Fronthaul-Constrained Uplink Cloud Radio-Access Networks: Capacity Analysis and Algorithm Design

by

Yuhan Zhou

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Electrical and Computer Engineering
University of Toronto

© Copyright 2016 by Yuhan Zhou

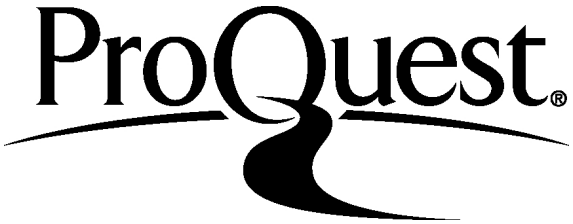
www.manaraa.com

ProQuest Number: 10141299

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest 10141299

Published by ProQuest LLC (2016). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

www.manaraa.com

Abstract

Fronthaul-Constrained Uplink Cloud Radio-Access Networks:
Capacity Analysis and Algorithm Design

Yuhan Zhou
Doctor of Philosophy
Graduate Department of Electrical and Computer Engineering
University of Toronto
2016

This thesis considers the uplink of a cloud radio access network (C-RAN), in which base-stations (BSs) are connected to a cloud-computing based central processor (CP) via noiseless fronthaul links with finite capacities. The compress-and-forward strategy is employed, where the BSs compress the received signals and send the quantized bits to the CP. Then, the CP performs either joint decoding of both the quantization and message codewords simultaneously, or generalized successive decoding of quantization and message codewords in an arbitrary order. Under this setup, this thesis establishes several information theoretic results and proposes a number of practical algorithm designs.

From a theoretical perspective, this thesis first proves that under joint decoding and Gaussian input, Gaussian quantization maximizes the achievable rate region. Second, it is shown that generalized successive decoding achieves the identical rate region as joint decoding under a sum fronthaul capacity constraint. Third, a particular successive decoding scheme, in which quantization codewords are decoded first followed by message codewords, referred to as the virtual multiple-access channel (VMAC) scheme, achieves the same maximum sum rate as joint decoding under individual fronthaul constraints. Furthermore, it is shown that under a sum fronthaul constraint, Wyner-Ziv coding, quantized at the background noise level, can achieve the sum-capacity to within a constant gap. A similar constant-gap result is shown for single-user compression under a diagonally dominant channel condition.

From an optimization perspective, this thesis investigates the optimization of beamforming design and fronthaul compression for the VMAC schemes. First, under a sum fronthaul constraint, this thesis proposes a novel alternating convex optimization algorithm to maximize the weighted sum-rate for single-antenna uplink C-RAN. It is shown that setting the quantization noise levels to be proportional to the background noise levels is near optimal when the signal-to-quantization-noise-ratio is high. Second, under individual fronthaul constraints, this thesis develops a weighted minimum mean-square-error successive convex approximation algorithm to jointly optimize beamforming and fronthaul compression for multi-antenna uplink C-RAN. The performances of the proposed algorithms are verified under practical multicell and heterogeneous networks through numerical evaluation.

www.manaraa.com

Acknowledgements

This dissertation would not have been possible without the support I received from my parents, teachers and friends. First and foremost, I am deeply indebted to my advisor Prof. Wei Yu for his constant guidance and support at every stage of my Ph.D. studies, without which this dissertation would be impossible. A superb teacher and mentor, he taught me the knowledge on wireless communications, information theory and convex optimization. He taught me how to find a valuable research problem in wireless communications and formulate it in the cleanest way. His brilliant insights, strong passion, and rigorous academic attitude will be an inspiring role model for my future career. In addition, I appreciate his endless patience and great effort in editing and improving my academic papers.

I thank Prof. Frank Kschischang, Prof. Stark Draper, Prof. Ravi Adve, and Prof. Ashish Khisti for serving as my thesis defense committee members. Their valuable feedbacks have greatly improved the quality of the thesis. I thank Prof. Vincent Lau for serving as the external appraiser for my PhD defense.

I also extend my thanks to those professors, who are the instructors of the classes that I have attended. In particular, I thank Prof. Jeffrey Rosenthal for offering a fantastic course on stochastic processes, which makes probability theory become my favourite subject at University of Toronto.

I am particularly indebted to Prof. Jun Chen, Dr. Yinfei Xu, and Dr. Dimitris Toumpakaris for fruitful collaborations. Joint work with Prof. Jun Chen and Dr. Yinfei Xu led to the main results in Chapter 2 of the thesis. The main content of Chapter 3 in the thesis was greatly influenced by the discussion with Dr. Dimitris Toumpakaris. I am also grateful to Dr. Chirag Patel and Dr. Farhad Meshkati for being my mentors during my internship at Qualcomm Research Center in San Diego, where I spent a wonderful summer in 2014.

I have been fortunate to have many friends and colleagues who made my five years at Toronto an unforgettable experience: Lei Zhou, Yicheng Lin, Gokul Sridharan, Soroush Tabatabaei, Qiang Xiao, Louis Tan, Lei Zhang among them. I thank my dearest friends Binbin Dai, Dan Fang, Wei Bao, Zhi Zeng, Siyu Liu, Wanyao Zhao, who are always there for me during both joyful and gloomy times. I thank Pratik Patil for the interesting discussions, during which I gained new knowledge. I would also like to thank my Chinese friends Chu Pang, Caiyi Zhu, Huiyuan Xiong, Weiwei Li, Chunpo Pan, Wei Wang, Alice Gao, Cecilia Liu for the joyful cards-game nights that we spent together.

Finally, it is my greatest honor to thank my family: my mother and my father. No words could possibly express my deepest gratitude for their endless love, self-sacrifice and unwavering support. To them I dedicate this dissertation.

I am grateful to the Graduate Student Fellowship program and Queen Elizabeth II Graduate Scholarship in Science and Technology (QEII-GSST) at the University of Toronto, the Natural Science and Engineering Research Council (NSERC) of Canada, and Huawei Technologies, Canada for providing financial assistance during my graduate studies.

www.manaraa.com

To my beloved parents whose love and support
made this thesis possible

www.manaraa.com

Contents

www.manaraa.com

www.manaraa.com

www.manaraa.com

List of Tables

www.manaraa.com

List of Figures

www.manaraa.com

www.manaraa.com

Chapter 1

Introduction

The data demand in wireless communication driven by smartphones, tablets, and video streaming is increasing dramatically. To satisfy such growing user demands, modern cellular communication systems are moving toward densely deployed heterogenous networks consisting of base-stations (BSs) covering progressively smaller areas. Advanced techniques such as multiuser multiple-input-multiple-output (MIMO), in which numerous antennas simultaneously serve a large number of users in the same time-frequency resource, have been proposed in order to increase network capacity by creating more degrees of freedom for data transmission. As a consequence, the growing inter-cell interference levels and high deployment cost become the dominant limiting factors for the current cellular systems.

Cloud radio access network (C-RAN) refers to the virtualization of BS functionalities by means of cloud computing, which has the potential to address the aforementioned problems. In a C-RAN architecture, the baseband and higher layer operations of the BSs are migrated to a cloud-computing based centralized processor (CP). By allowing coordination and joint signal processing across multiple cells, C-RAN provides a platform for implementing network MIMO, also known as coordinated multi-point (CoMP), which can achieve significantly higher data rates than conventional cellular networks [1]. Due to that fact that fewer baseband units are needed in C-RAN compared to the traditional distributed cellular architecture, C-RAN also has the potential to decrease the cost of network operation, because of the reduced power and energy consumption [2].

1.1 C-RAN Architecture

Conventional wireless communication systems are designed with a cellular architecture, in which a geographical area to be supplied with radio service is divided into non-overlapping *cells*. The BS deployed in each cell provides coverage to users in that cell; mobile devices communicate with their assigned BS within each cell. In this architecture, the BS is the direct interface between the mobile users and the backbone wireline network, which implements all the functions of baseband and radio processing including modulation/demodulation, encoding/decoding of user information, frequency filtering and power amplification, etc.

In the C-RAN architecture, all the BSs are connected to a reconfigurable, general-purpose CP, as shown in Fig. 1.1. In such a system, the BSs degenerate into remote antennas, implementing only radio functionalities, including transmission/reception, filtering, amplification, down- and up-conversion

www.manaraa.com

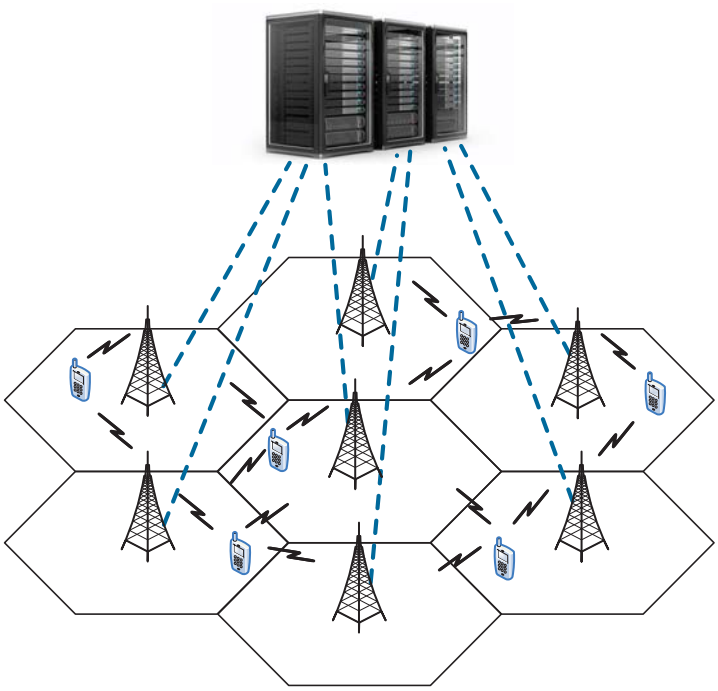


Figure 1.1: Illustration of the uplink of a cloud radio access network.

and possibly analog-to-digital conversion (ADC) and digital-to-analog conversion (DAC). The baseband operations at the BSs are migrated to the software-defined CP, which operates as a virtual BS to encode/decode user information and optimize the radio resource allocation [3]. In this scenario, subsets of adjacent cells can form cooperating clusters. The radio resources of the BSs within one cluster can be fully shared, thus forming a MIMO network system from the CPs perspective. The BS-CP bidirectional links that carry the information are referred to as *fronthaul* links, in contrast to the backhaul links connecting the CP to the backbone network. Fronthaul can be realized with different technologies, such as optical fiber communication [4], microwave communication [5], or even millimetre wave communication [6].

The novel architecture of C-RAN brings a number of benefits to the current wireless communication system: the centralized baseband processing reduces the power consumption and cost of BS operation, and provides increased flexibility in network upgrades and adaptability to non-uniform traffic. Advanced techniques such as network MIMO and CoMP, can be efficiently supported by C-RAN for mitigating inter-cell interference. In contrast to the traditional cellular system, the C-RAN architecture with co-located processing units eases network maintenance and upgrades.

The practical implementation of the C-RAN architecture restricts all the fronthaul links to have finite capacities. For example, the typical microwave fronthaul links in use today have the average capacity less than 100Mbps [7]. This finite-capacity constraint on the fronthaul makes both the theoretical analysis and the practical algorithm design for the C-RAN architecture challenging. To address this problem, this thesis studies the compress-and-forward scheme for uplink C-RAN with capacity-limited fronthaul. Through capacity analysis and algorithm design, this thesis aims to maximize the advantages of the C-RAN architecture by jointly optimizing the input signals at the users and the quantization at the BSs, and consequently, to enhance the performance of cellular systems to meet the requirement of the upcoming fifth generation (5G) wireless systems.

www.manaraa.com

1.2 Literature Survey

The uplink C-RAN model is also known in the literature as the multi-cell joint processing model, where the CP replaces the BSs in performing encoding/decoding functionalities. The high-capacity fronthaul links between the BSs and the central processor are used to exchange both the user data and the channel state information (CSI). Under the C-RAN architecture, joint transmission in the downlink and joint reception in the uplink can be performed to effectively mitigate the inter-cell interference.

From an information theoretic point of view, the uplink C-RAN model can be thought of as a two-hop relay network between the users and the central processor, with the BSs acting as relays. The study of limited fronthaul C-RAN model originates from the work on BS cooperation under infinite fronthaul [8] [9], where the network capacity has been shown to grow linearly with the number of BS antennas. This result has been extended to the case where BSs or transmitters are equipped with multiple antennas [10], but based on the assumption that the fading coefficients of the MIMO subchannels are completely uncorrelated. The effect of MIMO subchannel correlation on the capacity of the C-RAN system has been studied in [11].

The main coding strategy for the uplink C-RAN model is compress-and-forward, in which the BSs quantize their received signals, then forward the quantization codeword to the CP. In the decoding procedure, the CP may either jointly decode the quantization codewords and the user messages, or decode them successively, giving rise to different complexity-performance tradeoffs. The fundamental achievability scheme for the uplink C-RAN is first proposed for the uplink model under the individual fronthaul capacity constraint in [12–14]. In [12–14], the compress-and-forward relaying scheme and its achievable rates are derived for the scenario where multiple users communicate with a remote destination via multiple relays. The coding schemes of [12–14] assume joint decoding. Under the joint decoding framework, the uplink C-RAN model can also be considered as a special case of the multi-message multicast relay network whose entire capacity region can be achieved to within a constant gap, using the recently proposed quantize-map-and-forward or noisy network coding schemes [15–17].

To fully explore the benefits brought by the C-RAN architecture, it is important to efficiently utilize the finite-capacity fronthaul links between the BSs and CP. Substantial research works have made progress towards this direction from different perspectives such as fronthaul compression, CSI acquisition and processing, signal synchronization, and delay-aware resource allocation (See, e.g. [18] and references therein). For the optimization of fronthaul compression in uplink C-RAN, [19, 20] consider the optimal distributed compression strategies at the BSs with an emphasis on sum-rate maximization and robustness, respectively. The optimal channel training time to obtain the CSI at the BSs in uplink C-RAN with limited fronthaul is studied in [21]. The case of uplink multicell processing with imperfect CSI has been investigated in [22], where various relaying schemes, including decode-and-forward and compress-and-forward techniques, are compared under the assumption that there might be errors in channel estimation. To efficiently deliver CSI information to the central receiver, [23] proposes a compress-forward-estimate approach which jointly designs the fronthaul and CSI compression. Furthermore, if the CSI is only available at the CP side, the performance of compress-and-forward is evaluated for a two-user C-RAN model under limited individual fronthaul in [24]. The complexity of the CSI processing in C-RAN can be significantly reduced by utilizing the sparsity of the channel matrix [25]. The signal sparsity in C-RAN can also used to improve the performance of the fronthaul compression and user detection [26]. Additionally, by introducing the delay-optimal fronthaul allocation, the latency issue in the C-RAN design can be efficiently controlled [27]. The fronthaul compression can also be

www.manaraa.com

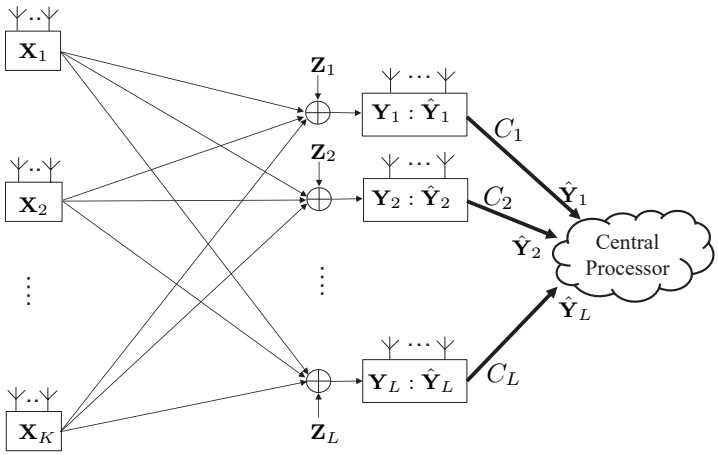


Figure 1.2: Illustration of a uplink cloud radio-access network with capacity-limited fronthaul.

designed for enhancing the performance of the synchronization in C-RAN [28].

Finally, we mention a completely different class of coding strategies called "compute-and-forward", in which the relaying BSs forward a function of the transmitted signals from the mobile subscribers to the CP. Under the assumption that the CSI is only available at the CP side, [29] demonstrated that a lattice-code-based relaying scheme can outperform the conventional decode-and-forward and compress-and-forward schemes in a certain signal-to-interference-and-noise-ratio (SINR) regime, for a symmetric Wyner model under a specific fronthaul constraint where each fronthaul link has the same capacity. This scheme is further applied to the C-RAN model with equal-capacity fronthaul links in [30], where it is shown that as compared with the compress-and-forward scheme, it achieves competitive performance but with significantly lower complexity. However, the construction of lattice codes for compute-and-forward is a nontrivial problem. Additionally, compute-and-forward can be very sensitive to channel estimation errors [31]. A sophisticated code design could potentially address this issue at the cost of high decoding complexity [32].

1.3 Overview of Thesis

This thesis focuses on the fundamental limits and system-level optimization of uplink C-RAN under practical finite-capacity fronthaul constraints. The channel model of the uplink C-RAN architecture is shown in Fig. 1.2, in which multiple distributed transmitters send information to a centralized receiver through a multi-access relay network. The main objective of this thesis is to optimize the input signaling at the transmitter, the quantization design at the relaying nodes, and the decoding strategy at the CP for maximizing the network utility.

As a first step toward practical implementation of C-RAN, we need to characterize the maximum achievable rate region of uplink C-RAN. Under the compress-and-forward scheme, the best known achievable region for uplink C-RAN is given by joint decoding of quantization codewords and user messages [13]. It is worth noting that this achievability scheme can be thought of as a particular instance of the noisy network coding scheme [15–17, 33] applied to a multi-message multicast relay network. Noisy network coding achieves the capacity region of the Gaussian multicast relay networks to within a constant gap. Consequently, the compress-and-forward scheme with joint decoding is also shown to achieve the ca-

www.manaraa.com

pacity region of the uplink C-RAN model to within a constant gap [34]. However, joint decoding is challenging to implement: the computational complexity of joint decoding scales exponentially with the total number of nodes in the network. Moreover, even a mere evaluation of the achievable rate could be computationally prohibitive. Under joint decoding, the achievable rate region for the uplink C-RAN model shown in Fig. 1.2 involves $2^K - 1$ constraints, each of which is a minimization over 2^L terms, where K and L are the number of users and relay nodes in the network, respectively. To overcome the above difficulty yet still take full advantage of the centralized processor is quite challenging.

When realistic design is considered for implementing the uplink C-RAN, a more practical successive decoding scheme [12, Theorem 1] could be adopted, in which the BSs perform Wyner-Ziv coding to compress the received signals and send the quantization bits to the CP; the CP decodes the quantization codes first, and then the user messages sequentially. The optimal quantization strategy and optimal decoding strategy for compress-and-forward in uplink C-RAN is the focus of Chapter 2 of this thesis. In Chapter 2, the compress-and-forward scheme is investigated, in which the BSs perform Wyner-Ziv coding to compress and send the received signals to the CP; the CP performs either joint decoding of both the quantization codewords and the user messages simultaneously, or successive decoding of the quantization and user message codewords according to a specific order that maximizes the network utility. Under this setup, this chapter makes progress toward the optimization of the fronthaul compression scheme by proving the following two results. First, it characterizes the rate region of generalized successive decoding which allows arbitrary decoding orders of the quantization and user message codewords, and shows that under a sum fronthaul capacity constraint, generalized successive decoding achieves the same rate region as joint decoding. It is also shown that the practical successive decoding which decodes the quantization codes first, and then the user messages, achieves the same performance as joint decoding for maximizing the sum rate of uplink C-RAN. Second, it is shown that if the input distributions are assumed to be Gaussian, then under joint decoding, the optimal quantization scheme for maximizing the achievable rate region is Gaussian.

For the implementation of the compress-and-forward strategy in C-RAN, it is important to choose appropriate quantization noise levels, such that the network utility can be maximized. Chapter 3 studies such optimization problems for uplink single-input-single-output (SISO) C-RAN. The compress-and-forward scheme with successive decoding is employed, in which the single-antenna BSs quantize the received signals and send the compressed bits to the CP using either distributed Wyner-Ziv coding or single-user compression. The CP decodes the quantization codewords first, and then decodes the user messages as if the remote users and the cloud center form a virtual multiple-access channel (VMAC). Chapter 3 formulates the problem of optimizing the quantization noise levels for the weighted sum rate maximization under a sum fronthaul capacity constraint. A novel alternating convex optimization approach is proposed to find a local optimum solution to the optimization problem. More importantly, it is established that setting the quantization noise levels to be proportional to the background noise levels is near optimal for sum-rate maximization, when the signal-to-quantization-noise ratio (SQNR) is high. In addition, with Wyner-Ziv coding, the approximately optimal quantization noise level is shown to achieve the sum-capacity of the uplink C-RAN model to within a constant gap. With single-user compression, a similar constant-gap result is obtained under a diagonal dominant channel condition. These results lead to an efficient algorithm for allocating the fronthaul capacities in C-RAN. The performance of the proposed scheme is evaluated for practical multicell and heterogeneous networks. It is shown that multicell processing with optimized quantization noise levels across the BSs can significantly improve

www.manaraa.com

Table 1.1: Summary of the Proposed Algorithms

	ACO scheme	Approx. opt. quantization noise level	WMMSE-SCA scheme	Separate design
Uplink C-RAN system	SISO	SISO	MIMO	MIMO
Fronthaul compression	Optimized	Approx. opt.	Jointly optimized	Approx. opt.
Transmit signal	Fixed	Fixed	Jointly optimized	Approx. opt.
Performance	High	Low	High	Low
Complexity	High	Low	High	Low

the performance of wireless cellular networks.

As an extension of Chapter 3 to the multi-antenna case, Chapter 4 investigates the fronthaul compression and transmit beamforming design for uplink MIMO C-RAN. A practical compress-and-forward scheme is employed, in which the CP performs successive decoding with either successive interference cancellation (SIC) receiver or linear minimum-mean-square-error (MMSE) receiver. Since conventional transmit beamforming strategies are designed not to be fronthaul-aware, they are not necessarily suitable for the multi-cell processing feature of the C-RAN architecture. Chapter 4 proposes a joint design of the transmit beamformers at the users and the quantization noise covariance matrices at the BSs for maximizing the network utility. A novel weighted minimum-mean-square-error successive convex approximation (WMMSE-SCA) algorithm is proposed for maximizing the weighted sum rate under the user transmit power and fronthaul capacity constraints with single-user compression first. Assuming a heuristic decompression ordering strategy, the proposed algorithm is then adapted for optimizing the transmit beamforming and fronthaul compression under distributed Wyner-Ziv coding. In addition, Chapter 4 also proposes a low-complexity separate design consisting of optimizing transmit beamformers for the Gaussian vector multiple-access channel along with per-antenna scalar quantizers with uniform quantization noise levels across the antennas at each BS. Numerical results show that the majority of the performance gain stems from the implementation of SIC at the CP. Furthermore, the low complexity separate design performs very close to the optimized joint design in the SQNR regime of practical interest.

Chapter 3 and Chapter 4 in the thesis propose four algorithms and their variants under the different scenarios. The proposed algorithms are the ACO scheme and the approximately optimal quantization noise level for optimizing the fronthaul compression with fixed transmit signal and under a sum fronthaul constraint, the WMMSE-SCA scheme and the separate design for jointly optimizing the beamforming and compression under individual fronthaul constraints. The differences between the above four algorithms are illustrated in Table 1.1. The ACO scheme vs. the approximately optimal quantizer (or the WMMSE-SCA scheme vs. the separate design) demonstrate the tradeoff between performance and complexity in uplink C-RAN design. In this sense, the above algorithms offer a great degree of flexibility in implementing uplink C-RAN under various design criterion.

The publications related to this thesis are as follows. Chapter 2 is joint work with Jun Chen (McMaster University, Canada) and Yinfei Xu (Southeast University, China), which is presented in part in [35]; Chapter 3 is presented in part in [36] and [37], respectively, and a complete version of Chapter 3 can be found in [38]; Chapter 4 is published in part in [39], and a complete version of Chapter 4 forms [40].

www.manaraa.com

1.4 Notations

Notation: Boldface letters denote vectors or matrices, where context should make the distinction clear. Superscripts $(\cdot)^\dagger$ and $(\cdot)^{-1}$ denote Hermitian transpose and matrix inverse operators; $\mathbb{E}[\cdot]$ and $\mathrm{Tr}(\cdot)$ denote expectation and matrix trace operators; $\mathrm{cov}(\cdot)$ denotes the covariance operation; $\mathrm{co}(\cdot)$ denotes the convex closure operation. We use $\mathbf{X}_i^j = (\mathbf{X}_i, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_j)$ to denote a matrix with $(j - i + 1)$ columns for $1 \leq i \leq j$. For a vector/matrix $\mathbf{X}$, $\mathbf{X}_{\mathcal{S}}$ denotes a vector/matrix with elements whose indices are elements of $\mathcal{S}$. Given matrices $\{\mathbf{X}_1, \ldots, \mathbf{X}_L\}$, $\mathrm{diag}\left(\{\mathbf{X}_\ell\}_{\ell=1}^L\right)$ denotes the block diagonal matrix formed with $\mathbf{X}_\ell$ on the diagonal. Denote by $\mathbf{J}(\mathbf{X})$ the Fisher information matrix of the random vector $\mathbf{X}$. $\mathbb{R}_+$ is used to denote non-negative real numbers. We let $\mathcal{K} = \{1, \cdots, K\}$ and $\mathcal{L} = \{1, \cdots, L\}$.

www.manaraa.com

Chapter 2

On the Optimal Compression and Decoding for Uplink C-RAN

2.1 Introduction

This chapter studies the information theoretic limit of uplink C-RAN under finite-capacity fronthaul constraints. The uplink C-RAN model shown in Fig. 1.2 is re-plotted here in Fig. 2.1 for the convenience of readers, which consists of multiple remote users sending independent messages to the CP through multiple BSs serving as relay nodes. Both the user terminals and the BSs are equipped with multiple antennas. The BSs and the CP are connected via noiseless fronthaul links with finite capacity. This channel model can be thought of as a two-hop relay network, with an interference channel between the users and the BSs, followed by a noiseless multiple-access channel between the BSs and the CP. This chapter assumes that the compress-and-forward relaying strategy is employed, in which the relaying BSs compress the received signals and forward the quantization bits to the CP through fronthaul links, and all the user messages are eventually decoded at the CP.

A key question in the design of compress-and-forward strategy in uplink C-RAN is the optimal input coding strategy at the user terminals, the optimal relaying strategy at the BSs, and the optimal decoding strategy at the CP. Toward this end, this chapter restricts attention to the strategy of compressing the received signals at the BSs, then either *joint decoding* of the quantization and message codewords simultaneously, or *generalized successive decoding* of the quantization and message codewords in some arbitrary order at the CP. Under this assumption, this chapter makes the following contributions toward revealing the structure of the optimal compress-and-forward strategy.

First, motivated by the fact that successive decoding is much easier to implement than joint decoding, we seek to understand whether successive decoding at the CP can perform as well as joint decoding. Toward this end, this chapter shows that the two schemes indeed achieve the same rate region for an uplink C-RAN model under a sum fronthaul constraint. Further, although not necessarily so for the general rate region, if one focuses on maximizing the sum rate, the strategy of successively decoding the quantization codewords first, then the user messages, achieves the optimal sum rate.

Second, we seek to understand the optimal input distribution and quantization schemes in uplink C-RAN. Although it is well known that joint Gaussian strategies are not necessarily optimal, this chapter shows that if we fix the input distribution to be Gaussian, then the optimal quantization scheme is

www.manaraa.com

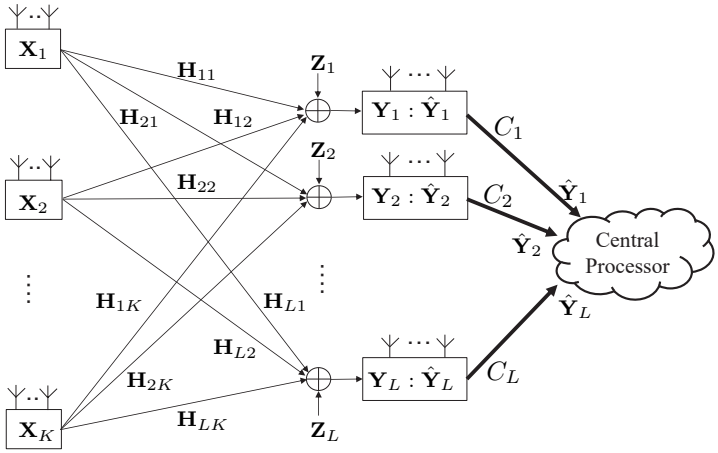


Figure 2.1: The uplink C-RAN model under finite-capacity fronthaul constraints

Gaussian under joint decoding, and vice versa. Moreover, joint Gaussian signaling can be shown to achieve the capacity region of the Gaussian multiple-input multiple-output (MIMO) uplink C-RAN model to within a constant gap. Finally, this chapter makes progress on the computational front by showing that under the joint Gaussian assumption, the optimization of the quantization covariance matrices for maximizing the sum rate can be formulated as a convex optimization problem. These results suggest that joint Gaussian input signaling and Gaussian quantization is a reasonable strategy for the practical implementation of uplink C-RAN.

2.1.1 Related Work

The achievable rate region of compress-and-forward with joint decoding for the uplink C-RAN model was first characterized in [13] for a single-transmitter model then in [14] for the multi-transmitter case. However, the number of rate constraints in the joint decoding rate region grows exponentially with the size of the network [13, Proposition IV.1], which makes the evaluation of the achievable rate computationally prohibitive. The achievable rate region of the compress-and-forward strategy with practical successive decoding, in which the quantization codewords are decoded first, then the user messages are decoded based on the recovered quantization codewords, has also been studied for the uplink C-RAN model [12, Theorem 1]. One of the objectives of this chapter is to illustrate the relationship between joint decoding and successive decoding. In the existing literature, the equivalence between these two decoding schemes is first demonstrated for single-source, single-destination, and single-relay networks [41, Appendix 16C], then shown for single-source, single-destination, and multiple-relay networks [42], under either block-by-block forward decoding or block-by-block backward decoding. This chapter further demonstrates that in the case of uplink C-RAN, which is a multiple-source, single-destination, multiple-relay network, the optimality of successive decoding still holds under suitable conditions.

