p_{s} \beta}, \]
\[ a_{3} = M \left( \frac{\varrho \sigma_{n}^{2}}{\tau_{1}} + \frac{\sigma_{n}^{2}}{(N-M) \tau_{1} \beta} \right) \sum_{i=1, i \neq k}^{K} \kappa_{i}, \]
and \( \beta \) is the smallest large-scale fading coefficient among all \( \beta_{m}, m = 1, \ldots, M \).

**Proof:** See Appendix D.

Then, the spectral efficiency with imperfect CSI is given by
\[ R_{s} = \frac{L - T}{L} \sum_{k=1}^{K} \tilde{R}_{k}. \] (58)

In (58), SNRs among UEs vary vastly due to different large-scale coefficients incorporated in (49), making the spectral efficiency difficult to yield a tractable expression suitable for choosing the training length. In principle, training length design can be straightforwardly derived by letting \( \tau_{1} \) and \( \tau_{2} \) dependant on \( k \). In this situation, new tradeoffs are introduced, e.g., the fairness between users near and far from RRHs needs to addressed. This tradeoff leads to another scheduling tradeoff of fairness against the total throughput, which is not the main point of our analysis in this paper. For the sake of tractability, it is assumed that the uplink rate of the \( k \)-th UE is the smallest, and a lower bound on the spectral efficiency can be derived as
\[ R_{s} \geq \tilde{R}_{s} = \frac{L - T}{L} K \tilde{R}_{k}. \] (59)

Then, the objective function can be reasonably chosen to maximize the lower bound on the spectral efficiency, namely, \( \tilde{R}_{k} \). Let us replace \( \tau_{2} \) by \( T - \tau_{1} \), and focus on finding the optimal \( \tau_{1} \) in terms of maximizing \( \tilde{R}_{k} \). Since \( \tilde{R}_{k} \) has only
γ factor, the derivative of efficiency with the optimal training can be written as

\[
\lim \frac{\partial \gamma}{\partial \tau_1} = \frac{1}{\tau_1 \ln 2} \frac{\partial \gamma_k}{\partial \tau_1}.
\]

(60)

Since the roots of \(\frac{\partial \gamma_k}{\partial \tau_1} = 0\) are equal to those of \(\frac{\partial \gamma_k}{\partial \tau_1} = 0\), it is more convenient to take the derivative of \(\gamma_k\) with respect to \(\tau_1\). After some tedious calculation and omitting the common factor, the derivative of \(\gamma_k\) with respect to \(\tau_1\) can be simplified as

\[
\frac{\partial \gamma_k}{\partial \tau_1} = \frac{\omega}{[\tau_1 + 1] \tau_1^2 - \tau_1^2} [c_1 (T - \tau_1)^2 - \tau_1^2] - [c_1 (T - \tau_1)^2 + c_2 (T - \tau_1) \tau_1] \frac{1}{(N-M)^2} \left[ \sum_{i=1}^{K} \tilde{\kappa}_i - \tilde{\kappa}_k + \frac{\sigma_n^2}{\varpi_p^2 \beta} \right],
\]

(61)

where

\[
\xi = \frac{(M-2)\tilde{\kappa}_k \varpi_p \sigma_n^2}{(N-M) \varpi_p^2}, \quad \omega = \frac{\varpi_p (M-1)(M-2)\tilde{\kappa}_k^2}{(N-M) \varpi_p},
\]

(62)

(63)

\[
c_2 = \frac{1}{\xi} \left[ \frac{\varpi_p^2 + \frac{\sigma_n^2}{(N-M) \beta}}{(N-M) \beta} \right] (M-2) \tilde{\kappa}_k,
\]

(64)

\[
c_1 = \frac{c_2 + M}{\xi} \left( \frac{\varpi_p^2 + \frac{\sigma_n^2}{(N-M) \beta}}{(N-M) \beta} \right) \sum_{i=1, i \neq k}^{K} \tilde{\kappa}_i.
\]

(65)

It can be readily derived that there are two roots through solving the equation \(\frac{\partial \gamma_k}{\partial \tau_1} = 0\) are \(\tau_{1,1} = \frac{\sqrt{\sigma} + \bar{\beta}}{\sqrt{\sigma} - \bar{\beta}}\) and \(\tau_{1,2} = \frac{\sqrt{\sigma} - \bar{\beta}}{\sqrt{\sigma} + \bar{\beta}}\). Moreover, it can also be affirmed that the optimal training length \(\tau_{1,2}\) is the local minimum, and the optimal training length \(\tau_1\) that maximizes \(R_k\) is given by

\[
\tau_{1,}\text{opt} = \frac{T}{1 + \sqrt{c_1}}.
\]

(66)

Substituting (66) into (59), the lower bound on the spectral efficiency with the optimal training can be written as

\[
\hat{R}_k = \frac{L - T}{L} K \log_2 \left( 1 + \frac{\omega_1 T}{T + \omega_2} \right),
\]