In general, it is challenging to find the optimal joint input and quantization noise distributions that maximize the achievable rate of the compress-and-forward scheme for uplink C-RAN. Gaussian signaling is not necessarily optimal—in particular, in a simple example of uplink C-RAN with one user and two BSs shown in [12], binary input is shown to outperform Gaussian input. However, Gaussian input and Gaussian quantization can be shown to be approximately optimal. In fact, the uplink C-RAN model is an example of a general Gaussian relay network with multiple sources and a single destination

www.manaraa.com

for which a generalization of compress-and-forward with joint decoding (referred to as noisy network coding scheme [15–17, 33]) and with Gaussian input and Gaussian quantization can be shown to achieve to within a constant gap to the information theoretical capacity of the overall network. Instead of using noisy network coding, our previous work [38] shows that successive decoding can achieve the sum capacity of uplink C-RAN to within constant gap, if the fronthaul links are subjected to a sum capacity constraint. In this work, we further demonstrate that the compress-and-forward scheme with joint decoding can achieve to within a constant gap to the entire capacity region of the uplink C-RAN model with individual fronthaul constraints; same is true for successive decoding under suitable condition.

An important theoretical result obtained in this chapter is that if the input distributions of the uplink C-RAN model are fixed to be Gaussian, then Gaussian quantizer is in fact optimal under joint decoding. Finding the optimal quantization for the C-RAN model is related to the mutual information constraint problem [43], for which entropy power inequality is used to show that Gaussian quantization is optimal for a three-node relay network with Gaussian input. However, it is challenging to extend this approach to the uplink C-RAN model, which has multiple sources. This chapter provides a novel proof of the optimality of Gaussian quantization based on the de Bruijn identity and the Fisher information inequality. A key insight here is a connection between the C-RAN model and the CEO problem in source coding [44], where a source is described to a central unit by remote agents with noisy observations. The solution to the CEO problem is known for the scalar Gaussian case [45], while significant recent progress has been made in the vector case, e.g., [46]. In this chapter, we use techniques for establishing the outer bound for the Gaussian vector CEO problem [47] to prove the optimality of Gaussian quantization. We also remark the connection between this quantization optimization problem and the information bottleneck method [48], for which joint Gaussian distribution is shown to be Pareto optimal. The technique used in this chapter is a significantly simpler alternative to the enhancement technique given in [49, 50].

This chapter also makes progress in observing that the optimization of Gaussian quantization noise covariance matrices for maximizing the (weighted) sum rate of uplink C-RAN can be reformulated as a convex optimization problem. The quantization noise covariance optimization problem has been considered in the literature, but only locally convergent algorithms are known previously [19, 20]. The convex formulation proposed in this chapter allows globally optimal Gaussian quantization noise covariance matrices to be found efficiently. In this chapter, the optimization of the quantization noise covariance matrix is performed under the fixed Gaussian input. The joint optimization of the input signal and quantization noise covariance matrices remains a computationally challenging difficult problem [40].

2.1.2 Main Contributions

This chapter establishes several information theoretic results on the uplink MIMO C-RAN model with finite-capacity fronthaul links. A summary of our main contributions is as follows:

- This chapter demonstrates that generalized successive decoding for compress-and-forward, which allows the decoding of the quantization and user message codewords in an arbitrary order, can achieve the same rate region as joint decoding under a sum fronthaul capacity constraint. Further, successive decoding of the quantization codewords first, then the user message codewords, can achieve the same maximum sum rate as joint decoding under individual fronthaul constraints;
- This chapter shows that Gaussian input and Gaussian quantization achieve to within a constant gap of the capacity region of the uplink MIMO C-RAN model under joint decoding. Combining

www.manaraa.com

with the result above, the same constant gap result also holds for generalized successive decoding under a sum fronthaul constraint and for successive decoding for sum rate maximization;

- This chapter shows that under fixed Gaussian input, Gaussian quantization maximizes the achievable rate region under joint decoding. Combining with the optimality result for successive decoding, this also implies that under fixed Gaussian input, Gaussian quantization is optimal for generalized successive decoding under a sum fronthaul constraint, and for successive decoding for sum rate maximization;

- Under joint Gaussian signaling and Gaussian quantization, the optimization of quantization noise covariance matrices for maximizing weighted sum rate under joint decoding and for maximizing sum rate under practical successive decoding can be formulated as convex optimization problems, which facilitate their efficient solution.

2.1.3 Chapter Organization

The rest of the chapter is organized as follows. Section 2.2 introduces the channel model for the uplink MIMO C-RAN and characterizes the achievable rate regions for compress-and-forward schemes with joint decoding and generalized successive decoding respectively. Section 2.3 demonstrates the rate-region optimality of generalized successive decoding under a sum fronthaul constraint and the sum-rate optimality of successive decoding. Section 2.4 focuses on establishing the optimality of Gaussian quantizers with joint decoding under Gaussian input. In addition, Section 2.4 also establishes the approximate capacity of the uplink MIMO C-RAN model to within constant gap, and shows the convex formulation of the (weighted) sum rate maximization problems over the quantization noise covariance matrices. Section 2.5 concludes the chapter.

2.2 Achievable Rate Regions for Uplink C-RAN

2.2.1 Channel Model

This chapter considers an uplink C-RAN model, where K mobile users communicate with a CP through L BSs, as shown in Fig. 2.1. The noiseless digital fronthaul link connecting the BS ℓ to the CP has the capacity of C_ℓ bits per complex dimension. The fronthaul capacity C_ℓ is the maximum long-term average throughput of the ℓth fronthaul link, i.e., $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} C_\ell(i) \le C_\ell$, where $C_\ell(i)$ represents the instantaneous transmission rate of the ℓth fronthaul link at the ith time slot. Each user terminal is equipped with M antennas; each BS is equipped with N antennas. Perfect channel state information (CSI) is assumed to be available to all the BSs and to the CP.

Let $\mathbf{X}_k \in \mathbb{C}^M$ be the signal transmitted by the kth user, which is subject to per-user transmit power constraint of P_k, i.e. $\mathbb{E}\left[\mathbf{X}_k\mathbf{X}_k^\dagger\right] \le P_k$. The signal received at the ℓth BS can be expressed as

$$\mathbf{Y}_\ell = \sum_{k=1}^{K} \mathbf{H}_{\ell,k}\mathbf{X}_k + \mathbf{Z}_\ell, \quad \ell = 1, 2, \ldots, L, \tag{2.1}$$

where $\mathbf{Z}_\ell \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_\ell)$ represents the additive Gaussian noise for BS ℓ and is independent across different BSs, and $\mathbf{H}_{\ell,k}$ denotes the complex channel matrix from user k to BS ℓ.

www.manaraa.com

We consider the compress-and-forward scheme [51,52] applied to the uplink C-RAN system, in which the BSs compress the received signals $\mathbf{Y}_\ell$, and forward the quantization bits to the CP for decoding. At the CP, the user messages are decoded using either joint decoding or some form of successive decoding. In joint decoding, the quantization codewords and the message codewords are decoded *simultaneously*, whereas, in a form of successive decoding, the quantization codewords and message codewords are decoded *successively* in some arbitrary order. Different orderings can potentially result in different achievable rates.

2.2.2 Achievable Rates for Joint Decoding, Successive Decoding, and Generalized Successive Decoding

In the following, we present the achievable rate region of compress-and-forward with joint decoding and different forms of successive decoding.

Proposition 2.1 ([13, Proposition IV.1]) *For the uplink C-RAN model shown in Fig. 2.1, the achievable rate-fronthaul region of compress-and-forward with joint decoding, $\mathcal{P}_{JD}^*$, is the closure of the convex hull of all $(R_1, \cdots, R_K, C_1, \ldots, C_L) \in \mathbb{R}_+^{K+L}$ satisfying*

$$\sum_{k\in\mathcal{T}} R_k < \sum_{\ell\in\mathcal{S}} \left[C_\ell - I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right)\right] + I\left(\mathbf{X}_\mathcal{T}; \hat{\mathbf{Y}}_{\mathcal{S}^c} | \mathbf{X}_{\mathcal{T}^c}\right) \tag{2.2}$$

for all $\mathcal{T} \subseteq \mathcal{K}$ and $\mathcal{S} \subseteq \mathcal{L}$, for some product distribution $\prod_{k=1}^K p(\mathbf{x}_k) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$.

Note that for the uplink C-RAN model, the rate region (2.2) given by compress-and-forward with joint decoding is identical to the rate region of the noisy network coding scheme [16], which is an extension of the compress-and-forward scheme to the general multiple access relay network by using joint decoding at the receiver and block Markov coding at the transmitters.

As a more practical decoding strategy, successive decoding of quantization codewords first, and then the user messages at the CP can also be used in uplink C-RAN. The following proposition states the rate-fronthaul region achieved by successive decoding.

Proposition 2.2 ([12, Theorem 1]) *For the uplink C-RAN model shown in Fig. 2.1, the achievable rate-fronthaul region of compress-and-forward with successive decoding, $\mathcal{P}_{SD}^*$, is the closure of the convex hull of all $(R_1, \cdots, R_K, C_1, \ldots, C_L) \in \mathbb{R}_+^{K+L}$ satisfying*

$$\sum_{k\in\mathcal{T}} R_k < I\left(\mathbf{X}_\mathcal{T}; \hat{\mathbf{Y}}_\mathcal{L} | \mathbf{X}_{\mathcal{T}^c}\right), \quad \forall\, \mathcal{T} \subseteq \mathcal{K}, \tag{2.3}$$

and

$$I\left(\mathbf{Y}_\mathcal{S}; \hat{\mathbf{Y}}_\mathcal{S} | \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) < \sum_{\ell\in\mathcal{S}} C_\ell, \quad \forall\, \mathcal{S} \subseteq \mathcal{L}. \tag{2.4}$$

for some product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$.

Note that (2.3) is the multiple-access rate region, (2.4) represents the Wyner-Ziv decoding constraint, while (2.2) incorporates the joint decoding of the quantization codewords and the user messages. Because of its lower decoding complexity, successive decoding is usually preferred for practical implementation of the uplink C-RAN systems [19, 20].

www.manaraa.com

It is possible to improve upon the successive decoding scheme by allowing arbitrary interleaved decoding orders between quantization codewords and user message codewords. We call this the generalized successive decoding scheme in this chapter. The generalized successive decoding scheme is first suggested in [34] under the name of joint base-station successive interference cancelation scheme. In such a successive decoding strategy, the set of potential decoding orders includes all the permutations of quantization and user message codewords. Denote π as a permutation on the set of quantization and user message codewords $\left(\hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2, \ldots, \hat{\mathbf{Y}}_L, \mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_K\right)$. For a given permutation π, the decoding order is given by the index of the elements in π, i.e. $\pi(1) \to \pi(2) \to \cdots \to \pi(L+K)$. For example, consider an uplink C-RAN model as shown in Fig. 2.1 with 2 BSs and 2 users. If $\pi = \left(\hat{\mathbf{Y}}_1, \mathbf{X}_1, \hat{\mathbf{Y}}_2, \mathbf{X}_2\right)$, then the decoding of $\hat{\mathbf{Y}}_2$ and $\mathbf{X}_2$ can use both previously decoded user messages and quantization codewords as side information. The resulting rate region is characterized as

$$\begin{cases} R_1 < I\left(\mathbf{X}_1; \hat{\mathbf{Y}}_1\right), \\ R_2 < I\left(\mathbf{X}_2; \hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2 | \mathbf{X}_1\right), \end{cases} \tag{2.5}$$

for some product distribution $p(\mathbf{x}_1)p(\mathbf{x}_2)p(\hat{\mathbf{y}}_1|\mathbf{y}_1)p(\hat{\mathbf{y}}_2|\mathbf{y}_2)$ that satisfies

$$\begin{cases} C_1 > I\left(\mathbf{Y}_1; \hat{\mathbf{Y}}_1\right), \\ C_2 > I\left(\mathbf{Y}_2; \hat{\mathbf{Y}}_2 | \hat{\mathbf{Y}}_1, \mathbf{X}_1\right). \end{cases} \tag{2.6}$$

Let $\mathcal{I}_{\mathbf{X}_k}$, $\mathcal{I}_{\mathbf{Y}_\ell}$ denote the indices of user messages that are decoded before $\mathbf{X}_k$ and $\mathbf{Y}_\ell$ under the permutation π, respectively. Likewise, let $\mathcal{J}_{\mathbf{X}_k}$, $\mathcal{J}_{\mathbf{Y}_\ell}$ denote the indices of quantization codewords that are decoded before $\mathbf{X}_k$ and $\mathbf{Y}_\ell$ under the permutation π, respectively. The rate-fronthaul region of generalized successive decoding for uplink C-RAN is stated in the following proposition.

Proposition 2.3 *For the uplink C-RAN model shown in Fig. 2.1, the achievable rate-fronthaul region of generalized successive decoding with decoding order π, $\mathcal{P}_{GSD}(\pi)$, is the closure of the convex hull of all $(R_1, \cdots, R_K, C_1, \ldots, C_L) \in \mathbb{R}_+^{K+L}$ satisfying*

$$R_k < I\left(\mathbf{X}_k; \hat{\mathbf{Y}}_{\mathcal{J}_{\mathbf{X}_k}} | \mathbf{X}_{\mathcal{I}_{\mathbf{X}_k}}\right), \quad \forall\, k \in \mathcal{K}, \tag{2.7}$$

and

$$C_\ell > I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \hat{\mathbf{Y}}_{\mathcal{J}_{\mathbf{Y}_\ell}}, \mathbf{X}_{\mathcal{I}_{\mathbf{Y}_\ell}}\right), \quad \forall\, \ell \in \mathcal{L}. \tag{2.8}$$

for some product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell)$. Define the generalized successive decoding region $\mathcal{P}_{GSD}^*$ to be the closure of the convex hull of the union of regions $\mathcal{P}_{GSD}(\pi)$ over all possible permutation π's, i.e.

$$\mathcal{P}_{GSD}^* = \mathrm{co}\left(\bigcup_\pi \mathcal{P}_{GSD}(\pi)\right). \tag{2.9}$$

2.3 Optimality of Successive Decoding

In general, we have $\mathcal{P}_{SD}^* \subseteq \mathcal{P}_{GSD}^* \subseteq \mathcal{P}_{JD}^*$. However, successive decoding is more desirable than joint decoding, not only because of its lower complexity, but also due to the fact that its rate region can be more easily evaluated. Thus, there is a tradeoff between complexity and performance in designing decoding

www.manaraa.com

strategies for uplink C-RAN. To further understand this tradeoff, this section establishes that: 1) By allowing arbitrary decoding orders of quantization and message codewords, the generalized successive decoding actually achieves the same rate region as joint decoding under a sum fronthaul constraint; 2) The practical successive decoding strategy in which the BSs decode the quantization codewords first, then the user messages, actually achieves the same maximum sum rate as joint decoding under individual fronthaul constraints.

2.3.1 Optimality of Generalized Successive Decoding under a Sum Fronthaul Constraint

This section shows that in the special case where the fronthaul links are subject to a sum capacity constraint, generalized successive decoding achieves the rate region as joint decoding. In this model, the fronthaul capacities are constrained by $\sum_{\ell=1}^{L} C_\ell \leq C$ and $C_\ell \geq 0$, as has been considered in [19, 38]. The sum fronthaul capacity constraint considered here is particularly suited to model the scenario where the fronthaul is implemented in a wireless shared medium. For example, as shown in Fig. 2.2, when the wireless fronthaul links are implemented using an orthogonal access scheme such as time/frequency division multiple access (TDMA or FDMA), and the total number of time/frequency slots that can be utilized by different access points can be shared, the sum-capacity constraint captures the essential feature of the fronthaul constraints.

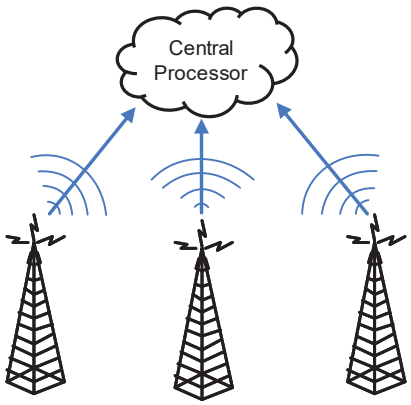


Figure 2.2: An illustration of wireless fronthaul links using TDMA/FDMA with power density constraint.

Under the sum fronthaul capacity constraint C, the rate regions achieved by with joint decoding $\mathcal{R}^*_{JD,s}$ is defined as

$$\mathcal{R}^*_{JD,s} = \left\{ (R_1, \ldots, R_K) : (R_1, \cdots, R_K, C_1, \ldots, C_L) \in \mathcal{P}^*_{JD}, \ \sum_{\ell=1}^{L} C_\ell \leq C, \ C_\ell \geq 0 \right\} \tag{2.10}$$

Likewise, the rate region achieved with generalized successive decoding $\mathcal{R}^*_{GSD,s}$ is given by

$$\mathcal{R}^*_{GSD,s} = \left\{ (R_1, \ldots, R_K) : (R_1, \cdots, R_K, C_1, \ldots, C_L) \in \mathcal{P}^*_{GSD}, \ \sum_{\ell=1}^{L} C_\ell \leq C, \ C_\ell \geq 0 \right\} \tag{2.11}$$

The following theorem states the main result of this section.

www.manaraa.com

Theorem 2.1 *For the uplink C-RAN model with the sum fronthaul capacity constraint* $\sum_{\ell=1}^{L} C_\ell \leq C$ *and* $C_\ell \geq 0$*, the rate region achieved by generalized successive decoding and joint coding are identical, i.e.,* $\mathcal{R}^*_{GSD,s} = \mathcal{R}^*_{JD,s}$.

Proof. See Appendix A. □

The roadmap for the proof of Theorem 2.1 shares the same idea as the characterization of the rate distortion region for the CEO problem under logarithmic loss [53] and the rate distortion region for the multiple-access channel [54], which uses the properties of submodular polyhedron (see Appendix B). Specifically, in order to show $\mathcal{R}^*_{GSD,s} = \mathcal{R}^*_{JD,s}$, we show that under fixed product distribution $\prod_{k=1}^{K} p(\mathbf{x}_k) \prod_{\ell=1}^{L} p(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell)$, every extreme point of the polyhedron $(\mathcal{R}^*_{JD,s}, C)$ is dominated by the points in the polyhedron defined by $(\mathcal{R}^*_{GSD,s}, C)$. We conjecture that Theorem 2.1 holds also for the case of individual fronthaul capacity constraints. However, in that case, finding the dominant faces of polyhedron $(\mathcal{R}^*_{JD,s}, C)$ becomes much more difficult, it appears non-trivial to extend the current proof to the case of individual fronthaul constraints.

2.3.2 Optimality of Successive Decoding for Maximizing Sum Rate

As a special instance of generalized successive decoding, successive decoding reconstructs quantization codewords first, then user message codewords in a sequential order. In what follows, we show that the optimal sum rate achieved by this special successive decoding is the same as that achieved by joint decoding.

Under fixed input distribution and fixed fronthaul capacities C_ℓ, for $\ell = 1, \ldots, L$, the sum rate achieved by joint decoding $R^*_{JD,SUM}$ is defined as

$$R^*_{JD,SUM} = \begin{cases} \max & \sum_{k=1}^{K} R_k \\ \text{s.t.} & (R_1, \cdots, R_K, C_1, \ldots, C_L) \in \mathcal{P}^*_{JD}. \end{cases} \tag{2.12}$$

Likewise, the sum rate for successive decoding $R_{SD,SUM}$ is given by

$$R^*_{SD,SUM} = \begin{cases} \max & \sum_{k=1}^{K} R_k \\ \text{s.t.} & (R_1, \cdots, R_K, C_1, \ldots, C_L) \in \mathcal{P}^*_{SD}. \end{cases} \tag{2.13}$$

The following theorem demonstrates the optimality of successive decoding for maximizing uplink C-RAN under individual fronthaul constraints.

Theorem 2.2 *For the uplink C-RAN model with fronthaul capacities* C_ℓ *shown in Fig. 2.1, the maximum sum rates achieved by successive decoding and joint decoding are the same, i.e.,* $R^*_{SD,SUM} = R^*_{JD,SUM}$.

Proof. See Appendix C. □

We remark that Theorem 2.2 can be thought as a generalization of a result in [42] that shows that under block-by-block forward decoding, the compress-and-forward scheme with compression-message successive decoding achieves the same maximum rate as that with compression-message joint decoding for a single-source, single-destination relay network. The uplink C-RAN is a multiple-source, single-destination relay network. If all the user terminals are regarded as one super transmitter, then it follows

www.manaraa.com

from [42] that successive decoding and joint decoding achieves the same maximum sum rate. However, the proof in [42] is quite complicated. In this chapter, we provide an alternative proof technique for showing the optimality of successive decoding for sum rate maximization in uplink C-RAN. The new proof utilizes the properties of submodular optimization, which is simpler than the proof provided in [42]. The proofs of Theorem 2.2 and Theorem 2.1 illustrate the usefulness of submodular optimization in establishing this type of results.

It is remarked that successive decoding and joint decoding achieve the same sum rate, but do not achieve the same rate region. The achievable rate region of generalized successive decoding is in general larger than that of successive decoding. For example, consider the compress-and-forward scheme for maximizing the rate of user 1, R_1, only. The optimal decoding order should be $\mathbf{X}_{\mathcal{K}\setminus\{1\}} \to \hat{\mathbf{Y}}_{\mathcal{L}} \to \mathbf{X}_1$. With this decoding order, user 1 can achieve larger rate than using the decoding order of $\hat{\mathbf{Y}}_{\mathcal{L}} \to \mathbf{X}_{\mathcal{K}}$, because the decoded user messages $\mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_K$ can serve as side information for the decoding of $\hat{\mathbf{Y}}_{\mathcal{L}}$. In general, to maximize a weighted sum rate, one needs to maximize over $(L+K)!$ orderings for generalized successive decoding. The main result of this section shows however that for maximizing the sum rate in uplink C-RAN, successive decoding of the quantization codewords first, and then the user messages is optimal; this reduces the search space considerably.

2.4 Uplink C-RAN with Gaussian Input and Gaussian Quantization

In this section, we specialize to the compress-and-forward scheme for uplink C-RAN with Gaussian input signal at the users and Gaussian quantization at the BSs. Although it is known that joint Gaussian distribution is suboptimal for uplink C-RAN [12], Gaussian input is desirable, because it leads to achievable rate regions that can be easily evaluated. In the following section, it is shown that with Gaussian input and Gaussian quantization, compress-and-forward with joint decoding can achieve the capacity region of uplink C-RAN to within a constant gap, which is independent of the channel gain matrix and the SNR in the network. We further establish the optimality of Gaussian compression at the relaying BSs for joint decoding, if the input is Gaussian. These results can be further extended to generalized successive decoding under a sum fronthaul constraint and successive decoding for the maximum sum rate. Additionally, under Gaussian signaling, the optimization of quantization noise covariance matrices for weighted sum-rate maximization under joint decoding and for sum rate maximization under practical successive decoding can be cast as convex optimization problems, thereby facilitating their efficient numerical solution. Throughout this section, we focus on the achievable rates under the fixed Gaussian input, and the fixed fronthaul capacity constraints C_ℓ for $\ell = 1, \ldots, L$.

2.4.1 Achievable Rate Regions under Gaussian Input and Gaussian Quantization

We let the input distribution to be Gaussian, i.e., $\mathbf{X}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{K}_k)$, then evaluate the rate regions for the compress-and-forward scheme with joint decoding and successive decoding under Gaussian quantization, denoted as $\mathcal{R}^G_{JD,GIn}$ and $\mathcal{R}^G_{SD,GIn}$, respectively. Set $\prod_{\ell=1}^{L} p(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell) \sim \mathcal{CN}(\mathbf{y}_\ell, \mathbf{Q}_\ell)$, where $\mathbf{Q}_\ell$ is the Gaussian quantization noise covariance matrix at the ℓth BS.

www.manaraa.com

With Gaussian input and Gaussian quantization, we have

$$I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}) = \log \frac{|\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell|}{|\mathbf{Q}_\ell|} \tag{2.14}$$

and

$$I\left(\mathbf{X}_\mathcal{T}; \hat{\mathbf{Y}}_{\mathcal{S}^c} | \mathbf{X}_{\mathcal{T}^c}\right) = \log \frac{\left|\mathbf{H}_{\mathcal{S}^c,\mathcal{T}} \mathbf{K}_\mathcal{T} \mathbf{H}_{\mathcal{S}^c,\mathcal{T}}^\dagger + \operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{S}^c}\right)\right|}{\left|\operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{S}^c}\right)\right|}. \tag{2.15}$$

The achievable rate region (2.2) for joint decoding can be evaluated as

$$\sum_{k \in \mathcal{T}} R_k < \sum_{\ell \in \mathcal{S}} \left[C_\ell - \log \frac{|\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell|}{|\mathbf{Q}_\ell|} \right] + \log \frac{\left|\mathbf{H}_{\mathcal{S}^c,\mathcal{T}} \mathbf{K}_\mathcal{T} \mathbf{H}_{\mathcal{S}^c,\mathcal{T}}^\dagger + \operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{S}^c}\right)\right|}{\left|\operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{S}^c}\right)\right|}, \tag{2.16}$$

for all $\mathcal{T} \subseteq \mathcal{K}$ and $\mathcal{S} \subseteq \mathcal{L}$.

Likewise the achievable rate expression (2.3) for successive decoding becomes

$$\sum_{k \in \mathcal{T}} R_k < \log \frac{\left|\mathbf{H}_{\mathcal{S}^c,\mathcal{K}} \mathbf{K}_\mathcal{K} \mathbf{H}_{\mathcal{L},\mathcal{K}}^\dagger + \operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{L}}\right)\right|}{\left|\operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{L}}\right)\right|}, \quad \forall\, \mathcal{T} \subseteq \mathcal{K}. \tag{2.17}$$

And the fronthaul constraint (2.4) is evaluated as

$$\begin{aligned} I\left(\mathbf{Y}_\mathcal{S}; \hat{\mathbf{Y}}_\mathcal{S} | \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) &\overset{(a)}{=} I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{S} | \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) + \sum_{\ell \in \mathcal{S}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}) \\ &\overset{(b)}{=} I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{L}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) + \sum_{\ell \in \mathcal{S}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}) \\ &= \log \frac{\left|\mathbf{H}_{\mathcal{L},\mathcal{K}} \mathbf{K}_\mathcal{K} \mathbf{H}_{\mathcal{L},\mathcal{K}}^\dagger + \operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{L}}\right)\right|}{\left|\mathbf{H}_{\mathcal{S}^c,\mathcal{K}} \mathbf{K}_\mathcal{K} \mathbf{H}_{\mathcal{S}^c,\mathcal{K}}^\dagger + \operatorname{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{S}^c}\right)\right|} - \sum_{\ell \in \mathcal{S}} \log |\mathbf{Q}_\ell| \le \sum_{\ell \in \mathcal{S}} C_\ell, \end{aligned}$$

for all $\mathcal{S} \subseteq \mathcal{L}$, where equalities (a) and (b) follow from the fact that

$$I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{S} | \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) + I\left(\mathbf{Y}_\mathcal{S}; \hat{\mathbf{Y}}_\mathcal{S} | \mathbf{X}_\mathcal{K} \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) = I\left(\mathbf{Y}_\mathcal{S}; \hat{\mathbf{Y}}_\mathcal{S} | \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) + I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{S} | \mathbf{Y}_\mathcal{S} \hat{\mathbf{Y}}_{\mathcal{S}^c}\right),$$

and the Markov chain

$$\hat{\mathbf{Y}}_i \leftrightarrow \mathbf{Y}_i \leftrightarrow \mathbf{X}_\mathcal{K} \leftrightarrow \mathbf{Y}_j \leftrightarrow \hat{\mathbf{Y}}_j, \quad \forall\, i \neq j.$$

Instead of parameterizing the rate expressions over $\mathbf{Q}_\ell$ as in above, in this section, we introduce the following reparameterization, which is crucial for proving our main results. Define

$$\mathbf{B}_\ell = \left(\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\right)^{-1}. \tag{2.18}$$

We represent the rate regions of joint decoding and successive decoding in terms of $\mathbf{B}_\ell$ in the following.

Proposition 2.4 *For the uplink C-RAN model shown in Fig. 2.1 and under fixed Gaussian input* $\mathbf{X}_\mathcal{K} \sim \mathcal{CN}(\mathbf{0}, \mathbf{K}_\mathcal{K})$ *with* $\mathbf{K}_\mathcal{K} = \operatorname{diag}\left(\{\mathbf{K}_k\}_{k \in \mathcal{K}}\right)$. *The rate-fronthaul region for joint decoding under Gaussian*

www.manaraa.com

quantization, $\mathcal{P}^{G}_{JD,GIn}$, is the closure of the convex hull of all $(R_1,\cdots,R_K,C_1,\ldots,C_L)$ satisfying

$$\sum_{k\in\mathcal{T}} R_k < \sum_{\ell\in\mathcal{S}}\left[C_\ell - \log\frac{|\mathbf{\Sigma}_\ell^{-1}|}{|\mathbf{\Sigma}_\ell^{-1}-\mathbf{B}_\ell|}\right] + \log\frac{\left|\sum_{\ell\in\mathcal{S}^c}\mathbf{H}_{\ell,\mathcal{T}}^{\dagger}\mathbf{B}_\ell\mathbf{H}_{\ell,\mathcal{T}}+\mathbf{K}_{\mathcal{T}}^{-1}\right|}{\left|\mathbf{K}_{\mathcal{T}}^{-1}\right|} \tag{2.19}$$

for all $\mathcal{T}\subseteq\mathcal{K}$ and $\mathcal{S}\subseteq\mathcal{L}$, for some $0\preceq\mathbf{B}_\ell\preceq\mathbf{\Sigma}_\ell^{-1}$, where $\mathbf{K}_{\mathcal{T}}=\mathbb{E}\left[\mathbf{X}_{\mathcal{T}}\mathbf{X}_{\mathcal{T}}^{\dagger}\right]$ is the covariance matrix of $\mathbf{X}_{\mathcal{T}}$, and $\mathbf{H}_{\ell,\mathcal{T}}$ denotes the channel matrix from $\mathbf{X}_{\mathcal{T}}$ to $\mathbf{Y}_\ell$. Furthermore, under the fixed fronthaul capacity constraints C_ℓ for $\ell=1,\ldots,L$, the rate regions achieved by with joint decoding $\mathcal{R}^{G}_{JD,GIn}$ is defined as

$$\mathcal{R}^{G}_{JD,GIn}=\left\{(R_1,\ldots,R_K):(R_1,\cdots,R_K,C_1,\ldots,C_L)\in\mathcal{P}^{G}_{JD,GIn}\right\} \tag{2.20}$$

Proposition 2.5 *For the uplink C-RAN model shown in Fig. 2.1 and under fixed Gaussian input $\mathbf{X}_{\mathcal{K}}\sim\mathcal{CN}(\mathbf{0},\mathbf{K}_{\mathcal{K}})$ with $\mathbf{K}_{\mathcal{K}}=\operatorname{diag}\left(\{\mathbf{K}_k\}_{k\in\mathcal{K}}\right)$. The rate-fronthaul region for successive decoding, $\mathcal{P}^{G}_{SD,GIn}$, is the closure of the convex hull of all $(R_1,\cdots,R_K,C_1,\ldots,C_L)$ satisfying*

$$\sum_{k\in\mathcal{T}} R_k < \log\frac{\left|\sum_{\ell=1}^{L}\mathbf{H}_{\ell,\mathcal{T}}^{\dagger}\mathbf{B}_\ell\mathbf{H}_{\ell,\mathcal{T}}+\mathbf{K}_{\mathcal{T}}^{-1}\right|}{\left|\mathbf{K}_{\mathcal{T}}^{-1}\right|}, \qquad \forall\,\mathcal{T}\subseteq\mathcal{K}, \tag{2.21}$$

and

$$\log\frac{\left|\sum\limits_{\ell=1}^{L}\mathbf{H}_{\ell,\mathcal{K}}^{\dagger}\mathbf{B}_\ell\mathbf{H}_{\ell,\mathcal{K}}+\mathbf{K}_{\mathcal{K}}^{-1}\right|}{\left|\sum\limits_{\ell\in\mathcal{S}^c}\mathbf{H}_{\ell,\mathcal{K}}^{\dagger}\mathbf{B}_\ell\mathbf{H}_{\ell,\mathcal{K}}+\mathbf{K}_{\mathcal{K}}^{-1}\right|} + \sum_{\ell\in\mathcal{S}}\log\frac{|\mathbf{\Sigma}_\ell^{-1}|}{|\mathbf{\Sigma}_\ell^{-1}-\mathbf{B}_\ell|} < \sum_{\ell\in\mathcal{S}} C_\ell, \qquad \forall\,\mathcal{S}\subseteq\mathcal{L}, \tag{2.22}$$

for some $0\preceq\mathbf{B}_\ell\preceq\mathbf{\Sigma}_\ell^{-1}$, where $\mathbf{K}_{\mathcal{T}}=\mathbb{E}\left[\mathbf{X}_{\mathcal{T}}\mathbf{X}_{\mathcal{T}}^{\dagger}\right]$ is the covariance matrix of $\mathbf{X}_{\mathcal{T}}$, and $\mathbf{H}_{\ell,\mathcal{T}}$ denotes the channel matrix from $\mathbf{X}_{\mathcal{T}}$ to $\mathbf{Y}_\ell$. Moreover, under the fixed fronthaul capacity constraints C_ℓ for $\ell=1,\ldots,L$, the rate regions achieved by with successive decoding $\mathcal{R}^{G}_{SD,GIn}$ is defined as

$$\mathcal{R}^{G}_{SD,GIn}=\left\{(R_1,\ldots,R_K):(R_1,\cdots,R_K,C_1,\ldots,C_L)\in\mathcal{P}^{G}_{SD,GIn}\right\} \tag{2.23}$$

2.4.2 Gaussian Input and Gaussian Quantization Achieve Capacity to within Constant Gap

With Gaussian input and Gaussian quantization, the rate region of joint decoding (2.19) can be shown to be within a constant gap to the capacity region of uplink C-RAN. This constant-gap result is stated in the following theorem.

Theorem 2.3 *For any rate tuple $(R_1,R_2,\ldots,R_K)$ within the cut-set bound for uplink C-RAN with fixed fronthaul capacities of C_ℓ shown in Fig. 2.1, the rate tuple $(R_1-\eta,R_2-\eta,\ldots,R_K-\eta)$, with $\eta=NL+M$ is achievable for compress-and-forward with Gaussian input, Gaussian quantization, and joint decoding, where L is the number of BSs in the network, M is the number of transmit antennas at user, and N is the number of receive antennas at BS, i.e., $(R_1-\eta,R_2-\eta,\ldots,R_K-\eta)\in\mathcal{R}^{G}_{JD,GIn}$.*

Proof. See Appendix D. □

Although the uplink C-RAN model is an example of a relay network for which noisy network coding approach applies and it is known that compress-and-forward with joint decoding achieves the same rate

www.manaraa.com

region as noisy network coding for uplink C-RAN, we remark that Theorem 2.3 does not immediately follow from the constant-gap optimality result of noisy network coding [16]. The constant-gap optimality of noisy network coding is proven for Gaussian relay networks, whereas the uplink C-RAN model contains fronthaul links which are digital connections and not Gaussian channels.

Combining with our earlier results on the optimality of successive decoding, constant-gap optimality results can also be obtained for compress-and-forward with generalized successive decoding and successive decoding. These results are summarized in the following corollary.

Corollary 2.1 *For the uplink C-RAN model as shown in Fig. 2.1, compress-and-forward with generalized successive decoding, under Gaussian input and Gaussian quantization achieves the capacity region to within $NL+M$ bits per complex dimension if the fronthaul links are subjected to a sum capacity constraint $\sum_{\ell=1}^{L} C_\ell \leq C$. Furthermore, compress-and-forward with successive decoding, under Gaussian input and Gaussian quantization, achieves the sum capacity of an uplink C-RAN model with individual fronthaul constraints to within $NL + MK$ bits per complex dimension.*

2.4.3 Optimality of Gaussian Quantization under Joint Decoding

For the Gaussian uplink MIMO C-RAN model, it is known that Gaussian input and Gaussian quantization are not jointly optimal [12]. However, if the quantization noise is fixed as Gaussian, then the optimal input distribution must be Gaussian. This is because the channel reduces to a conventional Gaussian multiple-access channel in this case. The main result of this section is that the converse is also true, i.e., under fixed Gaussian input, Gaussian quantization actually maximizes the achievable rate region of the uplink C-RAN model under joint decoding. The work in this section is done jointly with Jun Chen and Yinfei Xu. Theorem 2.4 and the proof are contributions of Yinfei Xu.

Under fixed fronthaul capacity constraints C_ℓ for $\ell = 1, \ldots, L$, we let $\mathcal{R}^*_{JD,GIn}$ denote the rate region of joint decoding under Gaussian input and optimal quantization. In the following, we first define Fisher information and state the two main tools for proving this result: the Bruijn identity and the Fisher information inequality. We then present the main theorem on the optimality of Gaussian quantization for joint decoding, i.e., $\mathcal{R}^G_{JD,GIn} = \mathcal{R}^*_{JD,GIn}$.

Definition 2.4.1. Let $\mathbf{X}$ be any random vector with probability density function $f(\mathbf{x})$. The Fisher information of the distribution of $\mathbf{X}$ is defined as

$$\mathbf{J}(\mathbf{X}) = \mathbb{E}\left[(\nabla \log f(\mathbf{x}))(\nabla \log f(\mathbf{x}))^T\right] \tag{2.24}$$

Lemma 2.1 (Fisher Information Inequality, [55] [47, Lemma 2]) *Let $(\mathbf{U}, \mathbf{X})$ be an arbitrary complex random vector, where the conditional Fisher information of $\mathbf{X}$ conditioned on $\mathbf{U}$ exists. We have*

$$\log\left|(\pi e)\mathbf{J}^{-1}(\mathbf{X}|\mathbf{U})\right| \leq h(\mathbf{X}|\mathbf{U}). \tag{2.25}$$

Lemma 2.2 (Bruijn Identity, [56] [47, Lemma 3]) *Let $(\mathbf{V}_1, \mathbf{V}_2)$ be an arbitrary random vector with finite second moments, and $\mathbf{N}$ be a zero-mean Gaussian random vector with covariance $\mathbf{\Lambda}_N$. Assume $(\mathbf{V}_1, \mathbf{V}_2)$ and $\mathbf{N}$ are independent. We have*

$$\mathrm{cov}(\mathbf{V}_2|\mathbf{V}_1, \mathbf{V}_2 + \mathbf{N}) = \mathbf{\Lambda}_N - \mathbf{\Lambda}_N \mathbf{J}(\mathbf{V}_2 + \mathbf{N}|\mathbf{V}_1)\mathbf{\Lambda}_N. \tag{2.26}$$

www.manaraa.com

Theorem 2.4 *For the uplink C-RAN under fixed Gaussian input distribution and assuming joint decoding, Gaussian quantization is optimal, i.e.* $\mathcal{R}^{G}_{JD,GIn} = \mathcal{R}^{*}_{JD,GIn}$.

Proof. Recall that the achievable rate region of the compress-and-forward scheme under joint decoding is given by the set of $(R_1, \ldots, R_K)$ derived from (2.2) under the joint distribution

$$p\left(\mathbf{x}_1, \ldots, \mathbf{x}_K, \mathbf{y}_1, \ldots, \mathbf{y}_L, \hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_L\right) = \prod_{k=1}^{K} p\left(\mathbf{x}_k\right) \prod_{\ell=1}^{L} p\left(\mathbf{y}_\ell | \mathbf{x}_1, \ldots, \mathbf{x}_K\right) \prod_{\ell=1}^{L} p\left(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell\right). \tag{2.27}$$

For fixed Gaussian input $\mathbf{X}_{\mathcal{K}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{K}_{\mathcal{K}})$ and fixed $\prod_{\ell=1}^{L} p(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell)$, choose $\mathbf{B}_\ell$ with $\mathbf{0} \preceq \mathbf{B}_\ell \preceq \mathbf{\Sigma}_\ell^{-1}$ such that

$$\operatorname{cov}\left(\mathbf{Y}_\ell | \mathbf{X}_{\mathcal{K}}, \hat{\mathbf{Y}}_\ell\right) = \mathbf{\Sigma}_\ell - \mathbf{\Sigma}_\ell \mathbf{B}_\ell \mathbf{\Sigma}_\ell, \quad \ell = 1, \cdots, L.$$

We proceed to show that the achievable rate region as given by (2.19) with a Gaussian $\prod_{\ell=1}^{L} p(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell) \sim \mathcal{CN}(\mathbf{Y}_\ell, \mathbf{Q}_\ell)$, where $\mathbf{Q}_\ell = \mathbf{B}_\ell^{-1} - \mathbf{\Sigma}_\ell$, is as large as that of (2.2) under Gaussian input.

First, note that

$$\begin{aligned} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}\right) &= \log\left|(\pi e)\mathbf{\Sigma}_\ell\right| - h\left(\mathbf{Y}_\ell | \mathbf{X}_{\mathcal{K}}, \hat{\mathbf{Y}}_\ell\right) \\ &\geq \log\left|(\pi e)\mathbf{\Sigma}_\ell\right| - \log\left|(\pi e)\operatorname{cov}\left(\mathbf{Y}_\ell | \mathbf{X}_{\mathcal{K}}, \hat{\mathbf{Y}}_\ell\right)\right| \\ &= \log \frac{\left|\mathbf{\Sigma}_\ell^{-1}\right|}{\left|\mathbf{\Sigma}_\ell^{-1} - \mathbf{B}_\ell\right|}, \quad \ell = 1, \cdots, L, \end{aligned} \tag{2.28}$$

where we use the fact that Gaussian distribution maximizes differential entropy.

Moreover, we have

$$\begin{aligned} I\left(\mathbf{X}_{\mathcal{T}}; \hat{\mathbf{Y}}_{\mathcal{S}^c} | \mathbf{X}_{\mathcal{T}^c}\right) &= h\left(\mathbf{X}_{\mathcal{T}}\right) - h\left(\mathbf{X}_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}^c}, \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) \\ &\leq \log\left|\mathbf{K}_{\mathcal{T}}\right| - \log\left|\mathbf{J}^{-1}\left(\mathbf{X}_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}^c}, \hat{\mathbf{Y}}_{\mathcal{S}^c}\right)\right|, \end{aligned}$$

where the inequality is due to Lemma 2.1. Since

$$\mathbf{Y}_{\mathcal{S}^c} = \mathbf{H}_{\mathcal{S}^c,\mathcal{T}} \mathbf{X}_{\mathcal{T}} + \mathbf{H}_{\mathcal{S}^c,\mathcal{T}^c} \mathbf{X}_{\mathcal{T}^c} + \mathbf{Z}_{\mathcal{S}^c},$$

it follows from the MMSE estimation of Gaussian random vectors that

$$\begin{aligned} \mathbf{X}_{\mathcal{T}} &= \mathbb{E}\left[\mathbf{X}_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}^c}, \mathbf{Y}_{\mathcal{S}^c}\right] + \mathbf{N}_{\mathcal{T},\mathcal{S}^c} \\ &= \sum_{\ell \in \mathcal{S}^c} \mathbf{G}_{\mathcal{T},\ell}\left(\mathbf{Y}_\ell - \mathbf{H}_{\ell,\mathcal{T}^c} \mathbf{X}_{\mathcal{T}^c}\right) + \mathbf{N}_{\mathcal{T},\mathcal{S}^c}, \end{aligned}$$

where

$$\mathbf{G}_{\mathcal{T},\ell} = \left(\mathbf{K}_{\mathcal{T}}^{-1} + \sum_{j \in \mathcal{S}^c} \mathbf{H}_{j,\mathcal{T}}^{\dagger} \mathbf{\Sigma}_j^{-1} \mathbf{H}_{j,\mathcal{T}}\right)^{-1} \mathbf{H}_{\ell,\mathcal{T}}^{\dagger} \mathbf{\Sigma}_\ell^{-1},$$

www.manaraa.com

and $\mathbf{N}_{\mathcal{T},\mathcal{S}^c} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Lambda_N})$ with covariance matrix

$$\mathbf{\Lambda_N} = \left(\mathbf{K}_{\mathcal{T}}^{-1} + \sum_{\ell \in \mathcal{S}^c} \mathbf{H}_{\ell,\mathcal{T}}^{\dagger} \mathbf{\Sigma}_{\ell}^{-1} \mathbf{H}_{\ell,\mathcal{T}} \right)^{-1}. \tag{2.29}$$

Here $\mathbb{E}[\mathbf{X}_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}^c}, \mathbf{Y}_{\mathcal{S}^c}]$ is the MMSE estimator of $\mathbf{X}_{\mathcal{T}}$ from $\mathbf{X}_{\mathcal{T}^c}, \mathbf{Y}_{\mathcal{S}^c}$. The error in estimation is $\mathbf{N}_{\mathcal{T},\mathcal{S}^c}$, and the MMSE matrix is $\mathbf{\Lambda_N}$.

By the matrix complementary identity between Fisher information matrix and MMSE in Lemma 2.2, we have

$$\begin{aligned}
\mathbf{J}\left(\mathbf{X}_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}^c}, \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) &= \mathbf{\Lambda}_{\mathbf{N}}^{-1} - \mathbf{\Lambda}_{\mathbf{N}}^{-1} \operatorname{cov}\left(\sum_{\ell \in \mathcal{S}^c} \mathbf{G}_{\mathcal{T},\ell} (\mathbf{Y}_{\ell} - \mathbf{H}_{\ell,\mathcal{T}^c} \mathbf{X}_{\mathcal{T}^c}) | \mathbf{X}_{\mathcal{K}}, \hat{\mathbf{Y}}_{\mathcal{S}^c} \right) \mathbf{\Lambda}_{\mathbf{N}}^{-1} \\
&= \mathbf{\Lambda}_{\mathbf{N}}^{-1} - \mathbf{\Lambda}_{\mathbf{N}}^{-1} \operatorname{cov}\left(\sum_{\ell \in \mathcal{S}^c} \mathbf{G}_{\mathcal{T},\ell} \mathbf{Y}_{\ell} | \mathbf{X}_{\mathcal{K}}, \hat{\mathbf{Y}}_{\mathcal{S}^c} \right) \mathbf{\Lambda}_{\mathbf{N}}^{-1} \\
&= \mathbf{\Lambda}_{\mathbf{N}}^{-1} - \mathbf{\Lambda}_{\mathbf{N}}^{-1} \left[\sum_{\ell \in \mathcal{S}^c} \mathbf{G}_{\mathcal{T},\ell} \operatorname{cov}\left(\mathbf{Y}_{\ell} | \mathbf{X}_{\mathcal{K}}, \hat{\mathbf{Y}}_{\ell} \right) \mathbf{G}_{\mathcal{T},\ell}^{\dagger} \right] \mathbf{\Lambda}_{\mathbf{N}}^{-1} \\
&= \mathbf{\Lambda}_{\mathbf{N}}^{-1} - \sum_{\ell \in \mathcal{S}^c} \mathbf{H}_{\ell,\mathcal{T}}^{\dagger} \left(\mathbf{\Sigma}_{\ell}^{-1} - \mathbf{B}_{\ell} \right) \mathbf{H}_{\ell,\mathcal{T}} \\
&= \mathbf{K}_{\mathcal{T}}^{-1} + \sum_{\ell \in \mathcal{S}^c} \mathbf{H}_{\ell,\mathcal{T}}^{\dagger} \mathbf{B}_{\ell} \mathbf{H}_{\ell,\mathcal{T}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
I\left(\mathbf{X}_{\mathcal{T}}; \hat{\mathbf{Y}}_{\mathcal{S}^c} | \mathbf{X}_{\mathcal{T}^c}\right) &\leq \log \frac{\left| J(\mathbf{X}_{\mathcal{T}} | \mathbf{X}_{\mathcal{T}^c}, \hat{\mathbf{Y}}_{\mathcal{S}^c}) \right|}{\left| \mathbf{K}_{\mathcal{T}}^{-1} \right|} \\
&= \log \frac{\left| \mathbf{K}_{\mathcal{T}}^{-1} + \sum_{\ell \in \mathcal{S}^c} \mathbf{H}_{\ell,\mathcal{T}}^{\dagger} \mathbf{B}_{\ell} \mathbf{H}_{\ell,\mathcal{T}} \right|}{\left| \mathbf{K}_{\mathcal{T}}^{-1} \right|}
\end{aligned} \tag{2.30}$$

for all $\mathcal{T} \subseteq \mathcal{K}$ and $\mathcal{S} \subseteq \mathcal{L}$. Combining (2.28) and (2.30), we conclude that $\mathcal{R}_{JD,GIn}^{G}$ as derived from (2.19) is as large as $\mathcal{R}_{JD,GIn}^{*}$. Therefore, $\mathcal{R}_{JD,GIn}^{G} = \mathcal{R}_{JD,GIn}^{*}$. □

2.4.4 Optimization of Gaussian Input and Gaussian Quantization Noise Covariance Matrices

This section addresses the numerical optimization of the Gaussian input and quantization noise covariance matrices for uplink MIMO C-RAN under given fronthaul capacity constraints. First, we note that even when restricting to Gaussian input and Gaussian quantization, the joint optimization of input and quantization noise covariance matrices is still a challenging problem for the uplink MIMO C-RAN. However, if we fix the quantization noise covariance, then the input optimization reduces to that of optimizing a conventional Gaussian multiple-access channel. In particular, the problem of maximizing the weighted sum rate can be formulated as a convex optimization, which can be readily solved [57].

Conversely, if we fix the transmit covariance matrix, the optimization of quantization noise covariance can in some cases be formulated as convex optimization. The key enabling fact is the reparameterization in term of $\mathbf{B}_{\ell}$ (2.18), instead of direct optimization over $\mathbf{Q}_{\ell}$. Consider first the case of joint decoding.

www.manaraa.com

Using (2.19) under the fixed C_ℓ for $\ell = 1, \ldots, L$, the weighted sum rate maximization problem can be formulated over $\{R_k, \mathbf{B}_\ell\}$ as follows:

$$\begin{aligned} \max_{R_k, \mathbf{B}_\ell} \quad & \sum_{k=1}^{K} \mu_k R_k \\ \text{s.t.} \quad & \sum_{k\in\mathcal{T}} R_k \le \sum_{\ell\in\mathcal{S}} \left[C_\ell - \log \frac{|\mathbf{\Sigma}_\ell^{-1}|}{|\mathbf{\Sigma}_\ell^{-1} - \mathbf{B}_\ell|} \right] \\ & \qquad + \log \frac{\left| \sum_{\ell\in\mathcal{S}^c} \mathbf{H}_{\ell,\mathcal{T}}^\dagger \mathbf{B}_\ell \mathbf{H}_{\ell,\mathcal{T}} + \mathbf{K}_{\mathcal{T}}^{-1} \right|}{\left|\mathbf{K}_{\mathcal{T}}^{-1}\right|}, \quad \forall\, \mathcal{T} \subseteq \mathcal{K}, \quad \forall \mathcal{S} \subseteq \mathcal{L}, \\ & \mathbf{0} \preceq \mathbf{B}_\ell \preceq \mathbf{\Sigma}_\ell^{-1}, \quad \forall\, \ell \in \mathcal{L}. \end{aligned} \tag{2.31}$$

where μ_k represents the weight associated with user k, which is typically determined from upper layer protocols. The key observation is that the above problem is convex in $\{R_k, \mathbf{B}_\ell\}$. However, we also note that because of joint decoding, the number of constraints is exponential in the size of the network. Consequently, the above optimization problem can only be solved for small networks in practice.

Note that the above formulation considers the optimization of instantaneous achievable rates R_k under instantaneous fronthaul capacity constraints C_ℓ in a fixed time slot. The solution obtained, however, also applies to the more general case of optimizing the average rates under average fronthaul. This is because if we consider a slightly more general formulation of optimizing an objective of

$$\max_{R_k, \mathbf{B}_\ell, C_\ell} \sum_{k=1}^{K} \mu_k R_k - \sum_{\ell=1}^{L} \nu_\ell C_\ell \tag{2.32}$$

under the same constraints as in (2.31). Such an optimization problem is convex, so time-sharing is not needed. For this reason, the rest of this section considers the formulation with instantaneous rates only.

We now consider the weighted sum-rate maximization problem for the case of successive decoding of the quantization codewords followed by the user messages. However, the direct characterization of successive decoding rate (2.21)-(2.22) does not give rise to a convex formulation. Nevertheless, for the special case of maximizing the sum rate (i.e., with $\mu_1 = \cdots = \mu_K = 1$), using Theorem 2.2, which shows that successive decoding achieves the same maximum sum rate as joint decoding, the sum-rate maximization problem with successive decoding can be equivalently formulated as follows:

Theorem 2.5 *For the uplink C-RAN model with individual fronthaul capacity constraint C_ℓ as shown in Fig. 2.1, the sum rate maximization problem under successive decoding can be formulated as the following convex problem:*

$$\begin{aligned} \max_{R, \mathbf{B}_\ell} \quad & R \\ \text{s.t.} \quad & R \le \sum_{\ell\in\mathcal{S}} \left[C_\ell - \log \frac{|\mathbf{\Sigma}_\ell^{-1}|}{|\mathbf{\Sigma}_\ell^{-1} - \mathbf{B}_\ell|} \right] \\ & \qquad + \log \frac{\left| \sum_{\ell\in\mathcal{S}^c} \mathbf{H}_{\ell,\mathcal{T}}^\dagger \mathbf{B}_\ell \mathbf{H}_{\ell,\mathcal{T}} + \mathbf{K}_{\mathcal{K}}^{-1} \right|}{\left|\mathbf{K}_{\mathcal{K}}^{-1}\right|}, \quad \forall \mathcal{S} \subseteq \mathcal{L}, \\ & \mathbf{0} \preceq \mathbf{B}_\ell \preceq \mathbf{\Sigma}_\ell^{-1}, \quad \forall\, \ell \in \mathcal{L}. \end{aligned} \tag{2.33}$$

www.manaraa.com

Furthermore, if the fronthaul links are subject to a sum capacity constraint of C, the sum rate maximization problem can be formulated as the following convex problem:

$$\begin{aligned} \max_{R,\mathbf{B}_\ell} \quad & R \\ \text{s.t.} \quad & R \le \log \frac{\left|\sum_{\ell=1}^{L} \mathbf{H}_{\ell,\mathcal{K}}^{\dagger} \mathbf{B}_\ell \mathbf{H}_{\ell,\mathcal{K}} + \mathbf{K}_{\mathcal{K}}^{-1}\right|}{\left|\mathbf{K}_{\mathcal{K}}^{-1}\right|}, \\ & R + \sum_{\ell=1}^{L} \log \frac{|\mathbf{\Sigma}_\ell^{-1}|}{|\mathbf{\Sigma}_\ell^{-1} - \mathbf{B}_\ell|} \le C, \\ & \mathbf{0} \preceq \mathbf{B}_\ell \preceq \mathbf{\Sigma}_\ell^{-1}, \quad \forall\, \ell \in \mathcal{L}. \end{aligned} \tag{2.34}$$

We remark that the formulation for uplink C-RAN with individual fronthaul capacities (2.33) has exponential number of constraints, because the CP in effect needs to search over $L!$ different decoding orders of quantization codewords at the BSs. In practical implementation, a heuristic method can be used to determine the decoding orders of quantization codewords for avoiding the exponential search [40, 58]. Alternatively, if the C-RAN has a sum fronthaul constraint, then the number of constraints is linear in network size, because we only need to consider the case of $\mathcal{S} = \mathcal{L}$ and $\mathcal{S} = \emptyset$ in (2.33). Consequently, the resulting quantization noise covariance optimization problem (2.34) can be solved in polynomial time. Note that convexity is a key advantage of the above problem formulations as compared to previous approaches in the literature (e.g. [19, 20]) that parameterize the optimization problem over the quantization noise covariance $\mathbf{Q}_\ell$, which leads to a nonconvex formulation.

We emphasize the importance of Gaussian input for the convex formulation in Theorem 2.5. Suppose that both input signal $\mathbf{X}_{\mathcal{K}}$ and compressed signal $\hat{\mathbf{Y}}_\ell$ are discrete random vectors with finite alphabet. For fixed input distribution, the sum-rate maximization problem under the sum fronthaul constraint can be written as

$$\begin{aligned} \max_{p(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell)} \quad & I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{\mathcal{L}}\right), \\ \text{s.t.} \quad & I\left(\mathbf{Y}_{\mathcal{L}}; \hat{\mathbf{Y}}_{\mathcal{L}}\right) \le C, \\ & p\left(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell\right) \ge 0, \ \sum_{\hat{\mathbf{y}}_\ell} p\left(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell\right) = 1, \ \forall\, \ell \in \mathcal{L}. \end{aligned} \tag{2.35}$$

The above problem can be thought as a variant of the information bottleneck method [48], which can be solved by a generalized Blahut-Arimoto (BA) algorithm [59, 60]. However, due to the non-convex nature of problem (2.35), the generalized BA algorithm can only converge to a local optimum.

2.5 Summary

This chapter provides a number of information theoretical results on the compress-and-forward scheme for an uplink MIMO C-RAN model with capacity-limited fronthaul. The relationship between different rate regions for compress-and-forward is illustrated in Fig. 2.3. It is shown that the generalized successive decoding scheme, which allows arbitrary decoding orders between quantization and message codewords, can achieve the same rate region as joint decoding under a sum fronthaul constraint. Moreover, the practical successive decoding of the quantization codewords followed by the user messages is

www.manaraa.com

shown to achieve the same maximum sum rate as joint decoding under individual fronthaul constraints. Additionally, if the input distribution is assumed to be Gaussian, it is shown that Gaussian quantization maximizes the achievable rate region of joint decoding. With Gaussian input signaling, the optimization of Gaussian quantization for maximizing the weighted sum rate under joint decoding and the sum rate under successive decoding can be cast as convex optimization problems, which facilitates its efficient numerical solution. Finally, Gaussian input and Gaussian quantization achieve the capacity region of the uplink C-RAN model to within constant gap.

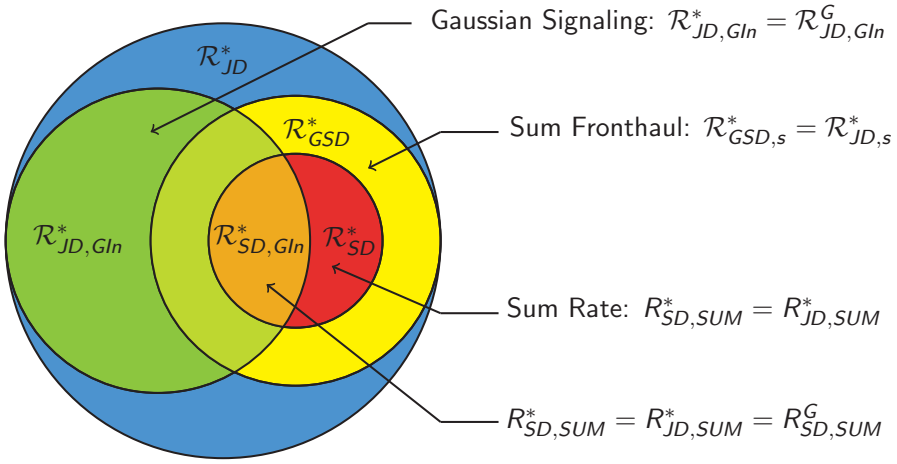


Figure 2.3: Relationship between the rate regions under compress-and-forward in uplink C-RAN

Collectively, these results provide justifications for the practical choice of using Gaussian input signals at the user terminals, Gaussian quantization at the relaying BSs, and successive decoding of quantization codewords followed by user messages at the CP for implementing uplink MIMO C-RAN. Under this choice of fronthaul compression and decoding strategies, the following chapters further study the optimization of quantization noise covariance and transmit signals for maximizing the network utility of the uplink C-RAN system with different fronthaul constraints. Specifically, Chapter 3 investigates the optimization of quantization noise levels for uplink C-RAN under a sum fronthaul capacity constraint. Chapter 4 further studies the joint optimization of transmit beamforming and fronthaul compression for uplink C-RAN under individual fronthaul capacity constraints.

www.manaraa.com

Chapter 3

Optimized Compression under a Sum Fronthaul Constraint

3.1 Introduction

In Chapter 2, we have shown that under compress-and-forward Gaussian quantization is optimal if the input is Gaussian and successive decoding of quantization codewords first, and then user messages, can achieve the maximum sum rate for uplink C-RAN. This chapter further deals with the practical fronthaul design for uplink C-RAN using the compress-and-forward scheme with Gaussian quantization at BS and successive decoding at CP. The uplink of C-RAN model, as shown in Fig. 3.1, consists of multiple remote users sending independent messages while interfering with each other at their respective BSs. The BSs are connected to the CP via noiseless fronthaul links with a finite sum capacity constraint C. The user messages are eventually decoded at the CP. This uplink C-RAN model can be thought of as a *virtual multiple-access channel* (VMAC) between the users and the CP, with the BSs acting as *relays*. The antennas of multiple BSs essentially become a virtual MIMO antenna array capable of spatially multiplexing multiple user terminals.