(67)

where

\[
\omega_1 = \frac{\varpi_p}{c_2} (M-1)(M-2) \tilde{\kappa}_k^2, \quad \omega_2 = \left( \sqrt{c_1 + \frac{1}{\varpi_p}} \right)^2.
\]

(68)

Then, the derivative of \(\hat{R}_k\) with respect to \(T\) can be expressed as

\[
\frac{\partial \hat{R}_k}{\partial T} = \frac{K \omega_1 \omega_2 (L-T)}{L(T + \omega_2)(T + \omega_1 + \omega_2)} - \frac{K}{L} \log_2 \left( 1 + \frac{\omega_1 T}{T + \omega_2} \right).
\]

(69)

(69) indicates that \(\lim_{T \to 0} \frac{\partial \hat{R}_k}{\partial T} > 0\) and \(\lim_{T \to L} \frac{\partial \hat{R}_k}{\partial T} < 0\). Hence, through solving the equation \(\frac{\partial \hat{R}_k}{\partial T} = 0\), one root in the interval \([0, L]\) can be derived, which must be the local maximum. Nonetheless, the closed-form solution to the equation \(\frac{\partial \hat{R}_k}{\partial T} = 0\) may not be obtained directly, and we should resort to standard numerical techniques to directly obtain the optimal symbol length \(T\) allocated for training.

Remark: With fixed \(T\), the training length design of \(\tau_1\) and \(\tau_2\) is equivalent to the power allocation between two training segments. If equal numbers of training symbols are assumed to be assigned to both the training matrices \(\Phi_1\) and \(\Phi\), i.e., \(\tau_1 = \tau_2 = \frac{T}{2}\), the optimal training length design for training matrix \(\Phi_1\) can be replaced by the optimal power allocation, which yields

\[
\Phi_1 = \left[ \sqrt{\varepsilon} \Phi, \sqrt{2 - \varepsilon} \text{CY}_{r_1} \right],
\]

(70)

where \(\varepsilon = \frac{2}{\sqrt{T}}\).

V. CRAMÉR-RAO BOUND

In this section, the CRB is analyzed to evaluate the performance of the proposed estimator. The numbers of training symbols for \(\Psi\) and \(\Phi\) are set to be equal, i.e., \(\tau_1 = \tau_2 = \tau\), and \(\Psi_1\) in (70) is applied at RRHs.

Define \(\theta = [h_{r_1}^H, h_{r_2}^H]^T\) as the unknown channel parameters to be estimated, where \(h_{r_0} = \text{vec}(H_{r_0})\). Let \(y = [y_r^T, y_d^T]^T\) represents the received training signal during the training phase, where \(y_r = \text{vec}(Y_r)\).

The Fisher information matrix (FIM) is define as [28]

\[
\mathcal{I} = \mathcal{E} \left\{ \frac{\partial \ln p(y, \theta)}{\partial \theta^*} \left( \frac{\partial \ln p(y, \theta)}{\partial \theta^*} \right)^H \right\},
\]

(71)

where the expectation in (71) is taken over the joint probability density function \(p(y, \theta)\).
For a deterministic \( \theta \), the FIM can be written as

\[
\mathcal{F} = \mathcal{E} \left\{ \frac{\partial \ln p(y_r|\theta)}{\partial \theta} \left( \frac{\partial \ln p(y_d|\theta)}{\partial \theta^*} \right)^H \right\}
\]

Using the property of the Kronecker product, the following equation can be derived

\[
\mathcal{E} \left\{ \frac{\partial \ln p(y_r|\theta)}{\partial \theta} \left( \frac{\partial \ln p(y_d|\theta)}{\partial \theta^*} \right)^H \right\} = \begin{bmatrix} \mathcal{F}_{11} & \mathcal{F}_{12} \\ \mathcal{F}_{21} & \mathcal{F}_{22} \end{bmatrix},
\]

(72)

Conditioned on a specific channel \( \theta \), the probability density functions \( p(y_r|\theta) \) and \( p(y_d|\theta) \) can be expressed as

\[
p(y_r|\theta) = \frac{1}{(\pi \sigma_n^2)^N} \exp \left\{ -\frac{(y_r - u_r)^2}{\sigma_n^2} \right\},
\]

(73)

\[
p(y_d|\theta) = \frac{1}{\pi N \mathbb{E} \left[ R_{n,i|\theta} \right]} \exp \left\{ -\frac{(y_d - u_d)^2}{\mathbb{E} \left[ R_{n,i|\theta} \right]} \right\},
\]

(74)

where \( u_r = \sqrt{\varepsilon} \left( (\Phi^T \Gamma_2^1) \otimes \mathbf{F} \mathbf{r} \right) \mathbf{h}_w \), \( u_d = \sqrt{2 - \varepsilon} \left( (\Psi^T \otimes \mathbf{G}_2 \mathbf{C}) \mathbf{D}^1 \right) \mathbf{h}_r, \)

and

\[
R_{n,i|\theta} = \sigma_n^2 \left[ (2 - \varepsilon) \mathbf{I}_r \otimes \mathbf{G}_2 \mathbf{C} \mathbf{H}_r \mathbf{G}_2^H + \mathbb{I}_{N \tau} \right].
\]

From [28], the \((i, j)\)-th elements of \( \mathcal{F}_1 \) and \( \mathcal{F}_2 \) are given by

\[
\mathcal{F}_{1,i,j} = \frac{1}{N \sigma_n^2} \frac{\partial \mathbf{h}_w}{\partial \theta^*} \frac{\partial \mathbf{u}_r}{\partial \theta^*},
\]

(76)

\[
\mathcal{F}_{2,i,j} = \frac{1}{N \sigma_n^2} \frac{\partial \mathbf{h}_w}{\partial \theta^*} \frac{\partial \mathbf{u}_d}{\partial \theta^*} \frac{\partial \mathbf{h}_r}{\partial \theta^*}
\]

\[
+ \mathbf{t} \left( \frac{\partial \mathbf{R}_n|\theta}{\partial \theta^*} \frac{\partial \mathbf{R}_n|\theta}{\partial \theta^*} \right).
\]

(77)

Let \( \mathcal{F}_2 = \mathcal{F}_s + \mathcal{F}_r \), and \( \mathcal{F}_r \) is a new matrix whose \((i, j)\)-th element is the second term in \( \mathcal{F}_{2,i,j} \). Using the property of the Kronecker product, the following equation can be derived as

\[
(\Psi^T \otimes \mathbf{G}_2 \mathbf{C}) \mathbf{D}^1 \mathbf{h}_r = \left( (\Psi^T \otimes \mathbf{G}_2 \mathbf{C}) \mathbf{D}^1 \otimes \mathbf{R}_r \right) \mathbf{h}_w.
\]

(78)

Using the above equation, the derivatives of \( \mathbf{u}_r \) and \( \mathbf{u}_d \) with respect to \( \mathbf{h}_w \) and \( \mathbf{h}_r \) are computed as

\[
\frac{\partial \mathbf{u}_r}{\partial \mathbf{h}_w} = \sqrt{\varepsilon} \left( (\Phi^T \Gamma_2^1) \otimes \mathbf{F} \mathbf{r} \right),
\]

(79)

\[
\frac{\partial \mathbf{u}_d}{\partial \mathbf{h}_w} = 0_{N \tau \times MK},
\]

(80)

\[
\frac{\partial \mathbf{u}_r}{\partial \mathbf{h}_r} = \sqrt{2 - \varepsilon} \left( (\Psi^T \otimes \mathbf{G}_2 \mathbf{C}) \mathbf{D}^1 \right),
\]

(81)

\[
\frac{\partial \mathbf{u}_d}{\partial \mathbf{h}_r} = \sqrt{2 - \varepsilon} \left( (\Psi^T \otimes \mathbf{G}_2 \mathbf{C}) \mathbf{D}^1 \right).
\]

(82)

The derivatives of \( R_{n,i|\theta} \) with respect to \( h_{r1,i} \) and \( h_{w,i} \) are given by

\[
\frac{\partial R_{n,i|\theta}}{\partial h_{r1,i}} = 0_{N \tau \times N \tau},
\]

(83)

\[
\frac{\partial R_{n,i|\theta}}{\partial h_{w,i}} = \sigma_n^2 (2 - \varepsilon) \left( I_r \otimes E_i \mathbf{C} \mathbf{H}_r^H \right),
\]

(84)

where \( E_i \) is an \( N \times M \) matrix with the \((i - N \tau | i |, i | N \tau + 1)\)-th element set to be one and other elements to be zero. Then, we have

\[
[N \tau \times N \tau],
\]

where \( B_i = I_r \otimes E_i \mathbf{C} \mathbf{H}_r^H \).

Using the expressions above, \( \mathcal{F}_{11} \) can be explicitly expressed in a compact form

\[
\mathcal{F}_{11} = \varepsilon \left( \Gamma_2^1 \mathbf{F}^2 \right) \mathbf{D}^1 \mathbf{h}_r = (\mathbf{2} - \varepsilon) \Xi R_{n,i|\theta} \mathbf{F}_v + \mathbf{F}_r,
\]

(85)

where \( \Xi = \left( \mathbf{C} \mathbf{H}_r^H \mathbf{R}_n|\theta \right) \mathbf{D}^1 \mathbf{h}_r \).