To explore the advantage of the C-RAN architecture, this chapter considers a compress-and-forward relay strategy in which the BSs send compressed version of their received signals to the CP through the fronthaul, and the CP either jointly or successively decodes all the user messages. Depending on the different compression strategies used at BSs, either with Wyner-Ziv (WZ) coding or with single-user (SU) compression, the coding strategies in this chapter are named VMAC-WZ or VMAC-SU respectively. A key parameter in fronthaul compression design is the level of quantization noise introduced by the compression operation. The main objective of this chapter is to identify efficient algorithms for the optimal setting of quantization noise levels for maximizing the network utility of uplink C-RAN with sum-capacity limited fronthaul.

3.1.1 Related Work

The achievable rates and the relay strategy of the uplink C-RAN architecture have been studied previously in the information theory literature. Under a Wyner model, the achievable rate of an uplink cellular network with BS cooperation is studied in [61] assuming unlimited cooperation, then extended

www.manaraa.com

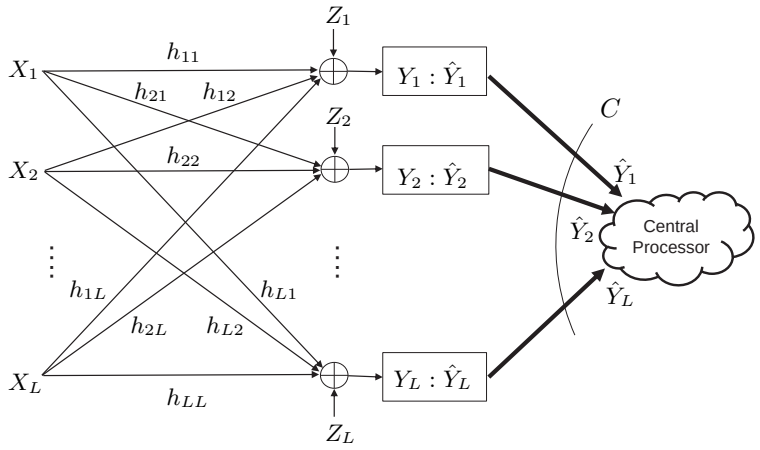


Figure 3.1: The uplink of a cloud radio access network with a finite sum fronthaul

to the limited cooperation case in [62], where the performances of relaying strategies such as decode-and-forward and compress-and-forward are evaluated.

The uplink C-RAN model considered in this chapter is closely related to that in [12–14], where the fundamental achievable rates using the compress-and-forward strategy are characterized under individual fronthaul capacity constraints. The achievable rates of [12–14] are derived assuming that the quantization codewords and the user messages are decoded jointly at the CP. However, such a joint decoding strategy is computationally complex. Further, the question of how to optimally set the quantization noise level is left open.

The uplink C-RAN model can be thought of as a particular instance of a general relay network with a single destination for which several recent works [15–17] have been able to characterize the information theoretical capacity to within a constant gap. The achievability schemes of [15–17] are still based on joint decoding, but with the new insight that in order to achieve to within a constant gap to the outer bound, the quantization noise level should be set at the background noise level.

This chapter goes one step further in identifying relaying and decoding schemes that have lower complexity than joint decoding, while maintaining certain optimality. Toward this end, this chapter shows that a *successive* decoding strategy in which the CP first decodes the quantization codewords, then decodes the user messages based on the quantized signals from all BSs can achieve to within a constant gap to the sum capacity of the network. We note that the proposed scheme is different and performs better than the per-BS successive interference cancellation (SIC) scheme of [34], where each user message is decoded based on the quantization codeword of its own BS only and the previously decoded messages.

A main focus of this chapter is the optimization of the quantization noise levels at the BSs for the uplink C-RAN model. In this direction, the present chapter is related to the work of [19], which uses a gradient approach to solve a quantization noise level optimization problem for a closely related problem. The present chapter is also closely related to [20], where the quantization noise level optimization problem is solved on a per-BS basis (and the robustness of the optimization procedure is addressed in addition). In contrast, the algorithm proposed in this chapter involves a more direct optimization objective where the quantization noise levels of all BSs are optimized jointly.

www.manaraa.com

3.1.2 Main Contributions

From a theoretical capacity analysis perspective, this chapter shows that VMAC-WZ with successive decoding can achieve the sum capacity of the C-RAN model to within a constant gap, while VMAC-SU achieves the sum capacity to within a constant gap under a channel diagonal dominant condition. Since the VMAC schemes have the advantage of low decoding complexity and low decoding delay as compared to joint decoding, the constant-gap results provide a strong motivation for the possible implementation of the VMAC schemes in practical C-RAN systems.

From an optimization perspective, this chapter proposes an alternating convex optimization algorithm for optimizing the quantization noise levels for weighted sum-rate maximization for the VMAC-WZ scheme, and proposes reformulation of the problem in term of optimizing fronthaul capacities for the VMAC-SU scheme. Further, this chapter shows that in the high signal-to-quantization-noise-ratio (SQNR) regime, the quantization noise level should be set to be proportional to the background noise level, regardless of the transmit power and the channel condition. Based on this observation, low-complexity algorithms are developed for the quantization noise level design in practical C-RAN scenarios.

Finally, this chapter evaluates the performance of the proposed VMAC schemes in multicell networks and in heterogeneous topologies where macro- and pico-cells may have significantly different fronthaul capacity constraints. Numerical simulations show that the C-RAN architecture can bring significant performance improvement, and that the proposed approximate quantization noise level setting can already realize much of the gains.

3.1.3 Chapter Organization

The rest of the chapter is organized as follows. Section 3.2 introduces the VMAC scheme with WZ compression and with SU compression. Section 3.3 focuses on optimizing the quantization noise level for the VMAC-WZ scheme, where an alternating convex optimization algorithm and an approximation algorithm are proposed. It is shown that the VMAC-WZ scheme achieves the sum capacity of the uplink C-RAN model to within a constant gap. Section 3.4 focuses on the optimization of quantization noise levels for the VMAC-SU scheme, and formulates an equivalent fronthaul capacity allocation problem. A constant-gap capacity result for the VMAC-SU scheme is demonstrated. The proposed VMAC schemes are evaluated numerically for practical multicell/picocell networks in Section 3.5. Conclusions are drawn in Section 3.6.

3.2 Preliminaries

3.2.1 System Model

This chapter considers the uplink C-RAN, where L single-antenna remote users send independent messages to L single-antenna BSs forming a fixed cluster, as shown in Fig. 3.1[1]. The BSs are connected to a CP through noiseless fronthaul links of capacities C_i, $i = 1, \ldots, L$. The user messages need to be eventually decoded at the CP. A key modelling assumption of this chapter is that the fronthaul capacities C_i can be adapted to the channel condition and user traffic demand, subject to an overall capacity

[1]For simple notation, this chapter assumes that both the number of users and the number of BSs are L. All the results in this chapter hold for the general case where there are K users and L BSs in the network.

www.manaraa.com

constraint, i.e. $\sum_{i=1}^{L} C_i \leq C$, which is justified in Chapter 2. For simplicity, both the remote users and the BSs are assumed to have a single antenna each here, but most results of this chapter can be extended to the MIMO case.

The uplink C-RAN model can be thought of as an $L \times L$ interference channel between the users and the BSs, followed by a noiseless multiple-access channel between the BSs and the CP. Alternatively, it can also be thought of as a virtual multiple-access channel between the users and the CP with the BSs serving as relay nodes. Let X_i denote the signal transmitted by the ith user. The signal received at the ith BS can be expressed as

$$Y_i = \sum_{j=1}^{L} h_{ij} X_j + Z_i \quad \text{for} \quad i = 1, 2, \ldots, L,$$

where $Z_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the independent background noise, and h_{ij} denotes the complex channel from the jth user to the ith BS. In this chapter, we assume that the user scheduling is fixed, and perfect CSI is available to all the BSs and to the centralized processor. Further, it is assumed that each user transmits at a fixed power, i.e., X_i's are complex-valued Gaussian signals with $\mathbb{E}\left[|X_i|^2\right] = P_i$, for $i = 1, \ldots, L$.

This chapter uses a compress-and-forward scheme in which the BSs quantize the received signals $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_L]^T$ into $\hat{\mathbf{Y}} = [\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_L]^T$ using either Wyner-Ziv coding or single-user compression and transmit the compressed bits to the CP through noiseless fronthaul links. A two-stage successive decoding strategy is employed, where the CP first recovers the quantized signals $\hat{\mathbf{Y}}$, and then decodes user messages $\mathbf{X} = [X_1, X_2, \ldots, X_L]^T$ based on the quantized signals $\hat{\mathbf{Y}}$. The successive decoding nature of the proposed scheme overcomes the delay and high computational complexity associated with joint decoding (e.g., [13,14]). Let $q_i = \mathbb{E}(\hat{Y}_i - Y_i)^2$ be the average squared-error distortion between Y_i and $\hat{Y}_i$. In this chapter, the distortion level q_i is referred to as the quantization noise level.

3.2.2 The VMAC-WZ Scheme

Because of the mutual interference between the neighboring users, the received signals at the different BSs are statistically correlated. Consequently, Wyner-Ziv compression can be used to achieve higher compression efficiency and to better utilize the limited fronthaul capacities than per-link single-user compression.

Proposition 3.1 *For the uplink C-RAN model with fronthaul sum capacity constraint C as shown in Fig. 3.1, the rate tuples $(R_1, R_2, \ldots, R_L)$ that satisfy the following set of constraints are achievable using the VMAC-WZ scheme:*

$$\sum_{i \in S} R_i \leq \log \frac{\left|\mathbf{H}_{\mathcal{S}} \mathbf{K}_{\mathcal{S}} \mathbf{H}_{\mathcal{S}}^{\dagger} + \mathbf{\Lambda}_q + \operatorname{diag}(\sigma_i^2)\right|}{|\mathbf{\Lambda}_q + \operatorname{diag}(\sigma_i^2)|} \tag{3.1}$$

such that

$$\log \frac{\left|\mathbf{H}_{\mathcal{L}} \mathbf{K}_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \mathbf{\Lambda}_q + \operatorname{diag}(\sigma_i^2)\right|}{|\mathbf{\Lambda}_q|} \leq C \tag{3.2}$$

for all $\mathcal{S} \subseteq \{1, 2, \ldots, L\}$, where $\mathbf{K}_{\mathcal{S}} = \mathbb{E}\left[\mathbf{X}_{\mathcal{S}} \mathbf{X}_{\mathcal{S}}^{\dagger}\right]$ is the covariance matrix of $\mathbf{X}_{\mathcal{S}}$, $\mathbf{\Lambda}_q = \operatorname{diag}(q_1, q_2, \ldots, q_L)$ is the covariance matrix of the quantization noise, and $\mathbf{H}_{\mathcal{S}}$ denotes the channel matrix from $\mathbf{X}_{\mathcal{S}}$ to $\mathbf{Y}$.

Proof. This theorem is a generalization of [12, Theorem 1], which treats the case of a single transmitter with multiple relays under individual fronthaul capacity constraints. In [12, Theorem 1], it has been

www.manaraa.com

shown that $R < I(\mathbf{X}; \hat{\mathbf{Y}})$ is achievable subject to

$$I(\mathbf{Y}_{\mathcal{S}}; \hat{\mathbf{Y}}_{\mathcal{S}} | \hat{\mathbf{Y}}_{\mathcal{S}^c}) \leq \sum_{i \in \mathcal{S}} C_i, \quad \forall \mathcal{S} \subseteq \{1, 2, \ldots, L\} \tag{3.3}$$

under a product distribution $p(\hat{\mathbf{y}}|\mathbf{y}) = \Pi_{i=1}^{L} p(\hat{y}_i|y_i)$. Note that under the sum fronthaul constraint $\sum_{i=1}^{L} C_i \leq C$, the constraint (3.3) simply becomes $I(\mathbf{Y}; \hat{\mathbf{Y}}) \leq C$. Now, with multiple users and considering the sum rate over any subset $\mathcal{S}$, we likewise have

$$\sum_{i \in \mathcal{S}} R_i \leq I(\mathbf{X}_{\mathcal{S}}; \hat{\mathbf{Y}} | \mathbf{X}_{\mathcal{S}^c}), \qquad \forall \mathcal{S} \subseteq \{1, 2, \ldots, L\} \tag{3.4}$$

subject to

$$I(\mathbf{Y}; \hat{\mathbf{Y}}) \leq C. \tag{3.5}$$

Let $p(\hat{y}_i|y_i)$ be defined by the test channel $\hat{Y}_i = Y_i + Q_i$, where $Q_i \sim \mathcal{CN}(0, q_i)$ is the quantization noise independent of everything else, and q_i is the quantization noise level. The achievable rate region (3.1) subject to (3.2) can now be derived by evaluating the mutual information expressions (3.4) and (3.5) assuming complex Gaussian distribution for X_i. □

3.2.3 The VMAC-SU Scheme

Although Wyner-Ziv coding represents a better utilization of the fronthaul, it is also complex to implement in practice. In this section, Wyner-Ziv coding is replaced by single-user compression. We derive the achievable rate region when the compression process does not take advantage of the statistical correlations between the received signals at different BSs. In this case, each BS simply quantizes its received signals using a vector quantizer.

Proposition 3.2 *For the uplink C-RAN model with L BSs and sum fronthaul capacity C shown in Fig. 3.1, the following rate tuple $(R_1, R_2, \ldots, R_L)$ is achievable using the VMAC-SU scheme:*

$$\sum_{i \in \mathcal{S}} R_i \leq \log \frac{\left| \mathbf{H}_{\mathcal{S}} \mathbf{K}_{\mathcal{S}} \mathbf{H}_{\mathcal{S}}^{\dagger} + \mathbf{\Lambda}_q + \mathrm{diag}(\sigma_i^2) \right|}{|\mathbf{\Lambda}_q + \mathrm{diag}(\sigma_i^2)|} \tag{3.6}$$

such that

$$\log \frac{\left| \mathrm{diag}(\mathbf{H}_{\mathcal{L}} \mathbf{K}_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger}) + \mathbf{\Lambda}_q + \mathrm{diag}(\sigma_i^2) \right|}{|\mathbf{\Lambda}_q|} \leq C \tag{3.7}$$

for all $\mathcal{S} \subseteq \{1, 2, \ldots, L\}$, where $\mathbf{K}_{\mathcal{S}} = \mathbb{E}\left[\mathbf{X}_{\mathcal{S}} \mathbf{X}_{\mathcal{S}}^{\dagger}\right]$ is the covariance matrix of $\mathbf{X}_{\mathcal{S}}$, $\mathbf{\Lambda}_q = \mathrm{diag}(q_1, \ldots, q_L)$ is the covariance matrix of the quantization noise, and $\mathbf{H}_{\mathcal{S}}$ denotes the channel matrix from $\mathbf{X}_{\mathcal{S}}$ to $\mathbf{Y}$.

Proposition 3.2 is a straightforward extension of Proposition 3.1, where the rate expression (3.6) is given by the achievable sum rate $I(\mathbf{X}_{\mathcal{S}}; \hat{\mathbf{Y}})$ and the constraint (3.7) follows from the fronthaul constraint $\sum_{i=1}^{L} I(Y_i; \hat{Y}_i) \leq C$. The rate expression implicitly assumes the successive decoding of the quantization codewords first, then the transmitted signals.

By comparing the expressions (3.2) with (3.7), it is not hard to find that the quantization noise levels supported by VMAC-WZ is smaller than that supported by VMAC-SU under the same fronthaul constraint. This is because that Wyner-Ziv coding is more efficient than single-user compression, which

www.manaraa.com

results lower compression distortion (which is corresponding to smaller quantization noise level) under the same compression rate (i.e., fronthaul capacity). Therefore, one can conclude that the VMAC-WZ scheme always achieves better performance than the VMAC-SU scheme. However, as it is shown in the later section, the gain obtained by Wyner-Ziv coding over the single-compression vanishes as the fronthaul capacity increases.

3.3 Quantization Noise Level Optimization for VMAC-WZ

The achievable rate regions for the VMAC schemes have an intuitive interpretation. The quantization process adds quantization noise to the overall multiple-access channel. Finer quantization results in higher overall rate, but also leads to higher fronthaul capacity requirements. To characterize the tradeoff between the achievable rate and the fronthaul constraint, this section formulates a weighted sum rate maximization problem over the quantization noise levels $\{q_1, \ldots, q_L\}$ under a sum fronthaul capacity constraint for VMAC-WZ.

3.3.1 Problem Formulation

Let μ_i be the weights representing the priorities associated with the mobile users typically determined from upper layer protocols. Without loss of generality, let $\mu_L \geq \mu_{L-1} \geq \cdots \geq \mu_1 \geq 0$. The boundary of the achievable rate region for VMAC-WZ can be attained using a successive decoding approach with a decoding order from user 1 to L. A weighted rate sum maximization problem that characterizes the VMAC-WZ achievable rate region can be written as:

$$\begin{aligned} \max_{\mathbf{\Lambda}_q} \quad & \sum_{i=1}^{L} \mu_i \log \frac{\left|\sum_{j=i}^{L} P_j \mathbf{h}_j \mathbf{h}_j^\dagger + \mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|}{\left|\sum_{j>i}^{L} P_j \mathbf{h}_j \mathbf{h}_j^\dagger + \mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|} \\ \text{s.t.} \quad & \log \frac{\left|\sum_{j=1}^{L} P_j \mathbf{h}_j \mathbf{h}_j^\dagger + \mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|}{|\mathbf{\Lambda}_q|} \leq C, \\ & \mathbf{\Lambda}_q(i,j) = 0, \quad \text{for} \quad i \neq j, \\ & \mathbf{\Lambda}_q(i,i) \geq 0, \end{aligned} \tag{3.8}$$

where $\mathbf{\Lambda}_q(i,j)$ is the (i,j)th entry of matrix $\mathbf{\Lambda}_q$, and the optimization is over the quantization noise levels $\mathbf{\Lambda}_q = \mathrm{diag}(q_i)$.

The objective function of (3.8) is a convex function of $\mathbf{\Lambda}_q$ (instead of concave). Consequently, finding the global optimum solution of (3.8) is challenging. In [19], an algorithm based on the gradient projection method together with a bisection search on the dual variable is proposed for a related problem, where the quantization noise levels are optimized one after another in a coordinated fashion. The above problem formulation is also related to that in [20] where the quantization noise levels at the BSs are optimized for sum-rate maximization on a per-BS basis. The advantage of the present formulation is that the quantization noise levels across the BSs are optimized jointly, resulting in better overall performance.

www.manaraa.com

3.3.2 Alternating Convex Optimization Approach

This section proposes an alternating convex optimization (ACO) scheme capable of arriving at a stationary point of the problem (3.8). The key observation is that the objective function of (3.8) is a difference of two concave functions. The idea is to linearize the second concave function to obtain a concave lower bound of the original objective function, then successively approximate the optimal solution by optimizing this lower bound. The ACO scheme is closely related to the block successive minimization method [63] or minorize-maximization algorithm [64], which can be used to solve a broad class of optimization problems with nonconvex objective functions over a convex set. These optimization techniques have also been previously applied for solving related problems in wireless communications; see [65, 66].

Before presenting the proposed algorithm, we first state the following lemma, which is a direct consequence of Fenchel's inequality for concave functions.

Lemma 3.1 *For positive definite Hermitian matrices* $\mathbf{\Omega}, \mathbf{\Gamma} \in \mathbb{C}^{L\times L}$,

$$\log|\mathbf{\Omega}| \le \log|\mathbf{\Gamma}| + \mathrm{Tr}\left(\mathbf{\Gamma}^{-1}\mathbf{\Omega}\right) - L \tag{3.9}$$

with equality if and only if $\mathbf{\Omega} = \mathbf{\Gamma}$.

Applying Lemma 3.1, we reformulate problem (3.8) as a double maximization problem:

$$\begin{aligned}
\max_{\mathbf{\Lambda}_q, \mathbf{\Gamma}\succeq\mathbf{0}} \quad & \sum_{i=1}^{L}(\mu_i - \mu_{i-1}) \log\left|\sum_{j=i}^{L} P_j\mathbf{h}_j\mathbf{h}_j^\dagger + \mathrm{diag}(\mathbf{\Gamma}_i^2) + \mathbf{\Lambda}_q\right| \\
& -\mu_L\left(\log|\mathbf{\Gamma}| + \mathrm{Tr}\left(\mathbf{\Gamma}^{-1}(\mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q)\right)\right) \\
\text{s.t.} \quad & \log\frac{\left|\sum_{i=1}^{L} P_i\mathbf{h}_i\mathbf{h}_i^\dagger + \mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|}{|\mathbf{\Lambda}_q|} \le C \\
& \mathbf{\Lambda}_q(i,j) = 0, \quad \text{for} \quad i \ne j, \\
& \mathbf{\Lambda}_q(i,i) \ge 0,
\end{aligned} \tag{3.10}$$

where $\mu_L \ge \mu_{L-1} \ge \cdots \ge \mu_1 > \mu_0 = 0$.

Although the maximization problem (3.10) is still nonconvex with respect to $(\mathbf{\Lambda}_q, \mathbf{\Gamma})$, the advantage of the reformulation is that fixing either $\mathbf{\Lambda}_q$ or $\mathbf{\Gamma}$, problem (3.10) is a convex optimization with respect to the other variable. This coordinate-wise convexity property enables us to use an iterative coordinate ascent algorithm. Specifically, when $\mathbf{\Lambda}_q$ is fixed, we solve

$$\min_{\mathbf{\Gamma}\succeq\mathbf{0}} \quad \log|\mathbf{\Gamma}| + \mathrm{Tr}\left(\mathbf{\Gamma}^{-1}(\mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q)\right). \tag{3.11}$$

Following Lemma 3.1, problem (3.11) has the following closed-form solution:

$$\mathbf{\Gamma}^* = \mathrm{diag}(\sigma_i) + \mathbf{\Lambda}_q. \tag{3.12}$$

www.manaraa.com

If $\mathbf{\Gamma}$ is fixed, problem (3.10) becomes

$$\begin{aligned}\max_{\mathbf{\Lambda}_q} \quad & \sum_{i=1}^{L}(\mu_i-\mu_{i-1})\log\left|\sum_{j=i}^{L}P_j\mathbf{h}_j\mathbf{h}_j^\dagger+\mathrm{diag}(\sigma_i^2)+\mathbf{\Lambda}_q\right| \\ & -\mu_L\mathrm{Tr}\left(\mathbf{\Gamma}^{-1}(\mathrm{diag}(\sigma_i^2)+\mathbf{\Lambda}_q)\right) \\ \text{s.t.} \quad & \log\frac{\left|\sum_{i=1}^{L}P_i\mathbf{h}_i\mathbf{h}_i^\dagger+\mathrm{diag}(\sigma_i^2)+\mathbf{\Lambda}_q\right|}{|\mathbf{\Lambda}_q|}\le C, \\ & \mathbf{\Lambda}_q(i,j)=0, \quad \text{for} \quad i\neq j, \\ & \mathbf{\Lambda}_q(i,i)\ge 0. \end{aligned} \tag{3.13}$$

It is easy to verify that the above problem is a convex optimization problem, as the objective function is now concave with respected to $\mathbf{\Lambda}_q$. So, it can be solved efficiently with polynomial complexity. We summarize the ACO algorithm below:

Algorithm 3.1 Alternating Convex Optimization (ACO)

1: Initialize $\mathbf{\Lambda}_q^{(0)}=\mathbf{\Gamma}^{(0)}=\gamma I$.
2: **repeat**
3: Fix $\mathbf{\Gamma}=\mathbf{\Gamma}^{(i)}$, solve the convex optimization problem (3.13) over $\mathbf{\Lambda}_q$. Set $\mathbf{\Lambda}_q^{(i+1)}$ to be the optimal point.
4: Update $\mathbf{\Gamma}^{(i+1)}=\mathrm{diag}(\sigma_i^2)+\mathbf{\Lambda}_q^{(i+1)}$.
5: **until** convergence

The ACO algorithm yields a nondecreasing sequence of objective values for problem (3.10). So the algorithm is guaranteed to converge. Moreover, it converges to a stationary point of the optimization problem.

Theorem 3.1 *From any initial point $(\mathbf{\Lambda}_q^{(0)},\mathbf{\Gamma}^{(0)})$, the limit point $(\mathbf{\Lambda}_q^*,\mathbf{\Gamma}^*)$ generated by the alternating convex optimization algorithm is a stationary point of the weighted sum-rate maximization problem (3.8).*

The proof of Theorem 3.1 is similar to that of [65, Proposition 1] and is also closely related to the convergence proof of successive convex approximation algorithm [66]. First, based on a result on block coordinate descent [67, Corollary 2], it can be shown that the ACO algorithm converges to a stationary point of the double maximization problem (3.10). Now, suppose that $(\mathbf{\Lambda}_q^*,\mathbf{\Gamma}^*)$ is a stationary point of (3.10), we have

$$\mathrm{Tr}\left(\nabla_{\mathbf{\Lambda}_q}F\left(\mathbf{\Lambda}_q^*,\mathbf{\Gamma}^*\right)^\dagger,(\mathbf{\Lambda}_q-\mathbf{\Lambda}_q^*)\right)\le 0,\forall\,\mathbf{\Lambda}_q\in\mathcal{W}, \tag{3.14}$$

where $F\left(\mathbf{\Lambda}_q,\mathbf{\Gamma}\right)$ denotes the objective function of (3.10). Using the same argument as the proof of [65, Proposition 1], we can substitute $\mathbf{\Gamma}^*=\mathrm{diag}(\sigma_i)+\mathbf{\Lambda}_q^*$ into (3.14) and verify that $\mathbf{\Lambda}_q^*$ is also a stationary point of (3.8).

We mention here that although the ACO algorithm is stated here for the SISO case, it is equally applicable to the MIMO case, where the BSs are equipped with multiple antennas, and the optimization is over quantization covariance matrices. In the following, we highlight the advantage of our approach as compared to that of [19, 20].

In [19], a gradient projection method together with a bisection search on the dual variable is used to solve the weighted sum-rate maximization for a related problem. Although the gradient projection

www.manaraa.com

approach also converges to a stationary point of the problem, it is slower than the proposed ACO algorithm. This is because the algorithm of [19] relies on per-BS block coordinate gradient descent, which has sublinear convergence [68], rather than joint optimization across all the BSs. The gradient-type approach used in [19] is also typically much slower than optimization techniques which use second-order Hessian information (e.g. Newton's method) that can be applied to convex problems. In [20], the optimization of the quantization noise covariance matrices for sum-rate maximization is solved on a per-BS basis in a greedy fashion, one BS at a time. This approach in general does not converge to a local optimal solution, (as has already been pointed out in [20]). It cannot be applied to the weighted sum-rate maximization problem considered in this chapter. In contrast, the ACO algorithm presented here is capable of solving the optimal quantization noise covariance matrices across all the BSs jointly, and the convergence to the stationary point is guaranteed.

3.3.3 Optimal Quantization Noise Level at High SQNR

Although locally optimal quantization noise level can be effectively found using the proposed ACO algorithm for any fixed user schedule, user priority, and channel condition, the implementation of ACO in practical systems can be computationally intensive, especially in a fast-fading environment or when the scheduled users in the time-frequency slots change frequently. In this section, we aim to understand the structure of the optimal solution by deriving the optimal quantization noise level in the high SQNR regime. The main result of this section is that setting the quantization noise level to be proportional to the background noise level is approximately optimal for maximizing the overall sum rate. This leads to an efficient way for setting the quantization noise levels in practice.

Consider the sum-rate maximization problem:

$$\begin{aligned} \max \quad & \log \frac{\left|\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \operatorname{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|}{\left|\operatorname{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|} \\ \text{s.t.} \quad & \log \frac{\left|\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \operatorname{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|}{\left|\mathbf{\Lambda}_q\right|} \le C \\ & \mathbf{\Lambda}_q(i,j) = 0, \quad \text{for} \quad i \neq j \\ & \mathbf{\Lambda}_q(i,i) \ge 0. \end{aligned} \tag{3.15}$$

This optimization problem is nonconvex, but its Karush-Kuhn-Tucker (KKT) condition still gives a necessary condition for optimality. To derive the KKT condition, form the Lagrangian

$$\begin{aligned} L(\mathbf{\Lambda}_q, \lambda, \mathbf{\Psi}) = (1-\lambda) \log \left|\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \operatorname{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right| \\ - \log \left|\operatorname{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right| + \lambda \log \left|\mathbf{\Lambda}_q\right| + \operatorname{Tr}(\mathbf{\Psi}\mathbf{\Lambda}_q) \end{aligned} \tag{3.16}$$

where $\mathbf{\Psi}$ is a matrix whose diagonal entries are zeros and the off-diagonal entries are the dual variables associated the constraint $\mathbf{\Lambda}_q(i,j) = 0$ for $i \neq j$, and λ is the Lagrangian dual variable associated with the fronthaul sum-capacity constraint.