The CRB is given by \( \text{CRB}_\theta = \text{tr} \left( \mathcal{F}^{-1} \right) \). Then, the CRBs for \( \mathbf{h}_w \) and \( \mathbf{h}_{r1} \) can be separately expressed as

\[
\text{CRB}_{\mathbf{h}_w} = \sum_{i=1}^{NM} \left| \mathcal{F}^{-1} \right|_{ii},
\]

(89)

\[
\text{CRB}_{\mathbf{h}_{r1}} = \sum_{i=N \tau + 1}^{NM+MK} \left| \mathcal{F}^{-1} \right|_{ii}.
\]

(90)

The CRB bounds derived in (89) and (90) are difficult to express in closed-form, and thus they are not useful as alternative criteria for designing the training by minimizing the CRB bounds. Nonetheless, they still can be used as lower bounds on the attainable MSE of unbiased estimators, by which the performance of the proposed estimator can be evaluated.

VI. NUMERICAL RESULTS

In this section, we numerically evaluate the performance of our proposed training design with the SMMSE estimator under different scenarios. The power bounds for UEs and RRHs are set to be equal, i.e., \( p_s = p_r = P \). The noise variance \( \sigma_n^2 \) is set to one, and the SNR is defined as \( P/\sigma_n^2 = P \). We assume \( K = M = 4 \) and \( N = 8 \). Totally \( 10^5 \) Monte-Carlo iterations are adopted. Considering that the comparison of the different training sequence lengths is numerically difficult since only integers can be selected, we assume that the lengths of training sequences \( \phi_1 \) and \( \phi_2 \) are equal, i.e., \( \tau_1 = \tau_2 = \tau \), and the optimal training design
equivalently depends on the power allocation factor $\varepsilon$. For the estimation of $H_{r2}$, each iteration contains 10 temporally and spatially correlated fading blocks. The elements of the correlation matrix $R$ are given as $|R_{i,j}| = r^{\|i-j\|}$, $|r| = 0.2$. The temporal correlation coefficient $\eta$ is set at 0.9881 [29]. The normalized MSEs of $H_{r1}$ and $H_{r2}$ are defined as $\frac{1}{T}E_{H_{r1}}[M_{r2}]$ and $\frac{1}{T}E_{H_{r2}}[M_{r2}]$, respectively. The expectation of the normalized MSE of $H_{r1}$ is computed by taking the average over 100 realizations of $H_{r2}$. To simplify the large-scale fading, the large-scale fading coefficients $\beta_m$ and $\kappa_{km}$ are modeled as $\beta_m = z_m/(r_m/\tau_0)^v$ and $\kappa_{km} = z_{km}/(r_{km}/\tau_0)^v$, respectively, where $z_m$ and $z_{mk}$ are the corresponding log-normal random variables with the standard deviation $\delta$ dB. $r_m$ is the distance between the $m$-th RRH and the BBU pool, and $r_{km}$ is the distance between the $k$-th user and the $m$-th RRH. $r_0$ is set at 100 m, and the path loss exponent $v = 3.8$.

The normalized MSE performance of $H_{r2}$ under different correlation coefficient values $|\rho|$ is evaluated in Fig. 3 to verify the optimality of orthogonal training ($|\rho| = 0$) in the estimation of $H_{r2}$. The first fading block with $i = 0$ and SNR = 0 dB. The correlation coefficients is defined as $|\rho|_{ij} = |\phi_{2}(i)\phi_{2}(j)|^2$, and the correlation coefficients of any two training sequences are set to be equal, i.e., $|\rho|_{ij} = |\rho|$, $i \neq j$. The optimal power allocation coefficient $\varepsilon$ given in (70) is applied. As $|\rho|$ increases, the normalized MSE performance degrades, and the lowest normalized MSE is achieved when $|\rho| = 0$ for the different total lengths of training sequences $T$. The simulation result shows that orthogonal training is the optimal, which is consistent with the theoretical analysis in Lemma 1. For the different total lengths of training sequences $T$, we see that the longer the training sequence is, the better MSE performance can be achieved. In order to improve the estimation accuracy, we can lengthen the training sequence, which degrades the system spectral efficiency in return. Therefore, the tradeoff of symbols allocated for training and data transmission is important for the overall system performance improvement.

In Fig. 4, we depict the normalized MSE curves versus fading block index $i$ under different total training sequence lengths $T$ ($T = 14$ and $T = 16$) with fixed SNRs, which is set to be 0 dB. Two power allocation schemes are compared. For the optimal power allocation, the value of $\varepsilon$ is generated from (70), and $\varepsilon$ is simply set to be one for the equip power allocation. Orthogonal training is assumed to be applied for both $\Phi$ and $\Psi$. It is easy to verify from Fig. 3 that the performance of SMMSE estimation becomes better as the fading block index $i$ increases. This is reasonable since the estimation of $H_{r2}$ is based on the previously received training symbols and the temporal correlation is taken into consideration to improve estimation accuracy. The MSE curves adopting the optimal power allocation are lower than those with the equip power allocation, which demonstrates that more power is allocated to $\Phi$ for the estimation of $H_{r2}$. Comparing with the MSE curves of $T = 14$, the larger $T$ can provide better MSE performances.

The normalized MSE of $H_{r2}$ versus SNR for the SMMSE estimator using the Kalman filter and the FB Kalman filter is evaluated in Fig. 5, where CRBs are presented as benchmarks and orthogonal training is used for all MSE curves. The length of training is set at 12. The Kalman filter and FB Kalman filter under optimal power allocation, the Kalman filter under equip power allocation, and CRB schemes are compared. The SMMSE estimator with the FB Kalman filter is considered only under the optimal power allocation scheme. It can be observed that the SMMSE estimator achieves lower MSE under the optimal power allocation scheme than under the equip-power allocation scheme, which matches the observations depicted in Fig. 4. Moreover, the normalized MSEs of $H_{r2}$ under optimal training are close to their corresponding CRBs. The SMMSE estimator using the FB Kalman filter achieves the best performance in terms of MSEs, verifying its ability to improve the performance of channel estimation. As the SNR increases, the FB Kalman filter provides better MSE performance than the Kalman filter does when SNR is relatively low.

Fig. 6 shows the normalized MSE of $H_{r1}$ versus SNRs
under different structures for the transformation matrix $C$. The CRB is adopted as a benchmark and the optimal power allocation scheme is applied for the estimation of $H_{r1}$. Here, optimal design means that $C$ is derived according to (28)-(30). For the sub-optimal scheme, $C$ is orthogonal and satisfies $C^H C = \frac{\rho}{\bar{\kappa} \sigma^2 + \sigma^2} I$, where $\bar{\kappa}$ is the average value of all $\kappa_{ij}$’s. As expected, the MSE of $H_{r1}$ is a decreasing function of SNRs, and the optimal design yields better MSE performance than the sub-optimal design does. It can be seen that the MSE curves are relatively worse than the corresponding CRBs in the high SNR region. This happens because the training power allocated to the estimation of $H_{r1}$ is relatively small and the estimation of $H_{r1}$ suffers from severe fading.

In Fig. 7, the performance of uplink rate per UE versus training length $\tau_1$ is evaluated to gain a better understanding of optimal training length design and its influence on the overall system performance. Three different SNRs are considered in this situation, i.e., 0 dB, 5 dB, and 10 dB. $T$ is fixed at 16. The spatial and temporal correlation coefficients are set to zero, and the optimal training structures for the estimation of $H_{r1}$ and $H_{r2}$ are adopted herein. The theoretical optimal training lengths $\tau_1$ are marked by triangles. As observed, when $\tau_1$ is small, the uplink rate monotonically increases with $\tau_1$ until it reaches the maximum point. After that, the uplink rate becomes a monotonically decreasing function of $\tau_1$. It can be seen that the optimal $\tau_1$ stays in a position close to the location of the theoretical optimal $\tau_1$, which indicates that the theoretical analysis matches quite well with the numerical results. The optimal training design lies not only in the training structure design, but also in the optimal training length design, which is also necessary for the overall system performance improvement.

In Fig. 8, the bit error rate (BER) performance of data detection at the BBU pool under three different power allocation schemes is presented, where the transmitted data are modulated with SPSK, and conventional maximum likelihood detection is used at the BBU pool. $T$ is set to be 16. The BER curves under random power allocation, equi-power allocation, optimal power allocation, and perfect CSIs are compared. The random power allocation is generated by scaling the optimal $\varepsilon$ in (70) by 0.2. The BER performance with perfect CSIs is provided as a benchmark. The optimal power allocation scheme outperforms the other two power allocation schemes, which proves the importance of optimal training design for channel estimation. Moreover, the BER performance yielded by the optimal power allocation scheme is close to that with perfect CSI. Therefore, the SMMSE estimator and the optimal training design are effective in suppressing performance degradation caused by estimation errors.

The spectral efficiency versus the total training sequence length $T$ for uplink C-RANs is illustrated in Fig. 9. The MRC-ZF receiver is used at the BBU pool. Three SNRs are configured, i.e., 0 dB, 5 dB, and 10 dB. The total number of symbols $L$ is fixed at 32. When $T$ is relatively small, the spectral efficiency curves increase with the increasing $T$. 