Setting $\partial L / \partial \mathbf{\Lambda}_q$ to zero, we obtain the optimality condition

$$(1-\lambda)(\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \operatorname{diag}(\sigma_i^2) + \mathbf{\Lambda}_q)^{-1} - (\operatorname{diag}(\sigma_i^2) + \mathbf{\Lambda}_q)^{-1} + \lambda \mathbf{\Lambda}_q^{-1} + \mathbf{\Psi} = 0 \tag{3.17}$$

www.manaraa.com

Recall that $\mathbf{\Psi}$ has zeros on the diagonal, but can have arbitrary off-diagonal entries. Thus, the above optimality condition can be simplified as

$$(1-\lambda)\text{diag}(\mathbf{H}_{\mathcal{L}}K_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}+\text{diag}(\sigma_i^2)+\mathbf{\Lambda}_q)^{-1}-(\text{diag}(\sigma_i^2)+\mathbf{\Lambda}_q)^{-1}+\lambda\mathbf{\Lambda}_q^{-1}=0 \tag{3.18}$$

First, it is easy to verify that the optimality condition can only be satisfied if $0 \leq \lambda < 1$. Second, since $\mathbf{\Lambda}_q + \text{diag}(\sigma_i^2)$ is the combined quantization and background noise, if the overall system is to operate at reasonably high spectral efficiency, we must have[2] $\text{diag}(\mathbf{H}_{\mathcal{L}}K_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}) \gg \text{diag}(\sigma_i^2)+\mathbf{\Lambda}_q$. Under this high SQNR condition, we have

$$\text{diag}(\mathbf{H}_{\mathcal{L}}K_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}+\text{diag}(\sigma_i)+\mathbf{\Lambda}_q)^{-1} \ll (\text{diag}(\sigma_i^2)+\mathbf{\Lambda}_q)^{-1}$$

in which case the optimality condition becomes

$$q_i \approx \frac{\lambda}{1-\lambda}\sigma_i^2 \tag{3.19}$$

where $\lambda \in [0,1)$ is chosen to satisfy the fronthaul sum-capacity constraint. Thus we see that under high SQNR, the optimal quantization noise level should be proportional to the background noise level. Note that $\lambda = 0$ corresponds to the infinite fronthaul capacity case where $q_i = 0$. As λ increases, the sum fronthaul capacity becomes increasingly constrained, and the optimal quantization noise level q_i also increases accordingly.

3.3.4 Sum Capacity to Within a Constant Gap

We now further justify the setting of the quantization noise level to be proportional to the background noise level by showing that this choice in fact achieves the sum capacity of the uplink C-RAN model with sum fronthaul capacity constraint to within a constant gap. The gap depends on the number of BSs in the network but is independent of the channel matrix and the signal-to-noise ratios (SNRs).

Theorem 3.2 *For the uplink C-RAN model with a sum fronthaul capacity C as shown in Fig. 3.1, the VMAC-WZ scheme with the quantization noise levels set to be proportional to the background noise levels achieves a sum capacity to within one bit per BS per channel use.*

Proof. See Appendix E. □

The proof of above theorem depends on a comparison of achievable rate with a cut-set outer bound. The basic idea is to set the quantization noise levels to be at the background noise levels if C is large, (specifically, $C \geq \log \frac{\left|\mathbf{H}\mathbf{K}_{\mathcal{L}}\mathbf{H}^{\dagger}+2\text{diag}(\sigma_i^2)\right|}{\left|\text{diag}(\sigma_i^2)\right|}$ as in the proof), resulting in at most 1 bit gap per channel use per BS. When C is small, scaling the quantization noise level by a constant turns out to maintain the constant-gap optimality.

This result is reminiscent of the more general constant-gap result for arbitrary multicast relay network [15,16], but this result is both more specific, as it only applies to the sum-capacity constrained fronthaul case, and also more practically useful, as it assumes successive decoding of quantization codeword first then user messages, rather than joint decoding.

[2]Here, "$\gg$" denotes component-wise comparison on the diagonal entries.

www.manaraa.com

A similar constant-gap result can be obtained in the case where both transmitters and receivers are equipped with multiple antennas. For example, considering the scenario where K users with M transmit antennas each send independent messages to L BSs with N receive antennas each. It can be shown that the constant gap for sum capacity is $\min\{KM, NL\}$ bits per channel use. In particular, when $K = NL$, i.e., when the degree of freedom in the system is fully utilized, the constant-gap result becomes one bit per BS antenna per channel use.

3.3.5 Efficient Algorithm for Setting Quantization Noise Level

The main observation in the previous section is that setting the quantization noise levels at different BSs to be proportional to the background noise levels is near sum-rate optimal under high SQNR and from a constant-gap-to-capacity perspective. This holds regardless of the transmit power, the channel matrix, and the user schedule, which is especially advantageous for practical implementation as no adaptation to the channel condition is needed.

In the following, we propose a simple algorithm for setting the quantization noise level to be $q_i = \alpha\sigma_i^2$ for some appropriate α. Note that with this setting of q_i, the fronthaul constraint becomes:

$$C_{WZ}(\alpha) \triangleq \log \frac{\left|\sum_{j=1}^{L} P_j \mathbf{h}_j \mathbf{h}_j^{\dagger} + (1+\alpha)\mathrm{diag}(\sigma_i^2)\right|}{|\alpha \mathrm{diag}(\sigma_i^2)|} \leq C. \tag{3.20}$$

Since the fronthaul constraint should be satisfied with equality and since $C_{WZ}(\alpha)$ is monotonic in α, a simple bisection search can be used to find the suitable α. The algorithm is summarized below as Algorithm 3.2. As simulation results later in the chapter show, Algorithm 3.2 performs very close to the optimized scheme (Algorithm 3.1) for practical channel scenarios.

Algorithm 3.2 Approximate Algorithm for VMAC-WZ

1: Set $\alpha = 1$.
2: **while** $C_{WZ}(\alpha) > C$ **do**
3: Set $\alpha = 2\alpha$.
4: **end while**
5: Set $\alpha_{\max} = \alpha$ and $\alpha_{\min} = 0$.
6: Use bisection in $[\alpha_{\min}, \alpha_{\max}]$ to solve $C_{WZ}(\alpha) = C$.
7: Set $q_i = \alpha\sigma_i$.

3.4 Optimal Fronthaul Allocation for VMAC-SU

3.4.1 Problem Formulation

We now turn to the VMAC-SU scheme and consider the weighted sum-rate maximization problem under a sum fronthaul constraint for the more practical single-user compression scheme. The optimization

www.manaraa.com

problem can be stated as follows:

$$\begin{aligned}
\max_{\mathbf{\Lambda}_q} \quad & \sum_{i=1}^{L} \mu_i \log \frac{\left|\sum_{j=i}^{L} P_j \mathbf{h}_j \mathbf{h}_j^\dagger + \mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|}{\left|\sum_{j>i}^{L} P_j \mathbf{h}_j \mathbf{h}_j^\dagger + \mathrm{diag}(\sigma_i^2) + \mathbf{\Lambda}_q\right|} \\
\text{s.t.} \quad & \sum_{i=1}^{L} \log\left(1 + \frac{\sum_{j=1}^{L} P_j |h_{ij}|^2 + \sigma_i^2}{q_i}\right) \leq C \\
& \mathbf{\Lambda}_q(i,j) = 0, \quad \text{for} \quad i \neq j, \\
& \mathbf{\Lambda}_q(i,i) \geq 0.
\end{aligned} \tag{3.21}$$

As mentioned earlier, the objective function in the above is convex in q_i (instead of concave), which is not easy to maximize. But the ACO algorithm proposed earlier can still be used here to find locally optimal q_i's. However for VMAC-SU, because the compression at each BS is independent, it is possible to re-parameterize the problem in term of the rates allocated to the fronthaul links. It is instructive to work with such a reformulation in order to obtain system design insight. Introduce the new variables

$$C_i = \log\left(1 + \frac{\sum_{j=1}^{L} P_j |h_{ij}|^2 + \sigma_i^2}{q_i}\right). \tag{3.22}$$

Let γ_i be the combined quantization and background noise, i.e., $\gamma_i = \sigma_i^2 + q_i$. Then,

$$\gamma_i = \frac{\sum_{j=1}^{L} P_j |h_{ij}|^2 + \sigma_i^2 2^{C_i}}{2^{C_i} - 1}. \tag{3.23}$$

Further, define $\mathbf{\Upsilon} = \mathrm{diag}(1/\gamma_i)$. By a variable substitution, it is straightforward to establish that the optimization problem (3.21) is equivalent to the following:

$$\begin{aligned}
\max \quad & \sum_{i=1}^{L} (\mu_i - \mu_{i-1}) \log \left|\mathbf{\Upsilon} \sum_{j=i}^{L} P_j \mathbf{h}_j \mathbf{h}_j^\dagger + \mathbf{I}\right| \\
\text{s.t.} \quad & \sum_{i=1}^{L} C_i \leq C, \\
& C_i \geq 0, \quad i = 1, \ldots, L,
\end{aligned} \tag{3.24}$$

where, without loss of generality, it has been assumed $\mu_L \geq \cdots \geq \mu_1 > \mu_0 = 0$. The above problem is easier to solve than (3.21), because the feasible set of the problem is a polyhedron with only linear constraints. For example, it is possible to dualize with respect to the sum fronthaul constraint, then numerically find a local optimum of the Lagrangian. A bisection on the dual variable can then be used in an outer loop to solve (3.24).

3.4.2 Optimal Quantization Noise Level at High SQNR

For the VMAC-WZ scheme under high SQNR assumption, setting $q_i = \alpha\sigma_i^2$ is approximately optimal for maximizing the overall sum rate. This section establishes a similar result for the VMAC-SU case. We first introduce Lagrange multipliers $\nu_i \geq 0$ for the positivity constraints $C_i \geq 0$, and $\beta \geq 0$ for the

www.manaraa.com

fronthaul sum-capacity constraint $\sum_{i=1}^{L} C_i \leq C$, we obtain the following KKT condition

$$\mathrm{Tr}\left[\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} \left(\boldsymbol{\Upsilon} \mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \mathbf{I}\right)^{-1} \frac{\partial \boldsymbol{\Upsilon}}{\partial C_i}\right] - \beta + \nu_i = 0. \tag{3.25}$$

Note that γ_i is the combined quantization and background noise. So, under the high SQNR assumption, where SNR $\gg 1$ and $C_i \gg 1$, we must have $\mathrm{diag}(\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger}) \gg \mathrm{diag}\,(\gamma_i)$. Thus

$$\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} \left(\boldsymbol{\Upsilon} \mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \mathbf{I}\right)^{-1} \approx \boldsymbol{\Upsilon}^{-1}. \tag{3.26}$$

After some manipulations, the optimality condition now becomes

$$\frac{\sum_{j=1}^{L} P_j |h_{ij}|^2 + \sigma_i^2}{\sum_{j=1}^{L} P_j |h_{ij}|^2 + \sigma_i^2 2^{C_i}} - \beta + \nu_i \approx 0 \tag{3.27}$$

where we also use the approximation $2^{C_i} - 1 \approx 2^{C_i}$. Note that $\nu_i = 0$ whenever $C_i > 0$. Solving (3.27) together with $\sum_{i=1}^{L} C_i = C$ yields the following approximately optimal fronthaul rate allocation:

$$C_i \approx \log\left(\frac{1-\beta}{\beta} \overline{\mathsf{SNR}}_i + \frac{1}{\beta}\right) \tag{3.28}$$

where $\overline{\mathsf{SNR}}_i = (\sum_{j=1}^{L} P_j |h_{ij}|^2)/\sigma_i^2$ and β is chosen such that $\sum_{i=1}^{L} C_i = C$. The corresponding quantization noise level is given by

$$q_i \approx \frac{\beta}{1-\beta} \sigma_i^2. \tag{3.29}$$

We point out here that the same result can also be derived from the KKT condition of (3.21).

The above result shows that setting the quantization noise level to be proportional to the background noise level is near optimal for maximizing the sum rate for VMAC-SU at high SQNR. This is similar to the VMAC-WZ case. Intuitively, in the VMAC schemes the intercell interference is completely nulled by multicell decoding. The achievable sum rate is only limited by the combined quantization noise and background noise. Thus, it is reasonable that the optimal quantization noise levels only depend on the background noise levels.

3.4.3 Sum Capacity of Diagonally Dominant Channels

This section provides further justification for choosing the quantization noise level to be proportional to the background noise level by showing that doing so achieves the sum capacity of the VMAC model to within a constant gap when the received signal covariance matrix satisfies a diagonally dominant channel criterion. The received signal covariance matrix is defined as $\mathsf{E}[\mathbf{Y}\mathbf{Y}^{\dagger}] = \mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \mathrm{diag}(\sigma_i^2)$. It is often diagonally dominant, because the path losses from the remote users to the BSs are distance dependent, and typically each user is associated with its strongest BS. In the following, we define a diagonally dominant condition for matrices, and state a constant-gap result for sum capacity for the VMAC-SU scheme under a sum fronthaul constraint.

Definition 3.4.1. For a fixed constant $\kappa > 1$, a $n \times n$ matrix $\boldsymbol{\Psi}$ is said to be κ-strictly diagonally

www.manaraa.com

dominant if

$$|\mathbf{\Psi}(i,i)| \geq \kappa \sum_{j \neq i}^{n} |\mathbf{\Psi}(i,j)| \quad \text{for all} \quad i = 1, \ldots, n,$$

where $\mathbf{\Psi}(i,j)$ is the (i,j)-th entry of matrix $\mathbf{\Psi}$.

Theorem 3.3 *For the uplink C-RAN model with a sum fronthaul capacity C as shown in Fig. 3.1, if the received covariance matrix $\mathbf{H}_{\mathcal{L}} K_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \mathrm{diag}(\sigma_i^2)$ is κ-strictly diagonally dominant for $\kappa > 1$, then the VMAC-SU scheme achieves the sum capacity of the uplink CRAN model to within $\left(1 + \log \frac{\kappa}{\kappa - 1}\right)$ bits per BS per channel use.*

Proof. See Appendix F. □

We note that the above result can be further strengthened when C is large. In this case, setting the quantization noise levels to be at the background noise levels results in at most 1 bit gap per channel use per BS to sum capacity. It is not hard to further verify that, in this case, the VMAC-SU scheme is actually approximately optimal for the entire capacity region of the uplink C-RAN model. Analogous to Wyner-Ziv compression, a similar constant-gap result for single-user compression can also be obtained in the case where both users and BSs are equipped with multiple antennas.

3.4.4 Fronthaul Allocation for Heterogeneous Networks

The fact that setting the quantization noise levels to be proportional to the background noise levels is approximately optimal gives rise to an efficient algorithm for allocating capacities across the fronthaul links. This section describes an approach similar to the corresponding algorithm for the VMAC-WZ case. In addition, we further generalize to the case of heterogeneous network with multiple tiers of BSs.

Consider a multi-tier heterogeneous network consisting of not only macro BSs, but also pico-BSs, coordinated together in a C-RAN architecture. The macro- and pico-BSs typically have very different fronthaul capacities, so they may be subject to different fronthaul constraints. Let C_m be the sum fronthaul capacity constraint across the macro-BSs, and C_p be the fronthaul constraint for pico-BSs. Assuming a VMAC-SU implementation, the fronthaul constraints can be expressed as:

$$\sum_{i \in \mathcal{S}_m} \log \left(1 + \frac{\sum_{j=1}^{L} P_j |h_{ij}|^2 + \sigma_i^2}{q_i}\right) \leq C_m \tag{3.30}$$

$$\sum_{i \in \mathcal{S}_p} \log \left(1 + \frac{\sum_{j=1}^{L} P_j |h_{ij}|^2 + \sigma_i^2}{q_i}\right) \leq C_p \tag{3.31}$$

where $\mathcal{S}_m$ and $\mathcal{S}_m$ are the sets of macro- and pico-BSs, respectively.

It can be shown that for multi-tier networks, it is also near optimal to set the quantization noise levels to be proportional to the background noise levels under high SQNR. However, different tiers may have different proportionality constants. Since the quantization noise level (or equivalently the fronthaul capacity) for each BS may be set independently without affecting other BSs for VMAC-SU, a simple bisection algorithm can be used to optimize the quantization noise level (or equivalently the fronthaul capacity) in each tier independently.

Let

$$C_{SU}(\beta) = \sum_{i \in \mathcal{S}} \log \left(\frac{1 - \beta}{\beta} \overline{\mathsf{SNR}}_i + \frac{1}{\beta}\right) \tag{3.32}$$

www.manaraa.com

be the sum fronthaul capacity across a particular tier (where $\mathcal{S}$ can be $\mathcal{S}_m$ for macro-BSs or $\mathcal{S}_p$ for pico-BSs). The bisection algorithm described in Algorithm 3.3 can run simultaneously in each tier.

Algorithm 3.3 Approximate Algorithm for VMAC-SU

1: Set $\beta_{\min} = 0$, $\beta_{\max} = 1$.
2: Use bisection in $[\beta_{\min}, \beta_{\max}]$ to solve $C_{SU}(\beta) = C$.
3: Set $q_i = \frac{\beta}{1-\beta}\sigma_i^2$, and $C_i = \log\left(\frac{1-\beta}{\beta}\overline{\mathsf{SNR}}_i + \frac{1}{\beta}\right)$.

We point out here that practical heterogeneous network may have other types of fronthaul structure. For instance, in practical implementation the pico-BSs may not have direct fronthaul links to the CP, but may connect to the macro-BSs first then to the CP. In this case, the fronthaul constraints can be formulated as

$$\begin{cases} \sum\limits_{i\in\mathcal{S}_m} \log\left(1+\frac{\sum_{j=1}^{L} P_j|h_{ij}|^2+\sigma_i^2}{q_i}\right) \le \tilde{C}_m \\ \sum\limits_{i\in\mathcal{S}_p} \log\left(1+\frac{\sum_{j=1}^{L} P_j|h_{ij}|^2+\sigma_i^2}{q_i}\right) \le \tilde{C}_p \\ \tilde{C}_m + \tilde{C}_p \le C, \quad \tilde{C}_p \le C_p \end{cases} \tag{3.33}$$

where the optimization variables are $\{q_i\}$, $\tilde{C}_m$ and $\tilde{C}_p$. Here C_p is the sum-capacity constraint for the fronthaul links connecting pico-BSs to the macro-BSs, and C is the total sum fronthaul constraint for both pico-BSs and macro-BSs. In this case, the fronthaul constraints for macro-BSs and pico-BSs are coupled together. However, Algorithm 3.3 can be still be helpful in finding the approximately optimal quantization noise levels. Specifically, for each fixed pair of $\tilde{C}_m$ and $\tilde{C}_p$, Algorithm 3.3 can be used to find the q_i's for the macro-BSs and the pico-BSs respectively. The problem is now simplified to finding the optimal partition of C between $\tilde{C}_m$ and $\tilde{C}_p$.

3.5 Simulations

3.5.1 Multicell Network

In this section, the performances of the VMAC-WZ and VMAC-SU schemes with different quantization noise level optimization strategies are evaluated in a wireless cellular network setup with 19 cells wrapped around, 3 sectors per cell, and 20 users randomly located in each sector. The central 7 BSs (i.e., 21 sectors) form a C-RAN cooperation cluster, where each BS is connected to the CP with noiseless fronthaul link with a sum capacity constraint across the 7 BSs. The users are associated with the sector with the strongest channel. Round-robin user scheduling is used on a per-sector basis. Perfect channel estimation is assumed, and the CSI is made available to all BSs and to the CP. In the simulation, fixed transmit power of 23dBm is used at all the mobile users. Various algorithms are run on fixed set of channels. Detailed system parameters are outlined in Table 3.5.1.

In the simulation, weighted rate-sum maximization is performed over the quantization noise levels, with weights equal to the reciprocal of the exponentially updated long-term average rate. In the implementation of VMAC schemes, successive interference cancelation (SIC) decoding is used at the CP. The decoding order of the users is determined by their weights, i.e., the user with high weight is decoded last. The baseline system is the conventional cellular networks without joint multicell processing at the CP. Cumulative distribution function (CDF) of the user rates is plotted in order to visualize the performance

www.manaraa.com

Table 3.1: Multicell Network System Parameters

Cellular Layout	Hexagonal, 19-cell, 3 sectors/cell
BS-to-BS Distance	500 m
Frequency Reuse	1
Channel Bandwidth	10 MHz
Number of Users per Sector	20
Total Number of Users	420
User Transmit Power	23 dBm
Antenna Gain	14 dBi
SNR Gap (with coding)	6 dB
Background Noise	-169 dBm/Hz
Noise Figure	7 dB
Tx/Rx Antenna No.	1
Distance-dependent Path Loss	$128.1 + 37.6 \log_{10}(d)$
Log-normal Shadowing	8 dB standard deviation
Shadow Fading Correlation	0.5
Cluster Size	7 cells (21 sectors)
Scheduling Strategy	Round-robin

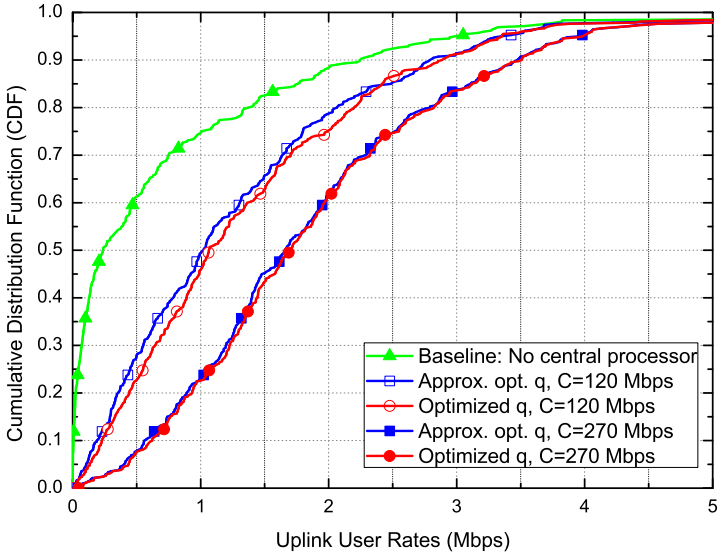


Figure 3.2: Cumulative distribution of user rates with the VMAC-WZ scheme

of various schemes.

Fig. 3.2 compares the performance of the baseline system with the VMAC-WZ scheme under the sum fronthaul capacities of 120Mbps per macro-cell (40Mbps per sector) and 270Mbps per cell (90Mbps per sector). The baseline system implements local decoding at the BSs without fronthaul capacity limit. The VMAC-WZ scheme is implemented with two choices of quantization noise levels: the approximately optimal q_i proportional to the background noise level as given by Algorithm 3.2 (labeled as "appro. opt. q") and the optimal q_i given by Algorithm 3.1 (labeled as "optimized q"). It is shown that the VMAC-WZ schemes significantly outperform the baseline system. The figure also shows that setting q_i to be proportional to the background noise level is indeed approximately optimal, especially when C is large. This confirms our earlier theoretical analysis on the approximately optimal q_i.

The VMAC schemes considered in this chapter is superior to the per-BS SIC scheme considered

www.manaraa.com

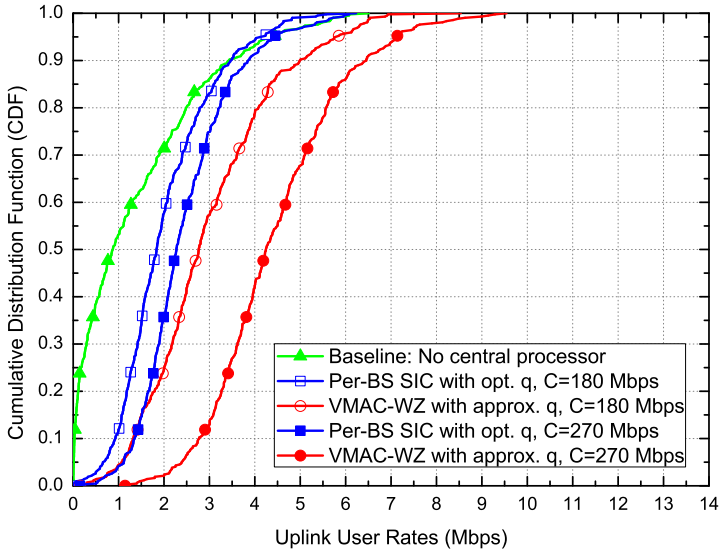


Figure 3.3: Performance comparison of the VMAC-WZ scheme with the per-BS interference cancellation scheme.

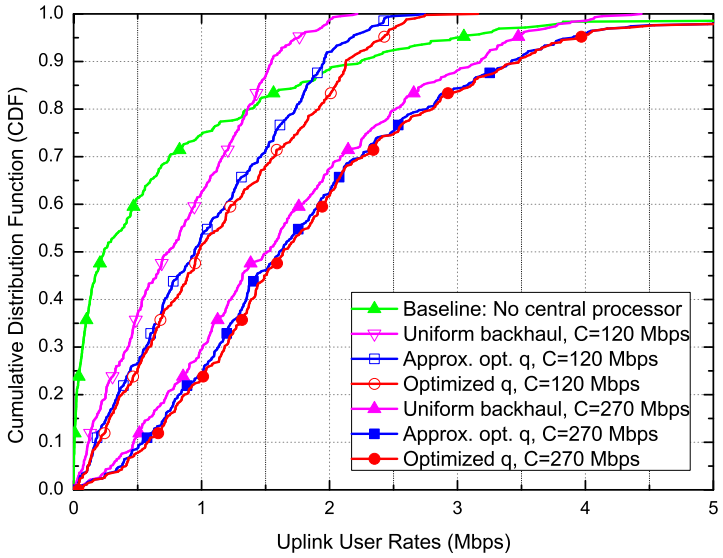


Figure 3.4: Cumulative distribution of user rates with the VMAC-SU scheme

in [34]. To illustrate this point, Fig. 3.3 compares the performance of the VMAC-WZ scheme under the approximately optimal q_i with the per-BS SIC scheme of [34] (labeled as "Per-BS SIC"). For fair comparison, we run the simulation over the users in the 7-cell cluster only, and ignore the out-of-cluster interference, which is the case considered in [34]. The figure shows that significant gain can be obtained by the VMAC-WZ scheme over the per-BS successive cancellation scheme.

Fig. 3.4 shows the CDF curves of user rates for the VMAC-SU scheme with three choices of quantization noise levels: the quantization noise levels given by allocating the fronthaul capacity equally across the BSs (labeled as "uniform fronthaul"), the approximately optimal q_i proportional to the background noise as given by Algorithm 3.3 (labeled as "approx. opt. q"), and the optimal q_i derived from the

www.manaraa.com

fronthaul capacity allocation formulation of the problem (labeled as "optimized q"). It can be seen that VMAC with single-user compression also significantly improves the performance of baseline system and that the approximately optimal q_i is near optimal, especially when C is large. The figure also shows that allocating fronthaul capacity uniformly across the BSs is strictly suboptimal.

To further compare the performance of the VMAC-SU scheme with different choices of quantization noise levels, Fig. 3.5 plots the average per-cell sum rate of the baseline and the VMAC-SU schemes as a function of the fronthaul capacity. The figure clearly shows the advantage of optimizing the quantization noise levels (or equivalently the allocation of fronthaul capacities). For example, to achieve 80Mbps per-cell sum rate, we need 200Mbps sum fronthaul if fronthaul capacities are allocated uniformly, 170Mbps sum fronthaul if q_i is chosen to be proportional to the background noise, and 150Mbps sum fronthaul if q_i is optimized. Thus, the optimization of the quantization noise level can save up to 25% in fronthaul capacity.

Further, it can be seen from Fig. 3.5 that under infinite sum fronthaul, the achieved per-cell sum rate saturates and approaches about 115Mbps for this cellular setting. But when the quantization noise level is optimized, a finite sum fronthaul capacity at about 200Mbps is already sufficient to achieve about 100Mbps user sum rate, which is 90% of the full benefit of uplink network MIMO. Note that the performance gap between the approximately optimal q_i and the optimal q_i becomes smaller as the sum fronthaul capacity increases, confirming the approximate optimality of $q_i = \alpha\sigma_i^2$ in the high SQNR regime.

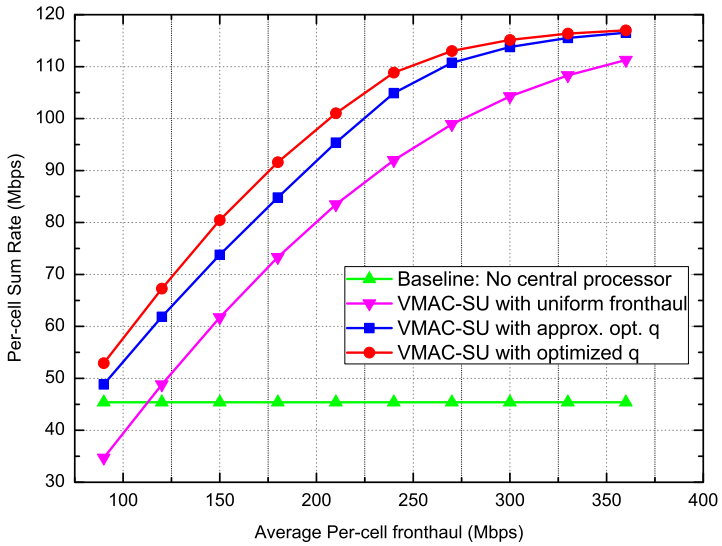


Figure 3.5: Per-cell sum rate vs. average per-cell fronthaul capacity of the VMAC-SU scheme.

Fig. 3.6 compares the performance of Wyner-Ziv coding and single-user compression for the VMAC scheme. It is observed that Wyner-Ziv coding is superior to single-use compression. However, as the sum fronthaul capacity becomes larger, the gain due to Wyner-Ziv coding diminishes.