![Normalized MSE of $H_{r2}$ versus SNRs.](image1)

![Normalized MSE of $H_{r1}$ versus SNRs.](image2)

![Uplink rate per UE versus $\tau_1$ with different SNRs.](image3)

![Spectral efficiency versus total training sequence length $T$.](image4)
This is reasonable because the estimation accuracy increases as well, which alleviates the uplink system capacity loss in return. Once $T$ is relatively large, the spectral efficiency becomes a decreasing function of $T$. This phenomenon is due to the fact that the improvement of spectral efficiency benefiting from the estimation accuracy enhancement can’t offset the decline brought by the reduction of transmitted data symbols. Therefore, it is vital to design the proportion of training symbol length to data symbol length to maximize the spectral efficiency of the overall uplink C-RAN.

**VII. Conclusion**

In this paper, training design for segment training based individual channel estimation in uplink C-RANs has been considered in order to improve estimation accuracy and data transmission quality. A sequential minimum mean-square-error estimator has been developed through the use of the Kalman filter, in which the channel fading of the radio access links is first estimated, followed by that of the wireless fronthaul links. Based on the SMMSE estimator, a training design for two training segments obtained by minimizing the estimation mean-square-error has been presented. Orthogonal training with the maximum transmit power has been shown to be optimal for the first training segment by using the generalized Karush-Kuhn-Tucker conditions. For the second training segment, training sequence design for both UEs and RRHs has been derived via eigenvalue decomposition. In order to adequately exploit the overall performance of C-RANs, the optimization design of the training sequence length has been considered by maximizing a lower bound on the uplink ergodic capacity. As an estimation performance benchmark, the Cramér-Rao bounds for both these two training segments have been derived. Simulation results have demonstrated that the SMMSE estimator can provide substantial improvements in estimation accuracy. Furthermore, the optimal training design also contributes significant performance improvements. The training design procedure and the related analysis developed in this paper provide useful guidelines for the system design in C-RANs. As an interesting topic for further study, energy efficiency should be adopted to study the influence of channel estimation and power constraints on data detection and the corresponding training design.

**APPENDIX A**

**Proof of Lemma 1**

First, the MSE $\mathcal{M}_0$ in the first fading block is discussed. Assuming that $\mathbf{F} = \mathbf{U}^H$ and using the matrix inversion lemma, i.e., $(\mathbf{I} + \mathbf{AB})^{-1} = \mathbf{I} - \mathbf{A} (\mathbf{I} + \mathbf{BA})^{-1} \mathbf{B}$, the resulting MSE $\mathcal{M}_0$ can be directly computed as

$$\mathcal{M}_0 = \text{tr} \left\{ (\mathbf{I}_M \otimes \mathbf{R}_c^H \mathbf{R}_c) \right\}$$

$$\left( \mathbf{I}_{MN} + \frac{1}{\sigma_n^2} (\mathbf{\Gamma}_2^1 \mathbf{\Phi}_0^* \mathbf{\Phi}_0^T \mathbf{\Gamma}_2^1) \otimes (\mathbf{R}_c^H \mathbf{R}_c)^{-1} \right)^{-1}. \quad (91)$$

Then, the corresponding Lagrangian function with the diagonal constraint can be written as

$$L(\mathbf{\Phi}_0, \mu) = \sum_k^N \lambda_{0,k} \text{tr} \left\{ \left( \mathbf{I}_M + \frac{\lambda_{0,k}}{\sigma_n^2} (\mathbf{\Gamma}_2^1 \mathbf{\Phi}_0^* \mathbf{\Phi}_0^T \mathbf{\Gamma}_2^1) \right)^{-1} \right\}$$

$$+ \mu \left( \text{tr}(\mathbf{\Phi}_0^* \mathbf{\Phi}_0^T) - \tau_2 P_r \right). \quad (92)$$

Since (92) is in quadratic form, only the derivative $\frac{\partial L}{\partial \mathbf{\Phi}_0^T}$ is only considered. According to the KKT conditions, we have

$$\frac{\partial L}{\partial \mathbf{\Phi}_0^T} = \mathbf{\Phi}_0^T \left\{ - \sum_k^N \frac{\lambda_{0,k}}{\sigma_n^2} (\mathbf{\Gamma}_2^1)^{-1} \right. \left. \left( \mathbf{I}_M + \frac{\lambda_{0,k}}{\sigma_n^2} (\mathbf{\Gamma}_2^1 \mathbf{\Phi}_0^* \mathbf{\Phi}_0^T \mathbf{\Gamma}_2^1)^{-1} \right)^{-2T} \mathbf{\Gamma}_2^1 + \mu \mathbf{I} \right\}, \quad (93)$$

$$\mu \left( \text{tr}(\mathbf{\Phi}_0^* \mathbf{\Phi}_0^T) - \tau_2 P_r \right) = 0. \quad (94)$$
Therefore, the optimal $\Phi_0$ can be obtained by setting (93) to zero, which is equivalent to

$$
\sum_k^N \frac{\lambda_k^2}{\sigma_n^2} \left( I_M + \frac{\lambda_k}{\sigma_n^2} (\Gamma_2^{\frac{1}{2}} \Phi_0^T \Phi_0^{\frac{1}{2}}) \right)^2 = \mu \Gamma_2^{-1}. 
$$

(95)

It is easy to observe from the equation above that the optimal training signal must satisfy $\Phi_0^\ast \Phi_0^T = \tau_2 p_r I_M$ for any $\mu > 0$.

When the optimal $\Phi_0$ is applied, $R_{1|0}$ can be calculated according to the Kalman filter

$$
R_{1|0} = \eta^2 R_{0|0} + (1-\eta^2) \tilde{R} 
$$

$$
= \tilde{U} \text{diag} \left( \lambda_0, 1 - \frac{\eta^2 \tau_2 p_r \beta_1 \lambda_0^2}{\sigma_n^2 + \tau_2 p_r \beta_1 \lambda_0}, \ldots, \right) \tilde{U}^T, 
$$

(96)

where $\tilde{R} = I \otimes R$, $\tilde{U} = I \otimes U$ and $\tilde{\Lambda} = I \otimes \Lambda$. As $R_{1|0}$ is similar to $\tilde{R}$ as well as all the following $R_{ij|i}$, $i > 1$, via recursive derivation, it follows that the optimal training signal $\Phi_i$ for $i \geq 1$ also satisfies $\Phi_i^\ast \Phi_i^T = \tau_2 p_r I_M$.

APPENDIX B

PROOF OF LEMMA 3

Note that the minimization of $R_{m|1}$ is equivalent to maximizing the second term on the RHS of (21), which can be simplified as

$$
\mathcal{M} = \text{tr} \left\{ \left[ \sigma_n^2 I_{N_r} + \sigma_n^2 I_1 \otimes \Lambda_2 + (\Lambda_1^2 \otimes \Lambda_2^2) (Q_1 \otimes Q_2) \right] \right\}. 
$$

It can be observed that the $\mathcal{M}$ is invariant to $U_1$ and $U_2$ and is dependent on $\Lambda_1$, $\Lambda_2$, $Q_1$ and $Q_2$. According to the lemmas in [30], the RRH power constraint can be written as

$$
\text{tr} \left( D(Q_1^H \otimes Q_2^H) (A_{2}^{\frac{1}{2}} \otimes \Lambda_2^2) (Q_1 \otimes Q_2) \right) 
$$

$$
+ \tau_1 \text{tr} \left( A_{2}^{-2} U_g^H U_2 A_2 U_2^H U_g \right) \geq \text{tr} \left( D(Q_1^H \otimes Q_2^H) \right). 
$$

$$
(A_1 \otimes \Lambda_2^2 A_2^{-2} \Lambda_2^2 \otimes A_2) (Q_1 \otimes Q_2) \right) + \tau_1 \text{tr} \left( A_{2}^{-2} A_2 \right), \quad (98)
$$

APPENDIX C

PROOF OF PROPOSITION 4

$G_r^H G_r$ has an $M \times M$ central complex Wishart distribution with $N$ degrees of freedom and covariance matrix $\Pi = \text{diag} \{ \beta_1, \beta_2, \ldots, \beta_M \}$, i.e., $G_r^H G_r \sim \mathcal{W}_M(N, \Pi)$. Using the identity in [25], we have

$$
\mathcal{E} \left\{ (G_r^H G_r)^{-1} \right\} = \frac{1}{N-M} \Pi^{-1}, \quad N \geq M + 1. 
$$

(99)

Similarly, $\tilde{G}_r^H \tilde{G}_r$ also has a complex central Wishart distribution, and we can also obtain

$$
\mathcal{E} \left\{ (\tilde{G}_r^H \tilde{G}_r)^{-1} \right\} = \frac{1}{N-M} \Pi^{-1}, \quad N \geq M + 1, 
$$

(100)

where $\tilde{G}_r = \text{diag} \left\{ \tau_2 p_r \beta_1^2, \tau_2 p_r \beta_1^2, \ldots, \tau_2 p_r \beta_M^2 \right\}$.

APPENDIX D

PROOF OF LEMMA 5

For the sake of tractability, we assume that all the large-scale fading coefficients of RRHs are equal to the smallest one, i.e., we set $\beta_m = \beta, m = 1, \ldots, M$. Since $g_i(i \neq k) \sim \mathcal{CN}(0, \kappa_i)$ is independent of $g_k$, the expectation of (51) can be taken over all the $g_i(i \neq k)$ first, and the uplink rate of the $k$-th UE is lower bounded by

$$
\hat{R}_k = \log_2 \left( 1 + \mathcal{E} \left\{ a_1 \| g_k \|^4 + a_2 \| g_k \|^2 + a_3 \right\} \frac{\eta_p}{\eta_g} \| g_k \|^2 \right)^{-1}. 
$$

(101)