3.5.2 Multi-Tier Heterogeneous Network

The performance of the VMAC-SU scheme is further evaluated for a two-tier heterogeneous network with 7 macro-cells wrapped around, 3 sectors per cell, 3 pico BSs randomly located in each sector, and

www.manaraa.com

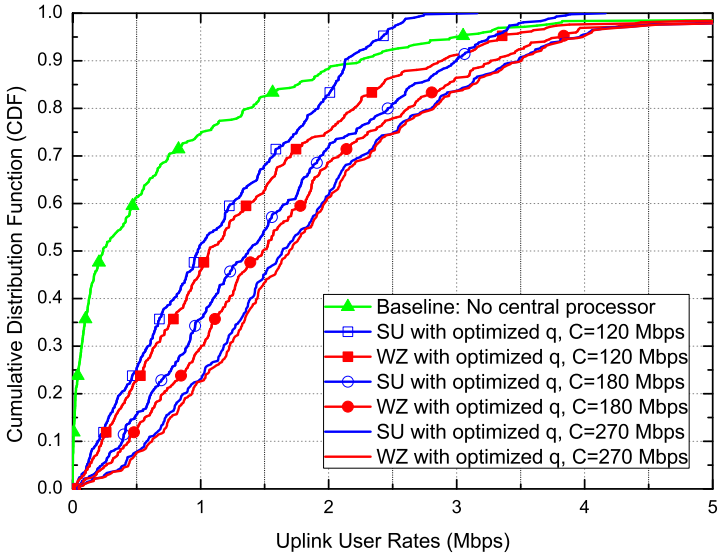


Figure 3.6: Comparison of the VMAC-SU and VMAC-WZ schemes

Table 3.2: Heterogeneous Network Channel Parameters

Cellular Layout	Hexagonal, wrapped around
BS-to-BS Distance	500 m
Number of Macro Cells	7 cells, 3 sectors/cell
Number of Pico Cells	3 pico cells per macro sector
Frequency Reuse	1
Channel Bandwidth	10 MHz
Number of Users per Macro Sector	20
User Transmit Power	23 dBm
Antenna Gain	14 dBi
SNR Gap (with coding)	6 dB
Background Noise	-169 dBm/Hz
Noise Figure	7 dB
Pico BS Antenna Pattern	Omni-directional
Tx/Rx Antenna No.	1
Path Loss Macro to User	$128.1 + 37.6 \log_{10}(d)$
Path Loss Pico to User	$140.7 + 36.7 \log_{10}(d)$
Log-normal Shadowing	8 dB standard deviation for macro-user link; 4 dB for pico-user link
Shadow Fading Correlation	0.5
Cluster Size	1 macro cell and 9 pico cells
Min. Dist. between BSs	75 m
Scheduling Strategy	Round-robin

20 mobile users per macro-cell sector. The cellular topology is shown in Fig. 3.7. Each user establishes connection with the macro/pico BS with the highest received SNR. Note that the number of users in each pico/macro-cell is not fixed. On average there are 8 users per macro-cell sector and 4 users per pico-cell. In this network, every macro-cell forms a C-RAN cluster, consisting of 3 macro-sectors and 9 pico-cells. The macro BSs and pico BSs are subject to different sum fronthaul capacity constraints. Specifically, the sum fronthaul capacity is set to be 189Mbps for the 3 macro-BSs and 81Mbps for the 9

www.manaraa.com

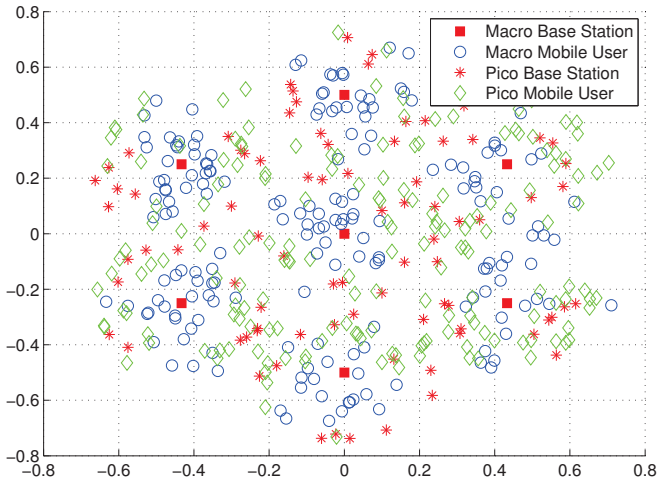


Figure 3.7: A picocell network topology with 7 cells, 3 sectors per cell, and 3 pico base-stations per sector placed randomly.

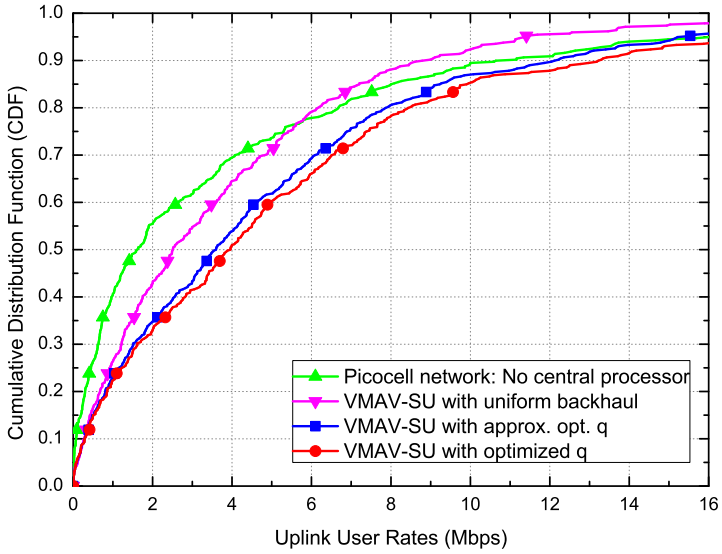


Figure 3.8: Cumulative distribution of user rates in the picocell network where the 3 macro-BSs and 9 pico-BSs within each 3-sector macrocell form a cluster. The VMAC-SU scheme is applied and the sum fronthaul constraints for macro and pico BSs are 189Mpbs and 81Mpbs per cluster, respectively.

pico BSs. Perfect CSI is made available to all the BSs and to the CP. System parameters are outlined in Table 3.2.

Fig. 3.8 shows the CDF plots of user rates achieved by the baseline scheme and the VMAC-SU scheme. It is clear that the C-RAN architecture significantly improves upon the baseline, more than doubling the 50-percentile rate. The optimization of the quantization noise level is important, as a naive uniform fronthaul allocation only achieves half of the potential gain for C-RAN. Finally, setting the quantization noise level to be proportional to the background noise is indeed approximately optimal. In this multi-tier heterogeneous network case, the proportionality constant is set independently for each tier using Algorithm 3.3.

www.manaraa.com

3.6 Summary

This chapter studies an uplink C-RAN model where the BSs within a cooperation cluster are connected to a cloud-computing based CP through noiseless fronthaul links of limited sum capacity. We employ two VMAC schemes where the BSs use either Wyner-Ziv compression or single-user compression to quantize the received signals and send the compressed bits to the CP. At the CP, quantization codewords are first decoded; subsequently the user messages are decoded as if the users form a virtual multiple-access channel.

The main findings of the chapter are concerned with efficient optimization of the quantization noise levels for both VMAC-WZ and VMAC-SU. We propose an alternating optimization algorithm for VMAC-WZ and a fronthaul capacity allocation formulation for VMAC-SU. More importantly, it is observed that setting the quantization noise levels to be proportional to the background noise levels is approximately optimal. This leads to efficient algorithms for optimizing the quantization noise levels, or equivalently, for allocating the fronthaul capacities.

From an analytic point of view, this chapter shows that setting quantization noise levels to be proportional to the background noise levels is near optimal for maximizing the sum rate when the system operates in the high SQNR regime. With such a choice of quantization noise levels, the VMAC-WZ scheme can achieve the sum capacity of the uplink C-RAN model to within a constant gap. A similar constant-gap result is also obtained for VMAC-SU under a diagonally dominant channel condition. From a numerical perspective, simulation results confirm that the proposed VMAC schemes can significantly improve the performance of wireless cellular systems. The improvement is maximized with optimized quantization noise levels or equivalently optimized fronthaul capacity allocations. The near optimal choice of quantization noise levels indeed performs very close to the optimal one over the SQNR region of practical interest.

www.manaraa.com

Chapter 4

Joint Beamforming and Compression for Uplink MIMO C-RAN

4.1 Introduction

In the previous chapter, we study the optimal fronthaul compression for the VMAC scheme in SISO uplink C-RAN under a sum fronthaul constraint. In this chapter, we further consider the problem of transmit beamforming and fronthaul design for the MIMO uplink C-RAN with individual fronthaul capacity constraints. The MIMO uplink C-RAN architecture is repeated Fig. 4.1 for convenience of discussion, where multi-antenna mobile users communicate with the CP with multi-antenna BSs serving as relay nodes. The BSs are connected with the CP via digital fronthaul links with finite capacities. We consider the VMAC scheme, in which the BSs quantize the received signals using either single-user compression or Wyner-Ziv coding and send the compressed bits to the CP. The CP performs successive decoding to decode the quantization codewords first, then the user messages sequentially. Under the VMAC scheme, this chapter studies the optimization of the transmit beamforming vectors and quantization noise covariance matrices for maximizing the weighted sum rate of the C-RAN system. Being different from the conventional multicell cellular systems, in which the optimal transmit beamforming only depends on the interfering signal strength and the channel gain matrices, in C-RAN, the finite fronthaul capacity also needs to be taken into account in the beamforming design. This chapter proposes a novel weighted minimum-mean-square-error successive convex approximation (WMMSE-SCA) algorithm to find the optimal transmit beamformers and quantization noise covariance matrices for maximizing the weighted sum rate of C-RAN. Moreover, a simple separate design consisting of optimizing transmit beamformers for the Gaussian vector multiple-access channel and per-antenna scalar quantizers with uniform quantization noise levels across the antennas at each BS is also developed, under the assumption that the signal-to-quantization-noise ratio (SQNR) is high and successive interference cancellation (SIC) is applied at the receiver. Numerical simulations show that the proposed separate design already performs very close to the optimized joint design in the SQNR regime of practical interest.

This chapter considers two different fronthaul compression strategies for C-RAN, namely *single-user*

www.manaraa.com

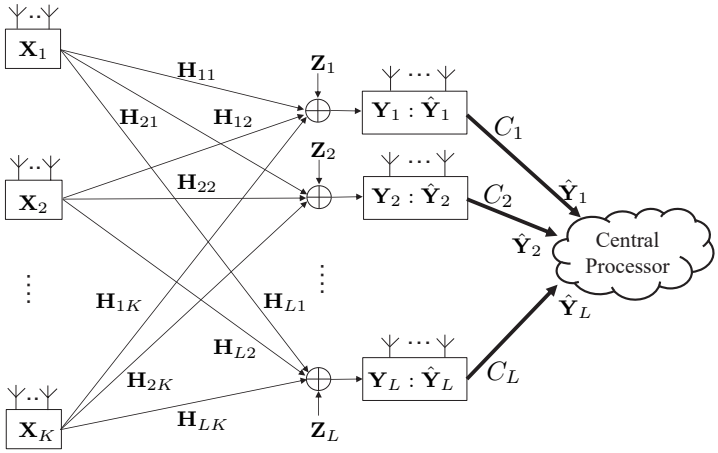


Figure 4.1: An uplink MIMO C-RAN system with capacity-limited fronthaul

compression and *Wyner-Ziv coding*. In single-user compression, which is also referred to as point-to-point compression in the literature [58], each BS uses vector quantization to compress the received signals but ignores the correlation between the received signals across different BSs. In contrast, Wyner-Ziv coding fully utilizes the correlation of the received signals for higher compression efficiency, thereby achieving better overall performance. The optimization strategy proposed in this chapter is first developed for single-user compression, then for the more complex Wyner-Ziv coding, assuming a heuristic ordering for decompression of the quantized signals at the BSs. The performance of the VMAC schemes with single-user compression and Wyner-Ziv coding are evaluated for practical multicell networks under linear *minimum-mean-square-error (MMSE) receiver* and *SIC receiver* respectively. It is shown that the implementation of SIC receiver significantly improves the performance achieved by linear MMSE receiver under both single-user compression and Wyner-Ziv coding. Furthermore, although single-user compression with SIC receiver can already realize majority of the benefit brought by the C-RAN architecture, Wyner-Ziv coding can further improve upon single-user compression when the fronthaul capacity is limited.

To precisely quantify the advantage of the C-RAN architecture, this chapter further evaluates the performance of optimized beamforming and fronthaul compression under two different types of BS clustering strategies: *disjoint clustering* and *user-centric clustering*. In disjoint clustering scheme, the entire network is divided into non-overlapping clusters and the BSs in each cluster jointly serve all the users within the coverage area [69]. In user-centric clustering, each user is served by an individually selected subset of neighboring BSs; different clusters for different users may overlap. The performance of user-centric clustering has been evaluated for the downlink of cooperative cellular networks [70] and C-RAN systems [71]. This chapter further shows numerically that in uplink C-RAN, with optimized beamforming and fronthaul compression, the user-centric clustering strategy significantly outperforms the disjoint clustering strategy, because the cell edges are effectively eliminated.

4.1.1 Related Work

One of the main issues in the implementation of C-RAN is how to optimally utilize the capacity-limited fronthaul links to efficiently reap the benefit of multicell processing. Substantial research works have made progress towards this direction [12, 13, 34, 39]. The compress-and-forward relaying scheme for the

www.manaraa.com

uplink C-RAN model is related to the noisy network coding scheme in information theory literature [15, 16], but noisy network coding involves high-complexity joint decoding at the decoder. In [38], a virtual multiple access channel (VMAC) scheme, which is a compress-and-forward strategy based on successive decoding, is proposed for the single-input single-output (SISO) C-RAN architecture. As compared to the noisy network coding scheme, the VMAC scheme has lower decoding complexity and shorter decoding delay, which makes it more desirable for practical implementation. Furthermore, it is shown in [35] that with Wyner-Ziv coding the successive decoding based VMAC scheme actually achieves the same maximum sum rate as noisy network coding for the uplink C-RAN model under a sum fronthaul constraint.

This chapter studies the linear transceiver and fronthaul compression design in the VMAC scheme for the uplink multiple-input-multiple-output (MIMO) C-RAN model. As a generalization of [38] which considers the SISO case only, this chapter considers the MIMO case where both the users and the BSs are equipped with multiple antennas. The main difference between the SISO case and the MIMO case is the impact of transmitter optimization at the user terminals. In the SISO case, since most of the intra-cluster interference is already nulled by multicell decoding, it is near optimal for the users to transmit at their maximum powers. In the MIMO case, the users are capable of doing transmit beamforming, so the optimal transmit beamforming design is more involved.

The fronthaul compression problem for the uplink C-RAN model has been considered extensively in the literature. Various algorithms such as alternating convex optimization [38], gradient projection [19], and the robust fronthaul compression approach [20] have been developed for maximizing the (weighted) sum rate under the fronthaul constraints. All of these algorithms focus only on the optimization of quantization noise covariance matrices across the BSs, with fixed transmit beamformers. This chapter goes one step further by considering the joint transmit beamformer and quantization noise covariance matrix optimization problem. Accounting for both the transmit beamforming and the quantization design problem together in the optimization framework is nontrivial, because the two are coupled through the fronthaul constraints. To tackle this problem, this chapter proposes a novel WMMSE-SCA algorithm for efficiently finding a local optimum solution to the weighted sum rate maximization problem. The proposed algorithm integrates the well-known WMMSE beamforming design strategy [72, 73], with the successive convex approximation technique [63, 64], to arrive at a stationary point of the maximization problem. The performance of optimized beamforing vectors and quantization noise covariance matrices for both Wyner-Ziv coding and single-user compression are evaluated under practical multicell networks with different receive beamforming schemes, i.e., the linear receiver and the SIC receiver. Simulation results show that the performance improvement of the SIC receiver as compared to the linear receiver is much larger than that of Wyner-Ziv coding as compared to single-user compression. Most of the performance gain brought by C-RAN can thus be obtained by single-user compression together with SIC receiver.

4.1.2 Chapter Organization

The rest of the chapter is organized as follows. Section 4.2 introduces the system model and the VMAC scheme. Section 4.3 considers the joint design of beamforming and fronthaul compression under single-user compression, where a novel WMMSE based successive convex optimization algorithm is proposed. The proposed joint design scheme is developed further in Section 4.4 for maximizing weighted sum rate under Wyner-Ziv coding. Section 4.5 is devoted to a low-complexity separate design, which is shown to

www.manaraa.com

be near-optimal at high SQNR regime. The proposed optimization algorithms are evaluated numerically for practical multicell and multicluster networks in Section 4.6 and concluding remarks are made in in Section 4.7.

4.2 Preliminaries

4.2.1 System Model

This chapter considers the uplink C-RAN, where K multi-antenna mobile users communicate with a CP through L multi-antenna BSs serving as relay nodes, as shown in Fig. 4.1. The noiseless fronthaul links connecting the BSs with the CP have per-link capacities C_ℓ. Each user terminal is equipped with M antennas; each BS is equipped with N antennas. Furthermore, it is assumed that perfect channel state information (CSI) is made available to all the BSs and to the CP.

We consider the VMAC scheme [38] applied for such a C-RAN system, in which the BSs quantize the received signals using either Wyner-Ziv coding or single-user compression, then forwards the compressed bits to the CP for decoding. In single-user compression, the compression process only involves the conventional vector quantizers, one for each BS, while in Wyner-Ziv coding, the correlation between the received signals across the BSs are fully utilized for higher compression efficiency. At the CP side, a two-stage successive decoding strategy is employed, where the quantization codewords are first decoded, and then the user messages are decoded sequentially.

Define $\mathbf{H}_{\ell,k}$ as the $N \times M$ complex channel matrix between the kth user and the ℓth BS. It is assumed that each user intends to transmit d parallel data streams to the CP. Let $\mathbf{V}_k \in \mathbb{C}^{M\times d}$ denote the linear transmit beamfomer that user k utilizes to transmit message signal $\mathbf{s}_k \in \mathbb{C}^{d\times 1}$ to the central receiver. We assume that each message signal $\mathbf{s}_k$ intended for user k is taken from a Gaussian codebook so that we have $\mathbf{s}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. Then the transmit signal at user k is given by $\mathbf{x}_k = \mathbf{V}_k\mathbf{s}_k$ with $\mathbb{E}[\mathbf{x}_k\mathbf{x}_k^\dagger] = \mathbf{V}_k\mathbf{V}_k^\dagger$. The transmit beamformers are subjected to per-user power constraints, i.e., $\mathrm{Tr}\left(\mathbf{V}_k\mathbf{V}_k^\dagger\right) \le P_k$ for $k \in \mathcal{K}$. The received signal at BS ℓ, $\mathbf{y}_\ell$, can be expressed as

$$\mathbf{y}_\ell = \sum_{k=1}^{K} \mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{s}_k + \mathbf{z}_\ell, \quad \forall\, \ell \in \mathcal{L}, \tag{4.1}$$

where $\mathbf{z}_\ell \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_\ell)$ represents the additive Gaussian noise for BS ℓ. Assuming Gaussian quantization test channel, the quantized received signal $\hat{\mathbf{y}}_\ell$ for the ℓth BS is given by

$$\hat{\mathbf{y}}_\ell = \mathbf{y}_\ell + \mathbf{q}_\ell \tag{4.2}$$

where $\mathbf{q}_\ell \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_\ell)$ represents the Gaussian quantization noise for the ℓth BS.

4.2.2 Achievable Rate of the VMAC scheme

The rate region of the VMAC scheme is characterized by that of a multiple-access channel, in which multiple users send information to a common CP. Following the results in [38], assuming that the linear MMSE receiver is applied at the CP, the transmission rate R_k for user k for the VMAC scheme is given by

$$R_k \le I(\mathbf{X}_k; \mathbf{Y}_1, \ldots, \mathbf{Y}_L) = \log\left|\mathbf{I} + \mathbf{V}_k^\dagger\mathbf{H}_{\mathcal{L},k}^\dagger\mathbf{D}_k^{-1}\mathbf{H}_{\mathcal{L},k}\mathbf{V}_k\right| \tag{4.3}$$

www.manaraa.com

where

$$\mathbf{D}_k = \mathbf{D}_k^{\mathrm{LE}} = \sum_{j\neq k}^{K} \mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^{\dagger}\mathbf{H}_{\mathcal{L},j}^{\dagger} + \mathbf{\Sigma} + \mathbf{Q}, \tag{4.4}$$

with $\mathbf{\Sigma} = \mathrm{diag}\left(\{\mathbf{\Sigma}_\ell\}_{\ell=1}^{L}\right)$ and $\mathbf{Q} = \mathrm{diag}\left(\{\mathbf{Q}_\ell\}_{\ell=1}^{L}\right)$. To achieve higher throughput, the SIC scheme can also be applied combined with the MMSE receiver. For simplicity, we here assume a decoding ordering of user messages $1, 2, \ldots, K$, while the case of an arbitrary decoding order of user messages can be obtained by simple analogy. The matrix $\mathbf{D}_k = \mathbf{D}_k^{\mathrm{LE}}$ in (4.3) is replaced by $\mathbf{D}_k^{\mathrm{SIC}}$ expressed as

$$\mathbf{D}_k = \mathbf{D}_k^{\mathrm{SIC}} = \sum_{j>k}^{K} \mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^{\dagger}\mathbf{H}_{\mathcal{L},j}^{\dagger} + \mathbf{\Sigma} + \mathbf{Q}. \tag{4.5}$$

The compression rates at the BSs should also satisfy the fronthaul link capacity constraints. Using information-theoretic formulation, the fronthaul constraints under single-user compression can be written as

$$I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell\right) \leq C_\ell, \ \forall\, \ell \in \mathcal{L}. \tag{4.6}$$

Evaluating the above mutual information expression with Gaussian input and Gaussian quantization noise, the fronthaul constraint (4.6) becomes [36]

$$\log \frac{\left|\sum_{k=1}^{K} \mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^{\dagger}\mathbf{H}_{\ell,k}^{\dagger} + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\right|}{|\mathbf{Q}_\ell|} \leq C_\ell, \tag{4.7}$$

for all $\ell = 1, 2, \ldots, L$. When Wyner-Ziv coding is implemented at BSs, the fronthaul constraints are given by the following mutual information expressions [35, 37]

$$I\left(\mathbf{Y}_{\mathcal{S}}; \hat{\mathbf{Y}}_{\mathcal{S}} | \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) \leq \sum_{\ell\in\mathcal{S}} C_\ell, \quad \forall\, \mathcal{S} \subseteq \mathcal{L}. \tag{4.8}$$

Utilizing the chain rule on mutual information and the Gaussian assumption, one can express the fronthaul constraint (4.8) for Wyner-Ziv coding as follows,

$$\log \frac{\left|\sum\limits_{k=1}^{K} \mathbf{H}_{\mathcal{L},k}\mathbf{V}_k\mathbf{V}_k^{\dagger}\mathbf{H}_{\mathcal{L},k}^{\dagger} + \mathrm{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell\in\mathcal{L}}\right)\right|}{\left|\sum\limits_{k=1}^{K} \mathbf{H}_{\mathcal{S}^c,k}\mathbf{V}_k\mathbf{V}_k^{\dagger}\mathbf{H}_{\mathcal{S}^c,k}^{\dagger} + \mathrm{diag}\left(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell\in\mathcal{S}^c}\right)\right|} - \sum_{\ell\in\mathcal{S}} \log|\mathbf{Q}_\ell| \leq \sum_{\ell\in\mathcal{S}} C_\ell, \quad \forall\, \mathcal{S} \subseteq \mathcal{L}. \tag{4.9}$$

4.3 Joint Beamforming and Compression Design under Single-User Compression

4.3.1 Weighted Sum Rate Maximization

This section investigates the joint beamforming and fronthaul compression design for the VMAC scheme with single-user compression. As shown in the achievable rate expression (4.3) and the fronthaul constraint expression (4.7), the beamforming vectors and quantization noise covariance matrices are coupled, and the two together determine the overall performance of a C-RAN system. To characterize the tradeoff between the achievable rates for the users and the system resources, we formulate the following weighted

www.manaraa.com

sum rate maximization problem:

$$\begin{aligned}
\max_{\mathbf{V}_k, \mathbf{Q}_\ell} \quad & \sum_{k=1}^{K} \mu_k \log \left| \mathbf{I} + \mathbf{V}_k^\dagger \mathbf{H}_{\mathcal{L},k}^\dagger \mathbf{D}_k^{-1} \mathbf{H}_{\mathcal{L},k} \mathbf{V}_k \right| && \text{(4.10a)} \\
\text{s.t.} \quad & \mathbf{D}_k = \mathbf{D}_k^{\text{LE}} \quad \text{or} \quad \mathbf{D}_k = \mathbf{D}_k^{\text{SIC}}, \\
& \log \frac{\left| \sum_{k=1}^{K} \mathbf{H}_{\ell,k} \mathbf{V}_k \mathbf{V}_k^\dagger \mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell \right|}{|\mathbf{Q}_\ell|} \leq C_\ell, \quad \forall \ell \in \mathcal{L}, && \text{(4.10b)} \\
& \mathbf{Q}_\ell \succeq \mathbf{0}, \quad \forall \ell \in \mathcal{L}, \\
& \text{Tr}\left(\mathbf{V}_k \mathbf{V}_k^\dagger \right) \leq P_k, \quad \forall k \in \mathcal{K}
\end{aligned}$$

where μ_k's are the weights representing the priorities associated with the mobile users typically determined from upper layer protocols. When SIC is implemented, to maximize the weighed sum rate, the user with larger weight should be decoded last. Without loss of generality, we assume $0 \leq \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$, which results in the decoding order of user messages $1, 2, \ldots, K$.

Due to the non-convexity of both the objective function and the fronthaul capacity constraints in problem (4.10), finding the global optimum solution of (4.10) is challenging. We point out here that the present formulation (4.10) actually implicitly includes the user scheduling strategy. More specifically, one can consider a weighted sum rate maximization problem over all the users in the network, where the beamformers for the users are set to be the zero vector if they are not scheduled. For simplicity in the following development, we focus on the active uses only and assume that the user scheduling is done prior to solving problem (4.10). Implicit scheduling is discussed later in the simulation part of the chapter.

4.3.2 The WMMSE-SCA Algorithm

In this section, we propose a novel algorithm to find a stationary point of the problem (4.10). The main difficulty in solving (4.10) comes from the fact that the objective function and fronthaul capacity constraints are both nonconvex functions with respect to the optimization variables. Inspired by the recent work of using the WMMSE approach for beamforming design [72, 73], we first reformulate the objective function in problem (4.10) as a convex function with respect to the MMSE matrix given by the user's target signal s_k and decoded signal $\hat{s}_k$. We then linearize the convex objective function and the compression rate expressions in the fronthaul constraints of (4.10) to obtain a convex approximation of the original problem. Finally we successively approximate the optimal solution by optimizing this convex approximation. The idea of convex approximation is rooted from modern optimization techniques including block successive minimization method and minorize-maximization algorithm, which have been previously applied for solving related problems in wireless communications [70, 74].

By applying Lemma 3.1 to the first log-determinant term in the fronthaul constraint expression (4.7) or (4.10b) and by setting

$$\mathbf{\Omega}_\ell = \sum_{k=1}^{K} \mathbf{H}_{\ell,k} \mathbf{V}_k \mathbf{V}_k^\dagger \mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell, \tag{4.11}$$

www.manaraa.com

we can approximate the fronthaul constraint (4.7) or (4.10b) with the following convex constraint:

$$\log|\mathbf{\Gamma}_\ell| + \mathrm{Tr}\left(\mathbf{\Gamma}_\ell^{-1}\left(\sum_{k=1}^{K}\mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\right)\right) - \log|\mathbf{Q}_\ell| \le C_\ell + N \tag{4.12}$$

for $\ell = 1, 2, \ldots, L$. It is not hard to see that the fronthaul constraint (4.7) or (4.10b) is always feasible when the convex constraint (4.12) is feasible. The two constraints are equivalent when

$$\mathbf{\Gamma}_\ell^* = \sum_{k=1}^{K}\mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell. \tag{4.13}$$

Now we approximate the objective function (4.10a) using the WMMSE approximation. Let $\mathbf{U}_k \in \mathbb{C}^{NL\times d}$ be the linear receiver applied at the CP for recovering $\mathbf{s}_k$. The transmission rate R_k in (4.3) can be expressed as the following [72] [73],

$$R_k = \max_{\mathbf{U}_k} \log|\mathbf{E}_k^{-1}| \tag{4.14}$$

where

$$\mathbf{E}_k = (\mathbf{I} - \mathbf{U}_k^\dagger\mathbf{H}_{\mathcal{L},k}\mathbf{V}_k)(\mathbf{I} - \mathbf{U}_k^\dagger\mathbf{H}_{\mathcal{L},k}\mathbf{V}_k)^\dagger + \mathbf{U}_k^\dagger\left(\sum_{j\neq k}^{K}\mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^\dagger\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}\right)\mathbf{U}_k. \tag{4.15}$$

By applying Lemma 3.1 again, we rewrite rate expression (4.14) as

$$R_k = \max_{\mathbf{W}_k,\mathbf{U}_k}\left(\log|\mathbf{W}_k| - \mathrm{Tr}(\mathbf{W}_k\mathbf{E}_k) + d\right) \tag{4.16}$$

where $\mathbf{W}_k$ is the weight matrix introduced by the WMMSE method. The optimal $\mathbf{W}_k$ is given by

$$\mathbf{W}_k^* = \mathbf{E}_k^{-1} = \mathbf{I} + \mathbf{H}_{\mathcal{L},k}^\dagger\mathbf{V}_k^\dagger\mathbf{U}_k^*, \tag{4.17}$$

where $\mathbf{U}_k^*$ is the MMSE receive beamformer given by

$$\mathbf{U}_k^* = \left(\sum_{j\neq k}^{K}\mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^\dagger\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}\right)^{-1}\mathbf{H}_{\mathcal{L},k}\mathbf{V}_k. \tag{4.18}$$

Using (4.16) and (4.12) to replace the objective function and the fronthaul constraints in problem (4.10), we reformulate the weighted sum-rate maximization problem as follows

$$\begin{aligned}
\max_{\substack{\mathbf{V}_k,\mathbf{Q}_\ell,\mathbf{U}_k,\\ \mathbf{W}_k,\mathbf{\Gamma}_\ell}} \quad & \sum_{k=1}^{K}\mu_k\left(\log|\mathbf{W}_k| - \mathrm{Tr}(\mathbf{W}_k\mathbf{E}_k)\right) + \rho\sum_{\ell=1}^{L}\|\mathbf{\Gamma}_\ell - \mathbf{\Omega}_\ell\|_F^2 \\
\text{s.t.} \quad & \log|\mathbf{\Gamma}_\ell| + \mathrm{Tr}\left(\mathbf{\Gamma}_\ell^{-1}\mathbf{\Omega}_\ell\right) - \log|\mathbf{Q}_\ell| \le C_\ell', \ \forall\ell\in\mathcal{L}, \\
& \mathbf{Q}_\ell \succeq \mathbf{0}, \quad \forall\ell\in\mathcal{L}, \\
& \mathrm{Tr}\left(\mathbf{V}_k\mathbf{V}_k^\dagger\right) \le P_k, \quad \forall k\in\mathcal{K},
\end{aligned} \tag{4.19}$$

where $\mathbf{\Omega}_\ell = \sum_{k=1}^{K}\mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell$, ρ is some positive constant, and $C_\ell' = C_\ell + N$. Note that

www.manaraa.com

Algorithm 4.1 WMMSE-SCA Algorithm

1: Initialize $\mathbf{Q}_\ell$ and $\mathbf{V}_k$ such that $\mathrm{Tr}\left(\mathbf{V}_k\mathbf{V}_k^\dagger\right) = P_k$.
2: **repeat**
3: $\quad\mathbf{\Gamma}_\ell \leftarrow \sum_{k=1}^K \mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell$.
4: $\quad\mathbf{U}_k \leftarrow \left(\sum_{j\neq k}\mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^\dagger\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}\right)^{-1}\mathbf{H}_{\mathcal{L},k}\mathbf{V}_k$.
5: $\quad\mathbf{W}_k \leftarrow \mathbf{I} + \mathbf{H}_k^\dagger\mathbf{V}_k^\dagger\mathbf{U}_k$.
6: $\quad$Fix $\mathbf{\Gamma}_\ell$, $\mathbf{U}_k$, and $\mathbf{W}_k$, solve the convex optimization problem (4.20). Set $(\mathbf{V}_k, \mathbf{Q}_\ell)$ to be its optimal solution.
7: **until** convergence

the last term in the objective function which involves a summation of Frobenius norms is a quadratic regularization term. It makes the optimization problem (4.19) strictly convex with respect to each optimization variable.