The covariance matrix of $g_k$ is $\kappa_i I_M$, and $\frac{\eta_p}{\eta_g} \| g_k \|^2$ has a chi-square distribution with $2M$ degrees of freedom. Denote $\| g_k \|^2$ as $x$, and the corresponding probability density function of $x$ is thus given by

$$
p(x) = \frac{1}{(M-1) \kappa_i} x^{M-1} \exp \left( -\frac{x}{\kappa_i} \right), \quad x \geq 0. 
$$

(102)

The expectation on the RHS of (101) is given by

$$
\mathcal{E} \left\{ a_1 \| g_k \|^4 + a_2 \| g_k \|^2 + a_3 \right\} \frac{\eta_p}{\eta_g} \| g_k \|^2 
$$

$$
a_1 \frac{\eta_p}{\eta_g} + a_2 \frac{\eta_p}{\eta_g} (M-1) \kappa_k + a_3 \frac{\eta_p}{\eta_g} (M-1)(M-2) \kappa_k^2, \quad M \geq 3. 
$$

(103)

Substituting (103) into (101), we can obtain the result (54).

REFERENCES


Qiang Hu received the B.S. degree in applied physics from the Beijing University of Posts & Communications (BUPT), China, in 2013. He is currently pursuing the Master degree at BUPT. His research interests include cooperative communications, such as cloud radio access networks (C-RANs), and statistical signal processing in largescale networks.

Mugen Peng (M’05–SM’11) received the B.E. degree in electronics engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2000, and the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. Afterward, he joined BUPT, where he has been a Full Professor with the School of Information and Communication Engineering since 2012. In 2014, he was an Academic Visiting Fellow with Princeton University, Princeton, NJ, USA. He leads a Research Group focusing on wireless transmission and networking technologies with the Key Laboratory of Universal Wireless Communications (Ministry of Education), BUPT. His main research areas include wireless communication theory, radio signal processing, and convex optimizations, with a particular interests in cooperative communication, self-organization networking, heterogeneous networking, cloud communication, and internet of things. He has authored/coauthored over 60 refereed IEEE journal papers and over 200 conference proceeding papers.

Dr. Peng was a recipient of the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the best paper award in IEEE WCNC 2015, WASA 2015, GameNets 2014, IEEE CIT 2014, ICTA 2011, IC–BNMT 2010, and IET CCWMC 2009. He received the First Grade Award of the Technological Invention Award in the Ministry of Education of China for the hierarchical cooperative communication theory and technologies, and the First Grade Award of Technological Invention Award from the China Institute of Communications for the contributions to the self-organizing networking technology in heterogeneous networks. He is on the Editorial/Associate Editorial Board of the *IEEE Communications Magazine*, *IEEE Access*, *IET Communications*, *International Journal of Antennas and Propagation (IJAP)*, and *China Communications*. He has been the guest leading editor for the special issues in the *IEEE Wireless Communications, IEEE Access*, and *IET Communications*. He is the Fellow Of the Institution of Engineering and Technology (IET).
Zhendong Mao received the B.S. degree in applied physics from the Beijing University of Posts & Communications (BUPT), China, in 2013. He is currently pursuing the Ph.D. degree at BUPT. His research interests focus on the channel estimation and training design in cloud radio access networks (C-RANs) and fog radio access networks.

Xinqian Xie received the B.E. degree and the Ph.D. degree in telecommunication engineering from the Beijing University of Posts and Communications (BUPT), China, in 2010 and 2015, respectively. He is currently working in the Huawei Company. His research interests include cooperative communications, estimation and detection theory.

H. Vincent Poor (S’72, M’77, SM’82, F’87) received the Ph.D. degree in electrical engineering and computer science from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Dean of Engineering and Applied Science, and the Michael Henry Strater University Professor of Electrical Engineering. He has also held visiting appointments at several other institutions, most recently at Imperial College and Stanford. His research interests are in the areas of information theory, stochastic analysis and statistical signal processing, and their applications in wireless networks and related fields. Among his publications in these areas is the recent book Mechanisms and Games for Dynamic Spectrum Allocation (Cambridge University Press, 2014).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of Academia Europaea and the Royal Society. He is also a fellow of the American Academy of Arts and Sciences, the Royal Academy of Engineering (U. K.), and the Royal Society of Edinburgh. In 1990, he served as President of the IEEE Information Theory Society, in 2004-07 as the Editor-in-Chief of these TRANSACTIONS, and in 2009 as General Co-chair of the IEEE International Symposium on Information Theory, held in Seoul, South Korea. He received a Guggenheim Fellowship in 2002 and the IEEE Education Medal in 2005. Recent recognition of his work includes the 2014 URSI Booker Gold Medal, the 2015 EURASIP Athanasios Papoulis Award, the 2016 John Fritz Medal, and honorary doctorates from Aalborg University, Aalto University, HKUST, and the University of Edinburgh.