It is easy to verify that problem (4.19) is convex with respect to any one of the optimization variables when the other optimization variables are fixed. Specifically, when the other variables are fixed, the optimal values of $\mathbf{\Gamma}_\ell$, $\mathbf{W}_k$, and $\mathbf{U}_k$ are given by equations (4.13), (4.17), and (4.18) respectively. When $\mathbf{\Gamma}_\ell$, $\mathbf{U}_k$, and $\mathbf{W}_k$ are fixed, the optimal values of $\mathbf{V}_k$ and $\mathbf{Q}_\ell$ are solutions to the following optimization problem:

$$\begin{aligned}\min_{\mathbf{V}_k,\mathbf{Q}_\ell}\quad & \sum_{k=1}^K \mu_k \mathrm{Tr}(\mathbf{W}_k\mathbf{E}_k) + \rho\sum_{\ell=1}^L \|\mathbf{\Gamma}_\ell - \mathbf{\Omega}_\ell\|_F^2 \\ \text{s.t.}\quad & \mathrm{Tr}\left(\mathbf{\Gamma}_\ell^{-1}\mathbf{\Omega}_\ell\right) - \log|\mathbf{Q}_\ell| \le C_\ell' - \log|\mathbf{\Gamma}_\ell|\,,\ \forall \ell\in\mathcal{L}, \\ & \mathbf{Q}_\ell \succeq \mathbf{0},\quad \forall \ell\in\mathcal{L}, \\ & \mathrm{Tr}\left(\mathbf{V}_k\mathbf{V}_k^\dagger\right) \le P_k,\quad \forall k\in\mathcal{K},\end{aligned} \tag{4.20}$$

where $\mathbf{\Omega}_\ell = \sum_{k=1}^K \mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell$. The above problem is convex with respect to $\mathbf{V}_k$ and $\mathbf{Q}_\ell$, and can be solved efficiently with polynomial complexity. Standard convex optimization solver such as CVX [75] can be used for solving problem (4.20) numerically. We summarize the proposed WMMSE-SCA algorithm for single-user compression in Algorithm 4.1.

4.3.3 Convergence and Complexity Analysis

The WMMSE-SCA algorithm yields a nondecreasing sequence of objective values for problem (4.10). So the algorithm is guaranteed to converge. Moreover, it converges to a stationary point of the optimization problem. The convergence result is stated in Theorem 4.1.

Theorem 4.1 *From any initial point $(\mathbf{V}_k^{(0)}, \mathbf{Q}_\ell^{(0)})$, the proposed WMMSE-SCA algorithm is guaranteed to converge. The limit point $(\mathbf{V}_k^*, \mathbf{Q}_\ell^*)$ generated by the WMMSE-SCA algorithm is a stationary point of the weighted sum-rate maximization problem (4.10).*

Proof. See Appendix G. □

We point out here that Theorem 4.1 can also be proved following a similar procedure as that for demonstrating the convergence of WMMSE algorithm [73]. Specifically, it follows from the general optimization theory [76, Theorem 2.7.1] that the WMMSE-SCA algorithm, which is the block coordinate

www.manaraa.com

descent method applied to the reformulated problem (4.19), converges to a stationary point of (4.19). Then one can show every stationary point of (4.19) is also a stationary point of the original maximization problem (4.10), thereby establishing the claim in Theorem 4.1. However such a proof is not as simple as the proof presented in this chapter which utilizes the convergence result of the successive convex approximation algorithm [77]. We also emphasize the importance of the regularization term involving sum of Frobenius norms in the objective function of (4.19). The regularization term makes the objective function in (4.19) a strongly convex function with respect to $(\mathbf{V}_k, \mathbf{Q}_\ell)$, therefore, guarantees the convergence of Algorithm 4.1.

Assuming a typical network with $K > L > N > M$, the computational complexity of the proposed WMMSE-SCA algorithm is dominated by the joint optimization of $(\mathbf{V}_k, \mathbf{Q}_\ell)$, i.e. Step 5 of Algorithm 4.1. Step 5 solves a convex optimization problem, which can be efficiently implemented by primal-dual interior point method with approximate complexity of $\mathcal{O}\left((KM + LN)^{3.5}\right)$ [78]. Suppose that Algorithm 4.1 takes T total number of iterations to converge, the overall computational complexity of Algorithm 4.1 is therefore $\mathcal{O}\left((KM + LN)^{3.5}T\right)$.

4.4 Joint Beamforming and Compression Optimization under Wyner-Ziv coding

In single-user compression, the compression procedures across different BSs take place independently and in parallel. This separate processing across the BSs neglects the key fact that the received signals $\mathbf{y}_\ell$ in (4.2) are statistically correlated across the BS index ℓ, since they are noisy observations of the same transmitted signals $\mathbf{x}_k$. Based on this fact, Wyner-Ziv coding, which jointly processes the signals received at the CP, is expected to be superior to the pre-link single-user compression in utilizing the limited fronthaul capacities. With fixed transmitters, the advantages of Wyner-Ziv coding has been demonstrated in [38, 58]. We take one step further in this section to study the problem of jointly optimizing transmit beamforming vectors and Wyner-Ziv quantization noise covariance matrices for the VMAC scheme in uplink C-RAN.

In the implementation of Wyner-Ziv coding, we decompress the quantization codeword $\hat{\mathbf{y}}_\ell$ sequentially from one BS to the other. To this end, we need to determine a decompression order on the BS indices $\{1, 2, \ldots, L\}$. The decompression order generally affects the achievable performance of the VMAC scheme and should be optimized. However, in order to determine the optimal order that results in the largest weighted sum rate (or the maximum network utility) for the C-RAN model shown in Fig. 4.1, we need to exhaustively search over $L!$ different decoding orders, which is impractical for large L. To tackle this problem, we propose a heuristic approach to select the decompression order of $\hat{\mathbf{y}}_\ell$'s. The proposed scheme decompresses first the signals from the BS with larger value of

$$C_\ell - \log\left|\mathbf{H}_{\ell,\mathcal{K}}\mathbf{H}_{\ell,\mathcal{K}}^\dagger + \mathbf{\Sigma}_\ell\right|, \quad \forall \ell \in \mathcal{L}. \tag{4.21}$$

The rationale of this approach is to let signals from the BSs with either larger fronthaul capacity or lower received signal power be recovered first, then the recovered signals can sever as side information in helping the decompression of other BSs. This decompression order attempts to make the quantization noise levels across the BSs small. It is shown by simulation in the later section that the proposed heuristic approach works rather well for implementing Wyner-Ziv coding in practical uplink C-RAN when the

www.manaraa.com

fronthaul capacities or the received signal powers at the BSs are different.

Assume that π is the decompression order of $\hat{\mathbf{y}}_\ell$ given by the heuristic approach. Denote the index set by $\mathcal{T}_\ell = \{\pi(1), \ldots, \pi(\ell)\}$, where $\pi(\ell)$ represents the ℓth component in π. Let $\mathbf{Q}_{\mathcal{T}_\ell} = \operatorname{diag}(\{\mathbf{Q}_\ell\}_{\ell \in \mathcal{T}_\ell})$. The weighted sum rate maximization problem under Wyner-Ziv coding can be formulated as follows:

$$\begin{aligned} \max_{\mathbf{V}_k, \mathbf{Q}_\ell} \quad & \sum_{k=1}^{K} \mu_k \log \left| \mathbf{I} + \mathbf{V}_k^\dagger \mathbf{H}_{\mathcal{L},k}^\dagger \mathbf{D}_k^{-1} \mathbf{H}_{\mathcal{L},k} \mathbf{V}_k \right| \\ \text{s.t.} \quad & \log \frac{|\mathbf{\Upsilon}_{\mathcal{T}_\ell} + \mathbf{Q}_{\mathcal{T}_\ell}|}{|\mathbf{\Upsilon}_{\mathcal{T}_{\ell-1}} + \mathbf{Q}_{\mathcal{T}_{\ell-1}}|} - \log |\mathbf{Q}_{\pi(\ell)}| \le C_{\pi(\ell)}, \ \forall \ell \in \mathcal{L}, \\ & \mathbf{Q}_\ell \succeq \mathbf{0}, \quad \forall \ell \in \mathcal{L}, \\ & \operatorname{Tr}\left(\mathbf{V}_k \mathbf{V}_k^\dagger \right) \le P_k, \quad \forall k \in \mathcal{K}, \end{aligned} \tag{4.22}$$

where $\mathbf{\Upsilon}_{\mathcal{T}_\ell} = \sum_{k=1}^{K} \mathbf{H}_{\mathcal{T}_\ell,k} \mathbf{V}_k \mathbf{V}_k^\dagger \mathbf{H}_{\mathcal{T}_\ell,k}^\dagger + \operatorname{diag}(\{\mathbf{\Sigma}_\ell\}_{\ell \in \mathcal{T}_\ell})$, μ_k's are the weights associated with the users, and $\mathbf{D}_k$ is given by either equation (4.4) for the MMSE receiver or equation (4.5) for the SIC receiver.

The above problem is again non-convex, which makes finding its global optimum challenging. To efficiently solve problem (4.22), we again utilize the successive convex approximation approach proposed in the WMMSE-SCA algorithm. An obstacle to applying the convex approximation procedure directly to problem (4.22) lies in the Wyner-Ziv fronthaul constraint, which contains three log-determinant functions. To facilitate the utilization of the WMMSE-SCA algorithm, we reformulate problem (4.22) as an equivalent problem as follows,

$$\begin{aligned} \max_{\mathbf{V}_k, \mathbf{Q}_\ell} \quad & \sum_{k=1}^{K} \mu_k \log \left| \mathbf{I} + \mathbf{V}_k^\dagger \mathbf{H}_{\mathcal{L},k}^\dagger \mathbf{D}_k^{-1} \mathbf{H}_{\mathcal{L},k} \mathbf{V}_k \right| \\ \text{s.t.} \quad & \log |\mathbf{\Upsilon}_{\mathcal{T}_\ell} + \mathbf{Q}_{\mathcal{T}_\ell}| - \sum_{\ell \in \mathcal{T}_\ell} \log |\mathbf{Q}_\ell| \le \sum_{\ell \in \mathcal{T}_\ell} C_\ell, \ \forall \ell \in \mathcal{L}, \\ & \mathbf{Q}_\ell \succeq \mathbf{0}, \quad \forall \ell \in \mathcal{L}, \\ & \operatorname{Tr}\left(\mathbf{V}_k \mathbf{V}_k^\dagger \right) \le P_k, \quad \forall k \in \mathcal{K}, \end{aligned} \tag{4.23}$$

where $\mathbf{\Upsilon}_{\mathcal{T}_\ell} = \sum_{k=1}^{K} \mathbf{H}_{\mathcal{T}_\ell,k} \mathbf{V}_k \mathbf{V}_k^\dagger \mathbf{H}_{\mathcal{T}_\ell,k}^\dagger + \operatorname{diag}(\{\mathbf{\Sigma}_\ell\}_{\ell \in \mathcal{T}_\ell})$. The advantage of reformulation (4.23) is that it has similar format as (4.10), so the successive convex approximation procedure can again be used directly. Similar to the single-user case, by approximating the objective function and the fronthaul constraints in (4.23) with (3.9) and (4.16) respectively, problem (4.23) can be rewritten as

$$\begin{aligned} \max_{\substack{\mathbf{V}_k, \mathbf{Q}_\ell, \mathbf{U}_k, \\ \mathbf{W}_k, \mathbf{\Sigma}_{\mathcal{T}_\ell}}} \quad & \sum_{k=1}^{K} \mu_k \left(\log |\mathbf{W}_k| - \operatorname{Tr}(\mathbf{W}_k \mathbf{E}_k) \right) + \rho \sum_{\ell=1}^{L} \| \mathbf{\Sigma}_{\mathcal{T}_\ell} - \mathbf{\Omega}_{\mathcal{T}_\ell} \|_F^2 \\ \text{s.t.} \quad & \log |\mathbf{\Sigma}_{\mathcal{T}_\ell}| + \operatorname{Tr}\left(\mathbf{\Sigma}_{\mathcal{T}_\ell}^{-1} \mathbf{\Omega}_{\mathcal{T}_\ell} \right) - \log |\mathbf{Q}_{\mathcal{T}_\ell}| \le C'_{\mathcal{T}_\ell}, \ \forall \ell \in \mathcal{L}, \\ & \mathbf{Q}_\ell \succeq \mathbf{0}, \quad \forall \ell \in \mathcal{L}, \\ & \operatorname{Tr}\left(\mathbf{V}_k \mathbf{V}_k^\dagger \right) \le P_k, \quad \forall k \in \mathcal{K}, \end{aligned} \tag{4.24}$$

where $\rho > 0$, $\mathbf{\Omega}_{\mathcal{T}_\ell} = \sum_{k=1}^{K} \mathbf{H}_{\mathcal{T}_\ell,k} \mathbf{V}_k \mathbf{V}_k^\dagger \mathbf{H}_{\mathcal{T}_\ell,k}^\dagger + \operatorname{diag}(\{\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\}_{\ell \in \mathcal{T}_\ell})$, and $C'_{\mathcal{T}_\ell} = \sum_{\ell \in \mathcal{T}_\ell} (C_\ell + N)$. Clearly, the proposed WMMSE-SCA algorithm can be applied for solving the above optimization problem. We describe the beamforming and fronthaul compression scheme for Wyner-Ziv coding in Algorithm 4.2.

www.manaraa.com

Algorithm 4.2 Beamforming and Fronthaul Compression Optimization under Wyner-Ziv coding

1: Determine a decompression order π of $\hat{\mathbf{y}}_\ell$'s by the value of $C_\ell - \log\left|\mathbf{H}_{\ell,\mathcal{K}}\mathbf{H}_{\ell,\mathcal{K}}^\dagger + \mathbf{\Sigma}_\ell\right|$.
2: Solve the optimization problem (4.24) using Algorithm 4.1. Set $(\mathbf{V}_k, \mathbf{Q}_\ell)$ to be its optimal solution.

4.5 Separate Design of Beamforming and Compression

Although locally optimal transmit beamformers and quantization noise covariance matrices can be found using the WMMSE-SCA algorithm for any fixed user schedule, user priority, and channel condition, the implementation of WMMSE-SCA in practice can be computationally intensive, especially when the channels are under fast fading or when the scheduled users in the time-frequency slots change frequently. In this section, we aim at deriving near optimal transmit beamformers and quantization noise covariance matrices in the high SQNR regime. The main result of this section is that a simple separate design which involves optimizing transmit beamformers for the Gaussian vector multiple-access channel at the user side and using scalar quantizers with uniform quantization noise levels across the antennas at each BS is approximately optimal if an appropriate set of users are scheduled. This leads to an efficient way for the transmit beamforming and fronthaul compression design in the practical uplink C-RAN systems.

4.5.1 Quantization Noise Design Under High SQNR

The proposed approximation scheme is derived by assuming that SIC is implemented at the central receiver. Without loss of generality, let $0 = \mu_0' \le \mu_1' \le \mu_2' \le \cdots \le \mu_{K-1}' \le \mu_K' = 1$ represent the user weights μ_k normalized by their maximum value. With these weights, the users are decoded in the order of $1, 2, \ldots, K$. The decoded user messages facilitate the decoding of subsequent user messages by serving as side information.

Denote the transmit signal covariance matrix for the jth user as $\mathbf{K}_j = \mathbf{V}_j\mathbf{V}_j^\dagger$. Under single-user compression, the weighted sum rate maximization problem can be formulated as follows,

$$\begin{aligned}
\max_{\mathbf{K}_j, \mathbf{Q}_\ell} \quad & \sum_{k=1}^{K} \mu_k' \log \frac{\left|\sum_{j=k}^{K} \mathbf{H}_{\mathcal{L},j}\mathbf{K}_j\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}\right|}{\left|\sum_{j>k}^{K} \mathbf{H}_{\mathcal{L},j}\mathbf{K}_j\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}\right|} \\
\text{s.t.} \quad & \log \frac{\left|\sum_{j=1}^{K} \mathbf{H}_{\ell,j}\mathbf{K}_j\mathbf{H}_{\ell,j}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\right|}{|\mathbf{Q}_\ell|} \le C_\ell, \ \forall \ell \in \mathcal{L}, \\
& \mathbf{Q}_\ell \succeq \mathbf{0}, \quad \forall \ell \in \mathcal{L}, \\
& \mathrm{Tr}\left(\mathbf{K}_j\right) \le P_j, \quad \forall j \in \mathcal{K},
\end{aligned} \tag{4.25}$$

where $\mathbf{\Sigma} = \mathrm{diag}\left(\{\mathbf{\Sigma}_\ell\}_{\ell=1}^L\right)$ and $\mathbf{Q} = \mathrm{diag}\left(\{\mathbf{Q}_\ell\}_{\ell=1}^L\right)$.

In the following, we provide a justification that the optimal quantization noise levels should be set to be uniform across the antennas at each BS under high SQNR. Towards this end, we derive the Karush-Kuhn-Tucker (KKT) condition for the optimization problem (4.25) under the high SQNR assumption.

www.manaraa.com

To obtain the KKT condition, form the Lagrangian

$$L(\mathbf{K}_j, \mathbf{Q}_\ell, \lambda_\ell, \nu_j) = \sum_{k=1}^{K} \left(\mu'_k - \mu'_{k-1}\right) \log \left| \sum_{j=k}^{K} \mathbf{H}_{\mathcal{L},j} \mathbf{K}_j \mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q} \right| - \log |\mathbf{\Sigma} + \mathbf{Q}| \\ - \sum_{\ell=1}^{L} \lambda_\ell \log \left| \sum_{j=1}^{K} \mathbf{H}_{\ell,j} \mathbf{K}_j \mathbf{H}_{\ell,j}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell \right| + \sum_{\ell=1}^{L} \lambda_\ell \log |\mathbf{Q}_\ell| - \sum_{j=1}^{K} \nu_j \mathrm{Tr}\left(\mathbf{K}_j\right), \quad (4.26)$$

where λ_ℓ is the Lagrangian dual variable associated with the ℓth fronthaul constraint, and ν_j is Lagrangian multiplier for the jth transmit power constraint.

Setting $\partial L / \partial \mathbf{Q}_\ell$ to zero, we obtain the optimality condition as follows,

$$\sum_{k=1}^{K} \left(\mu'_k - \mu'_{k-1}\right) \mathbf{F}_\ell \left(\sum_{j=k}^{K} \mathbf{H}_{\mathcal{L},j} \mathbf{K}_j \mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q} \right)^{-1} \mathbf{F}_\ell^T - \lambda_\ell \left(\sum_{j=1}^{K} \mathbf{H}_{\ell,j} \mathbf{K}_j \mathbf{H}_{\ell,j}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell \right)^{-1} \\ - \left(\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\right)^{-1} + \lambda_\ell \mathbf{Q}_\ell^{-1} = \mathbf{0}, \quad (4.27)$$

where $\mathbf{F}_\ell = [\mathbf{0}, \ldots, \mathbf{0}, \mathbf{I}_N, \mathbf{0}, \ldots, \mathbf{0}]$ with only the ℓth $N \times N$ block being nonzero. It is easy to verify that the above optimality condition can only be satisfied if $0 \leq \lambda_\ell < 1$. Furthermore, if the overall system is to operate at reasonably high spectral efficiency, the received signal-to-noise ratios (SNRs) are likely to be high and the fronthaul capacities are likely to be large. In this case, we must have $\sum_{j=k}^{K} \mathbf{H}_{\mathcal{L},j} \mathbf{K}_j \mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q} \gg \mathbf{\Sigma} + \mathbf{Q}$ and $\sum_{j=1}^{K} \mathbf{H}_{\ell,j} \mathbf{K}_j \mathbf{H}_{\ell,j}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell \gg \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell$. Under this high SQNR condition, $\left(\sum_{j=k}^{K} \mathbf{H}_{\mathcal{L},j} \mathbf{K}_j \mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q} \right)^{-1} \approx \mathbf{0}$ and $\left(\sum_{j=1}^{K} \mathbf{H}_{\ell,j} \mathbf{K}_j \mathbf{H}_{\ell,j}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell \right)^{-1} \approx \mathbf{0}$. Then the optimality condition becomes $\left(\mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\right)^{-1} \approx \lambda_\ell \mathbf{Q}_\ell^{-1}$, i.e.,

$$\mathbf{Q}_\ell \approx \frac{\lambda_\ell}{1 - \lambda_\ell} \mathbf{\Sigma}_\ell \quad (4.28)$$

where $\lambda_\ell \in [0, 1)$ is chosen to satisfy the fronthaul capacity constraints for single-user compression. Following the same analysis, similar conclusion can also be obtained under Wyner-Ziv coding.

The above result implies that scalar quantizers with uniform quantization noise levels across the antennas at each BS are nearly optimal at high SQNR, although the quantization noise level may differ from BS to BS depending on the background noise levels and the fronthaul constraints. Note that this line of reasoning is very similar to the corresponding condition for the SISO case derived in Chapter 3.

4.5.2 Beamforming Design Under High SQNR

We next consider the optimal transmit beamforming and power allocation under high SQNR. Intuitively speaking, for maximizing the sum rate, each user should align its signaling direction with the strongest eigenmode of the effective channel and allocate power along this direction in a "water-filling" fashion. For this, we need to whiten the combined quantization and background noise and interference, then diagonalize the resulting channel to find its eigenmodes, and iteratively perform the water-filling process among the users [57]. As we see from (4.28), at high SQNR, the optimal quantization noise covariance matrices are proportional to the background noise covariance matrices. Further, if we choose $d = \min\{M, NL/G\}$, i.e., if we let the total number of user data streams be equal to the number of degrees

www.manaraa.com

of freedom in the system, then multiuser interference would be reasonably contained.

Based on the above intuition on beamforming design, we propose a simple beamforming design in which each user selects its transmit beamformers by ignoring the affect of fronthaul capacity limitation. Specifically, we consider the following weighted sum rate maximization problem for a Gaussian vector multiple-access channel:

$$\begin{aligned} \max_{\mathbf{K}_j} \quad & \sum_{k=1}^{K} \mu_k' \log \frac{\left|\sum_{j=k}^{K} \mathbf{H}_{\mathcal{L},j}\mathbf{K}_j\mathbf{H}_{\mathcal{L},j}^{\dagger} + \mathbf{\Sigma}\right|}{\left|\sum_{j>k}^{K} \mathbf{H}_{\mathcal{L},j}\mathbf{K}_j\mathbf{H}_{\mathcal{L},j}^{\dagger} + \mathbf{\Sigma}\right|} \\ \text{s.t.} \quad & \mathrm{Tr}\left(\mathbf{K}_j\right) \le P_j, \quad \forall j \in \mathcal{K}, \\ & \mathbf{K}_j \succeq \mathbf{0}, \quad \forall j \in \mathcal{K}, \end{aligned} \tag{4.29}$$

It is easy to verify that the above optimization problem is convex, and it can be efficiently solved by the interior-point method [79]. Given the optimum solution $\mathbf{K}_j^*$ to problem (4.29), each user can obtain the optimal beamformers by performing eigenvalue decomposition on $\mathbf{K}_j^*$. Let γ_i represent the ith largest eigenvalue of $\mathbf{K}_j^*$, and $\mathbf{\Psi}_i$ represent the normalized eigenvector corresponding to ith eigenvalue γ_i. Then the optimal transmit beamforming matrix for user j is just

$$\mathbf{V}_j = \sum_{i=1}^{d} \sqrt{\frac{P_k\gamma_i}{\zeta_d}}\mathbf{\Psi}_i \tag{4.30}$$

where $\zeta_d = \sum_{i=1}^{d}\gamma_i$ represents the sum of d largest of eigenvalues $\mathbf{K}_j^*$.

4.5.3 Separate Beamforming and Compression Design

The above beamforming strategy together with per-antenna scalar quantizer provide us a low-complexity separate design for transmit beamforming and fronthaul compression. With single-user compression, define

$$C^{\mathrm{SU}}\left(\beta_\ell\right) = \log \frac{\left|\sum_{k=1}^{K} \mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^{\dagger}\mathbf{H}_{\ell,k}^{\dagger} + (1+\beta_\ell)\mathbf{\Sigma}_\ell\right|}{\left|\beta_\ell\mathbf{\Sigma}_\ell\right|}. \tag{4.31}$$

With Wyner-Ziv coding, assuming without loss of generality a decoding order of $\hat{\mathbf{y}}_\ell$ from 1 to L, define

$$C^{\mathrm{WZ}}(\beta_1,\ldots,\beta_j) = \log \frac{\left|\sum_{k=1}^{K} \mathbf{H}_{\mathcal{T}_j,k}\mathbf{V}_k\mathbf{V}_k^{\dagger}\mathbf{H}_{\mathcal{T}_j,k}^{\dagger} + \mathrm{diag}\left(\{(1+\beta_\ell)\mathbf{\Sigma}_\ell\}_{\ell\in\mathcal{T}_j}\right)\right|}{\left|\mathrm{diag}\left(\{\beta_\ell\mathbf{\Sigma}_\ell\}_{\ell\in\mathcal{T}_j}\right)\right|} \tag{4.32}$$

where $\mathcal{T}_j = \{1,\ldots,j\}$. The separate transmit beamforming and fronthaul compression design scheme is summarized as Algorithm 4.3.

There are two differences between the joint design scheme and the separate design scheme. First, in the joint design, transmit beamforming are chosen to be fronthaul-aware, while the impact of limit fronthaul is ignored in the separate design. Second, in the joint design, vector quantization is applied at each BS while separate design adopts scalar quantization on each receive antenna of the BSs. It is shown by simulation in later section that the separate design performs very well in the high SQNR regime. In other regimes, the difference between the joint design and separate design represents a tradeoff between complexity and performance in implementing uplink C-RAN.

www.manaraa.com

Algorithm 4.3 Separate Design

1: Solve the convex optimization problem (4.29) and set $\mathbf{K}_j$ to be its optimal solution.
2: Perform eigen-decomposition on $\mathbf{K}_j$ to obtain eigenvalues γ_i and eigenvectors $\mathbf{\Psi}_i$. Set $\mathbf{V}_j = \sum_{i=1}^{d} \sqrt{\frac{P_k \gamma_i}{\zeta_d}} \mathbf{\Psi}_i$ for $j = 1, \ldots, K$.
3: For $\ell = 1, \ldots, L$, use bisection in $[\beta_{\min}, \beta_{\max}]$ to solve for β_ℓ in $C^{\mathrm{SU}}(\beta_\ell) = C_\ell$ for single-user compression, or $C^{\mathrm{WZ}}(\beta_1, \ldots, \beta_j) = \sum_{\ell=1}^{j} C_\ell$ for Wyner-Ziv coding.
4: Set $\mathbf{Q}_\ell = \beta_\ell \mathbf{\Sigma}_\ell$ for $\ell = 1, \ldots, L$.

Table 4.1: Multicell Network System Parameters

Cellular Layout	Hexagonal, 19-cell, 3 sectors/cell
BS-to-BS Distance	500 m
Frequency Reuse	1
Channel Bandwidth	10 MHz
Number of Users per Sector	20
Total Number of Users	420
Max Transmit Power	23 dBm
Antenna Gain	14 dBi
SNR Gap (with coding)	6 dB
Background Noise	-169 dBm/Hz
Noise Figure	7 dB
Tx/Rx Antenna No.	2×2
Distance-dependent Path Loss	$128.1 + 37.6 \log_{10}(d)$
Log-normal Shadowing	8 dB standard deviation
Shadow Fading Correlation	0.5
Cluster Size	7 cells (21 sectors)
Scheduling Strategy	WMMSE based scheduling

4.6 Simulation Results

4.6.1 Single-Cluster Network

In this section, the performances of the proposed WMMSE-SCA schemes with different compression strategies (i.e., Wyner-Ziv coding and single-user compression) and different receiving schemes (i.e., linear MMSE receiver and SIC receiver) are evaluated on a 19-cell 3-sector/cell wireless network setup with center 7 cells (i.e., 21 sectors) forming a cooperating cluster. The users are randomly located and associated with the strongest BS. The proposed WMMSE-SCA algorithm is applied to all the users within the cluster, which automatically schedules the users with non-zero beamforming vectors. Each BS is equipped with $N = 2$ antennas, each user is equipped with $M = 2$ antennas, and each user sends one data stream (i.e., $d = 1$) to the CP. Perfect channel estimation is assumed, and the CSI is made available to all BSs and to the CP. Various algorithms are run on fixed set of channels. Detailed system parameters are outlined in Table 4.1.

Under single-user compression, Fig. 4.2 and Fig. 4.3 compare the performance of the WMMSE-SCA and separate design schemes implemented either with SIC (labeled as "SIC receiver" in the figures) or without SIC (labeled as "linear receiver" in the figures) at the receiver under two different fronthaul constraints. It is shown that both the WMMSE-SCA scheme and the separate design scheme significantly outperform the baseline scheme without multicell processing. Fig. 4.2 and Fig. 4.3 show that the SIC receiver achieves significant gain as compared to the linear receiver. The performance improvement is more significant for the users with low rate.

www.manaraa.com

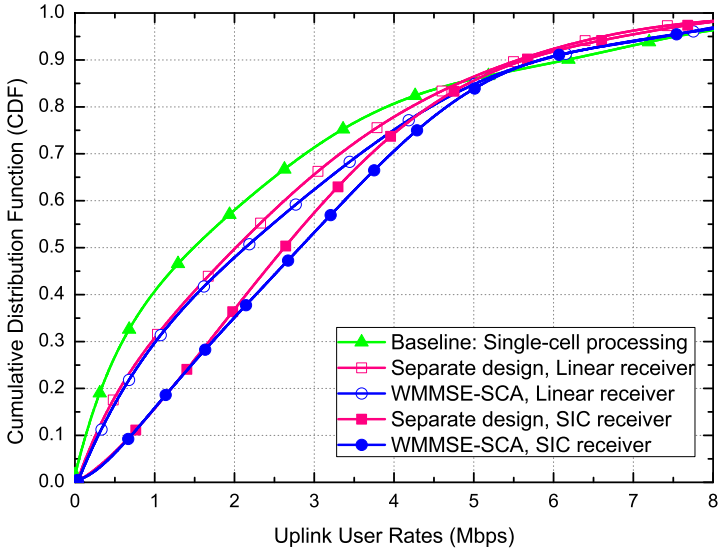


Figure 4.2: Cumulative distribution of user rates with single-user compression for a 19-cell network with center 7 cells forming a single cluster under the fronthaul capacity of 120Mbps per sector.

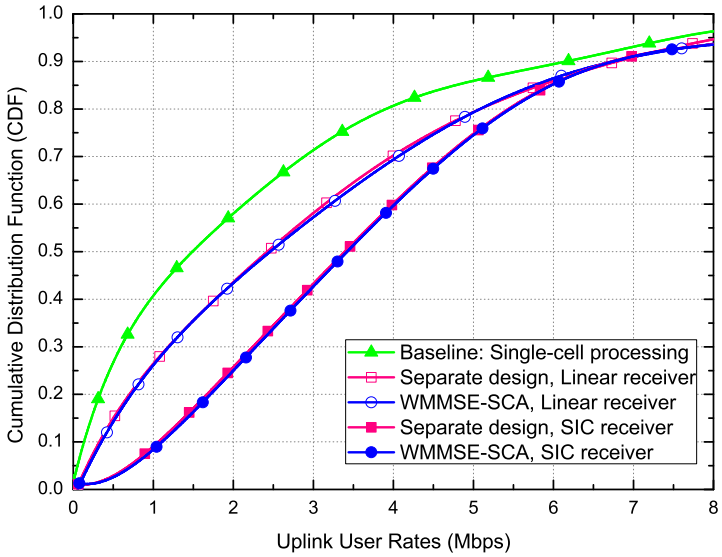


Figure 4.3: Cumulative distribution of user rates with single-user compression for a 19-cell network with center 7 cells forming a single cluster under the fronthaul capacity of 320Mbps per sector.

To further compare the performance of the proposed two schemes, Fig. 4.4 plots the average per-cell sum rate of the WMMSE-SCA scheme and the low-complexity separate design as a function of the fronthaul capacity. As the fronthaul capacity increases, the performance gap between these two schemes becomes smaller. This demonstrates the approximate optimality for separate design of transmit beamforming and fronthaul compression in the high SQNR regime.

Fig. 4.5 and Fig. 4.6 show the CDF curves of user rates for the WMMSE-SCA scheme implemented with four different choices of coding schemes: with either single-user or Wyner-Ziv compression at the BSs and with either linear MMSE or SIC receiver at the CP. It can be seen from Fig. 4.5 that under

www.manaraa.com

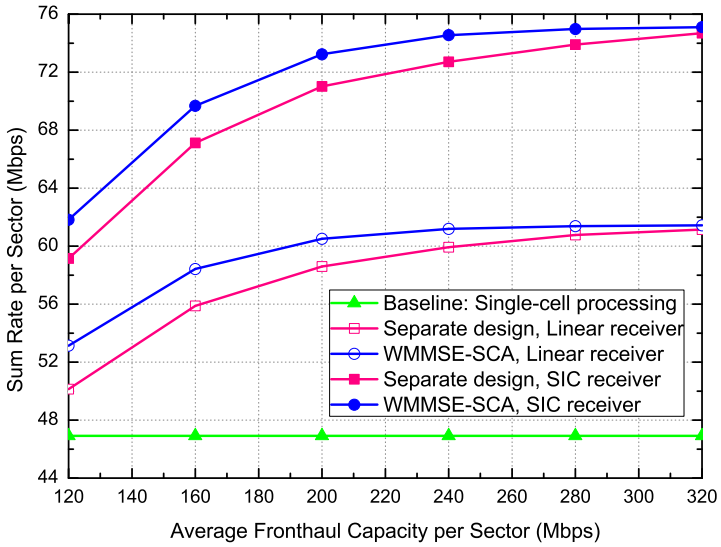


Figure 4.4: Per-cell sum rate vs. average per-sector fronthaul capacity with linear receiver and with SIC receiver for a 19-cell network with center 7 cells forming a single cluster.

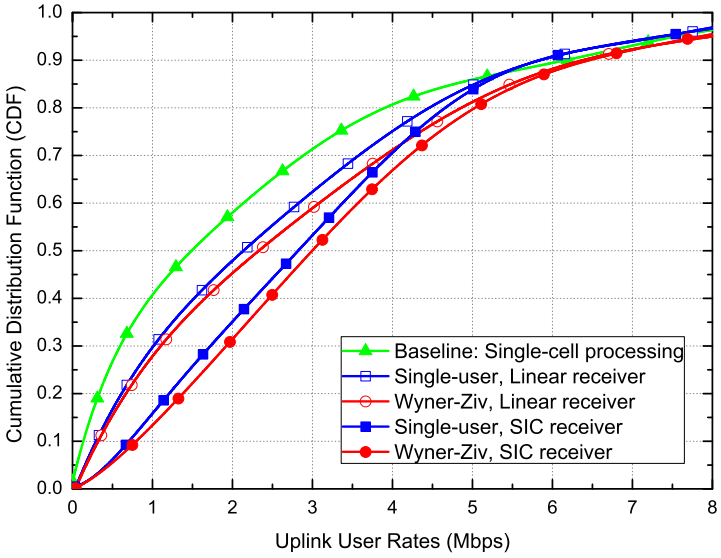


Figure 4.5: Cumulative distribution of user rates with either single-user compression or Wyner-Ziv coding for a 19-cell network with center 7 cells forming a single cluster under the fronthaul capacity of 120Mbps per sector.

the fronthaul capacity of 120Mbps, single-user compression with SIC receiver significantly improves the performance of linear MMSE receiver. Further gain on performance can be obtained if one replaces single-user compression by Wyner-Ziv coding. As the capacity of fronthaul increases to 320Mbps, as shown in Fig. 4.6, the gain due to Wyner-Ziv coding becomes negligible. In this high fronthaul scenario, SIC receiver still achieves a very large gain.

In order to quantify the performance gain brought by Wyner-Ziv coding and SIC receiver, Fig. 4.7 shows the average per-cell sum rate obtained by different schemes as the average capacity of fronthaul increases. It is observed that, under fronthaul capacity of 320Mpbs, both SIC receiver and Wyner-Ziv

www.manaraa.com

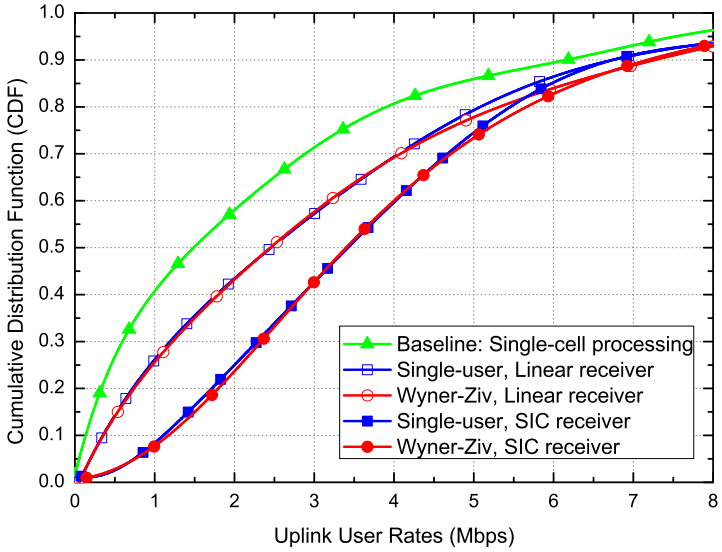


Figure 4.6: Cumulative distribution of user rates with either single-user compression or Wyner-Ziv coding using WMMSE-SCA algorithm for a 19-cell network with center 7 cells forming a single cluster under the fronthaul capacity of 320Mbps per sector.

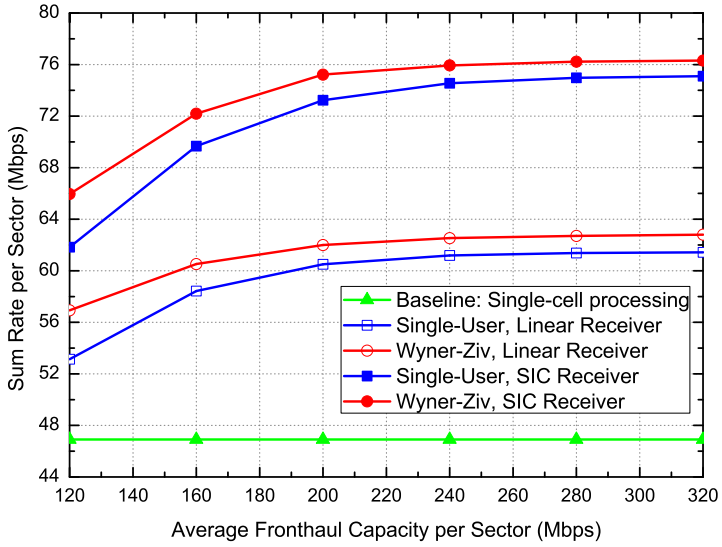


Figure 4.7: Per-cell sum rate vs. average per-cell fronthaul capacity with either single-user compression or Wyner-Ziv coding using WMMSE-SCA algorithm for a 19-cell network with center 7 cells forming a single cluster.

coding outperform the linear receiver and single-user compression respectively. But the performance improvement of SIC receiver upon linear receiver is much larger than the gain of Wyner-Ziv coding over single-user compression.

4.6.2 Multi-Cluster Network

The performance of the proposed WMMSE-SCA scheme is further evaluated for a large-scale multicell network with 65 cells and 10 mobile users randomly located within each cell. The BS to BS distance

www.manaraa.com

Table 4.2: Multi-Cluster Network Parameters

Cellular Layout	Hexagonal
BS-to-BS Distance	200 m
Frequency Reuse	1
Channel Bandwidth	10 MHz
Number of Users per Cell	10
Number of Cells	65
Total Number of Users	650
Max Transmit Power	23 dBm
Antenna Gain	14 dBi
SNR Gap (with coding)	6 dB
Background Noise	-169 dBm/Hz
Noise Figure	7 dB
Tx Antenna No.	2
Rx Antenna No.	4
Distance-dependent Path Loss	$128.1 + 37.6 \log_{10}(d)$
Log-normal Shadowing	8 dB standard deviation
Shadow Fading Correlation	0.5
Scheduling Strategy	Round-robin

is set to be 200m, each user is equipped with 2 transmit antennas, and each BS is equipped with 4 receive antennas. The channel is assumed to be flat-fading. Round-robin user scheduling is used on a per-cell basis and system is operated with loading factor 0.5, i.e., in each time slot, BS schedules two users. Detailed system parameters are outlined in Table 4.2. Two different clustering strategies, i.e., disjoint clustering and user-centric clustering, are applied to form clusters within the network. Disjoint clustering partitions the BSs in the network into nonoverlapping sets of cooperating clusters. In user-centric clustering, each user chooses a set of nearest BSs to form a cooperation cluster, and cooperating clusters overlap, which makes the implementation of Wyner-Ziv coding and SIC receiver under fronthaul capacity constraints of (4.9) more difficult. Therefore, for fair comparison, we only consider here the case where single-user compression and linear MMSE receiver are employed.

Fig. 4.8 and Fig. 4.9 show the CDF plots of user rates achieved with both disjoint clustering and user-centric clustering with WMMSE-SCA. It is clear that with optimized beamforming and fronthaul compression, the user-centric clustering significantly improves over disjoint clustering, and both of these two schemes improve as the cluster size increases. As the capacity of fronthaul links increases from 120Mbps to 360Mbps, the performance gap between the two clustering schemes becomes larger. Further, for disjoint clustering, increasing the cluster size from 2 to 6 achieves 60% performance improvement for the 50-percentile rate. This gain doubles when we further replace disjoint clustering with user-centric clustering.

Fig. 4.10 plots the average per-cell sum rate as the fronthaul capacity increases. The result again shows that user-centric clustering achieves significant performance gain over disjoint clustering. When cluster size increases to 6, to achieve per-cell sum rate of 110Mbps, disjoint clustering needs fronthaul capacity of 360Mbps, while user-centric needs 220Mbps, which is more than 60% improvement on the fronthaul requirement.

Finally, the performance of the two different clustering strategies are compared as a function of cluster size in Fig. 4.11. It is shown that for both disjoint clustering and user-centric clustering, the average per-cell sum rate increases as either the cluster size or fronthaul capacity increases. As expected, user-centric clustering always outperforms disjoint clustering. If we compare the performance of disjoint clustering

www.manaraa.com

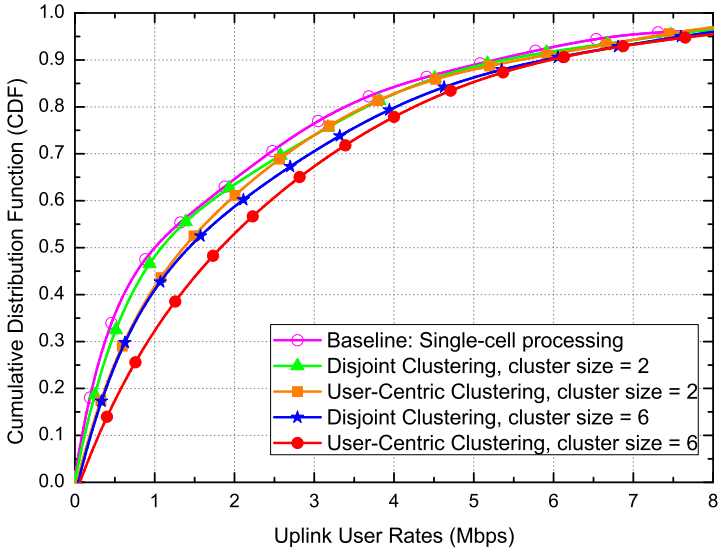


Figure 4.8: Cumulative distribution of user rates for the WMMSE-SCA algorithm with single-user compression under the average fronthaul capacity of 120Mbps with either disjoint or user-centric clustering for a multi-cluster network.

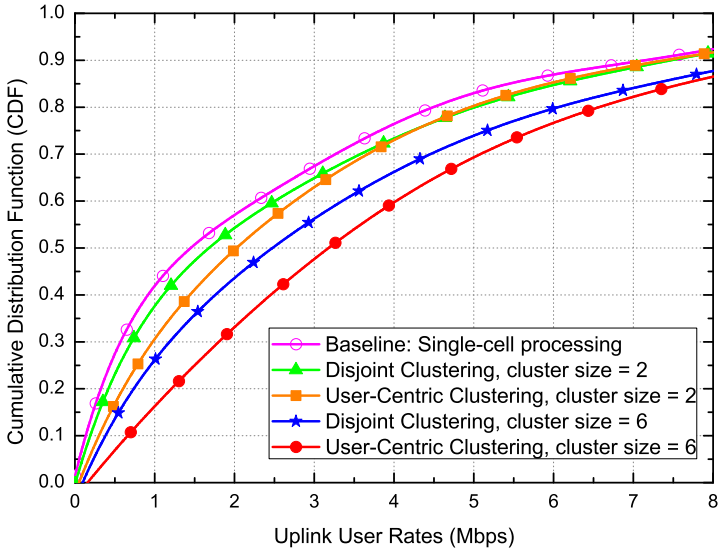


Figure 4.9: Cumulative distribution of user rates for the WMMSE-SCA algorithm with single-user compression under the average fronthaul capacity of 360Mbps with either disjoint or user-centric clustering for a multi-cluster network.

with fronthaul capacity of 360Mbps with user-centric clustering with fronthaul capacity of 240Mbps, we see that even with 120Mbps lower fronthaul capacity, user-centric clustering already achieves higher per-cell sum rate. This improvement on per-cell sum rate becomes larger as the cluster size increases.

www.manaraa.com

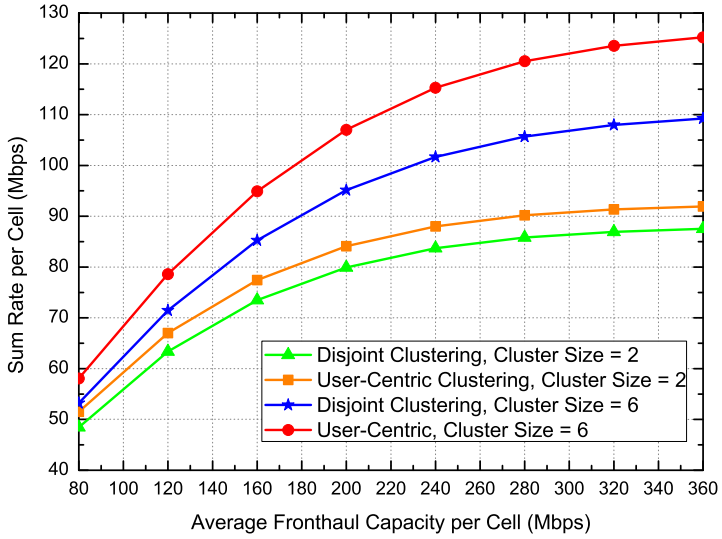


Figure 4.10: Per-cell sum rate vs. average per-cell fronthaul capacity of the WMMSE-SCA algorithm with single-user compression for a multi-cluster network under different clustering strategies and different cluster size.

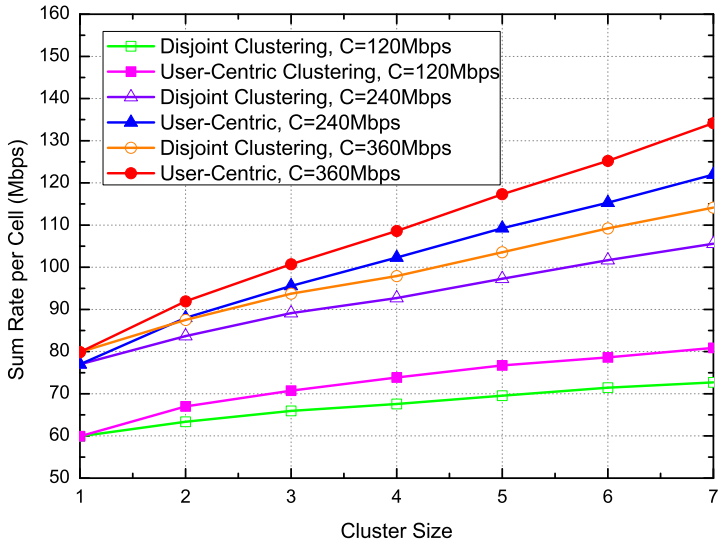


Figure 4.11: Per-cell sum rate vs. cluster size for the WMMSE-SCA algorithm with single-user compression for a multi-cluster network under different clustering strategies and different fronthaul capacity constraints.

4.7 Summary

This chapter studies the fronthaul compression and transmit beamforming design for an uplink MIMO C-RAN system. From algorithm design perspective, we propose a novel WMMSE-SCA algorithm to efficiently optimize the transmit beamformer and quantization noise covariance matrix jointly for maximizing the weighted sum rate with either Wyner-Ziv coding or single-user compression. Further, we propose a separate design consisting of transmit beamforming optimized for the Gaussian vector multiple-access channel without accounting for compression together with scalar quantization with uniform quantiza-

www.manaraa.com

tion noise levels across the antennas at each BS. This low-complexity separate design is shown to be near optimal for maximizing the weighted sum rate when the SQNR is high. The performance of optimized beamforming and fronthaul compression is evaluated for practical multicell networks with different compression strategies, different receiving schemes, and different clustering methods. Numerical results show that, with optimized beamforming and fronthaul compression, C-RAN can significantly improve the overall performance of MIMO cellular networks. Most of the performance gain are due to the implementation of SIC at the central receiver. Finally, user-centric clustering significantly outperforms disjoint clustering in terms of fronthaul capacity saving.

www.manaraa.com

Chapter 5

Conclusion

This thesis studies an uplink C-RAN model under a practical implementation constraint of capacity-limited fronthaul links. From the theoretical analytic point of view, this thesis provides a justification on the optimality of Gaussian input signals at the user, Gaussian quantization at the BSs, and successive decoding at the CP for implementing compress-and-forward in uplink C-RAN. Specifically, this thesis shows that generalized successive decoding achieves the same rate region as joint decoding under a sum fronthaul capacity constraint and successive decoding of the quantization codewords first, and then the user message codewords achieves the same maximum sum rate as joint decoding. Furthermore, it is shown that with Gaussian input distribution, Gaussian quantizer maximizes the achievable rate region for joint decoding. Additionally, with Gaussian input and Gaussian quantization, by setting quantization noise levels to be proportional to the background noise levels, successive decoding with Wyner-Ziv compression can achieve the sum capacity of the uplink C-RAN model to within a constant gap. A similar constant-gap result is also obtained for the single-user compression under a diagonally dominant channel condition.

From algorithm design perspective, the thesis first investigates the optimization of fronthaul compression for uplink C-RAN under a sum fronthaul constraint and proposes a novel alternating convex optimization algorithm for efficiently finding the optimal quantization noise levels that maximizes the weighted sum rate. The thesis further studies the joint optimization of transmit beamforming and fronthaul compression for uplink C-RAN under individual fronthaul constraints. We propose the joint optimization of the transmit beamformers and the quantization noise covariance matrices for maximizing the network utility and develop a novel WMMSE-SCA algorithm for maximizing the weighted sum rate under the user transmit power and fronthaul capacity constraints with either Wyner-Ziv coding or single-user compression. The performance of the proposed schemes are evaluated for practical multicell and heterogeneous networks. Numerical results show that with optimized beamforming and fronthaul compression, C-RAN can significantly improve the overall performance of conventional cellular networks.

We finally conclude this thesis by pointing out some possible future research directions. For the relaying strategies at the BSs, instead of compression, the BSs can also perform decoding and computation, which result in the partial decode-forward scheme [51] and the compute-and-forward scheme. In partial decode-forward, user messages are decoded locally at the BSs, which brings low latency but also poor performance. The tradeoff between latency and performance for partial decode-forward in C-RAN is an interesting problem for future study. On the other hand, in compute-and-forward each BS

www.manaraa.com

computes a linear combination of message codewords and send it to the CP for decoding. Due to the fact of no noise accumulation, compute-and-forward can outperform compress-and-forward for the C-RAN system under certain channel conditions and fronthaul constraints [30]. However, the performance compute-and-forward could be sensitive to the channel gain matrix. The construction of good codes for compute-and-forward with competitive performance under arbitrary channel conditions and fronthaul constraints in uplink C-RAN is still an open problem.

As a counterpart of uplink C-RAN, the downlink of the C-RAN system can thought as a broadcast relay network, where the CP sends multiple data streams to different user terminals. A number of pratical coding schemes have been studied for the downlink of C-RAN in the literature, which includes the message sharing scheme [71], the pure compression scheme [74], and the hybrid scheme [80]. From information theoretical point of view, the state-of-the-art coding strategy for downlink C-RAN is the so-called distributed decode-forward [81], which has be shown to achieve the capacity region of Gaussian broadcast relay networks to within a constant gap. However, to achieve such a constant-gap result, one needs to employ multiple block coding at the transmitter, which is computationally prohibitive for practical implementation. Exploring the tradeoff between complexity and performance for downlink C-RAN is an attractive but challenging problem for future research.

www.manaraa.com

Appendix A

Optimality of Generalized Successive Decoding

In this appendix, we prove Theorem 2.1, which states the equivalence between generalized successive decoding and joint decoding under a sum-capacity fronthaul constraint. We begin by introducing an outer bound for the achievable rate region of joint decoding under a sum fronthaul constraint. Under the sum fronthaul capacity constraint, define the rate-fronthaul region for joint decoding $\mathcal{P}^o_{JD,s}$ as the closure of the convex hull of all $(R_1, R_2, \ldots, R_K, C)$ satisfying

$$\begin{cases} \sum\limits_{k\in\mathcal{T}} R_k < \min\left\{ C - \sum\limits_{\ell\in\mathcal{L}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}\right),\ I\left(\mathbf{X}_{\mathcal{T}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{\mathcal{T}^c}\right) \right\}, \quad \forall\, \mathcal{T} \subseteq \mathcal{K}, \\ C > \sum_{\ell\in\mathcal{L}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}\right) \end{cases} \tag{A.1}$$

for some product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$. Under fixed sum fronthaul constraint C, define the region $\mathcal{R}^o_{JD,s}$ as follows

$$\mathcal{R}^o_{JD,s} = \left\{ (R_1, \ldots, R_K) : (R_1, \cdots, R_K, C) \in \mathcal{P}^o_{JD,s} \right\} \tag{A.2}$$

Note that the rate region $\mathcal{R}^o_{JD,s}$ is an outer bound for joint decoding rate region (2.10) because only the constraints corresponding to $\mathcal{S} = \emptyset$ and $\mathcal{S} = \mathcal{L}$ are included. These constraints turn out to be the only active ones under the sum fronthaul constraint $\sum_{\ell=1}^L C_\ell \le C$ and $C_\ell \ge 0$.

Under the sum fronthaul constraint, the generalized successive decoding region $\mathcal{P}_{GSD,s}(\pi)$ for decoding order π can be derived from (2.2) by letting $\sum_{\ell=1}^L C_\ell = C$. More specifically, $\mathcal{P}_{GSD,s}(\pi)$ is the closure of the convex hull of all $(R_1, R_2, \ldots, R_K, C)$ satisfying

$$\begin{cases} R_k < I\left(\mathbf{X}_k; \hat{\mathbf{Y}}_{\mathcal{J}_{\mathbf{X}_k}} | \mathbf{X}_{\mathcal{I}_{\mathbf{X}_k}}\right), \quad \forall\, k \in \mathcal{K}, \\ C > \sum\limits_{\ell=1}^L I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \hat{\mathbf{Y}}_{\mathcal{J}_{\mathbf{Y}_\ell}}, \mathbf{X}_{\mathcal{I}_{\mathbf{Y}_\ell}}\right), \end{cases} \tag{A.3}$$

for some product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$, where $\mathcal{I}_{\mathbf{X}_k}, \mathcal{I}_{\mathbf{Y}_\ell}$ are the indices of user messages that are decoded before $\mathbf{X}_k$ and $\mathbf{Y}_\ell$ under the permutation π, and $\mathcal{J}_{\mathbf{X}_k}$, $\mathcal{J}_{\mathbf{Y}_\ell}$ are the indices of quanti-

www.manaraa.com

zation codewords that are decoded before $\mathbf{X}_k$ and $\mathbf{Y}_\ell$ under decoding order π. Define $\mathcal{P}^*_{GSD,s}$ to be the closure of the convex hull of all $\mathcal{P}_{GSD,s}(\pi)$'s over decoding order π's, i.e. $\mathcal{P}^*_{GSD,s} = \text{co}\left(\bigcup_\pi \mathcal{P}_{GSD,s}(\pi)\right)$.

We say a point $(R_1, \ldots, R_K, C)$ is *dominated* by a point in $\mathcal{P}^*_{GSD,S}$ if there exists some $(R'_1, \ldots, R'_K, C')$ in $\mathcal{P}^*_{GSD,s}$ for which $R_k \le R'_k$ for $k = 1, 2, \ldots, K$, and $C \ge C'$.

Given the definitions of $\mathcal{R}^*_{GSD,s}$, $\mathcal{R}^*_{JD,s}$ and $\mathcal{R}^o_{JD,s}$, it is easy to see that $\mathcal{R}^*_{GSD,s} \subseteq \mathcal{R}^*_{JD,s} \subseteq \mathcal{R}^o_{JD,s}$. To show $\mathcal{R}^*_{GSD,s} = \mathcal{R}^*_{JD,s}$, it suffices to show $\mathcal{R}^o_{JD,s} \subseteq \mathcal{R}^*_{GSD,s}$, which is equivalent to show that if a point $(R_1, R_2, \ldots, R_K, C) \in \mathcal{P}^o_{JD,s}$, then the same point $(R_1, R_2, \ldots, R_K, C) \in \mathcal{P}^*_{GSD,s}$ also. To show this, it suffices to show that for any fixed product distribution $\prod_{k=1}^K p(\mathbf{x}_k) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell|\mathbf{y}_\ell)$ and fixed C, each extreme point $(R_1, \ldots, R_K, C)$ as defined by (A.1) is dominated by a point in $\mathcal{P}^*_{GSD,s}$ with the average sum fronthaul capacity requirement at most C.

To this end, define a set function $f : 2^{\mathcal{K}} \to \mathbb{R}$ as follows:

$$f(\mathcal{T}) := \min\left\{C - \sum_{\ell\in\mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell|\mathbf{X}_\mathcal{K}),\ \ I\left(\mathbf{X}_\mathcal{T}; \hat{\mathbf{Y}}_\mathcal{L}|\mathbf{X}_{\mathcal{T}^c}\right)\right\},$$

for each $\mathcal{T} \subseteq \mathcal{K}$. It can be verified that the function f is a submodular function (Appendix B, Lemma B.1). By construction, $(R_1, R_2, \ldots, R_K)$ as defined by (A.2) satisfies

$$\sum_{k\in\mathcal{T}} R_k \le f(\mathcal{T}),$$

which is a submodular polyhedron associated with f.

It follows by basic results in submodular optimization (Appendix B, Proposition B.1) that, for a linear ordering $i_1 \prec i_2 \prec \cdots \prec i_K$ on the set $\mathcal{K}$, an extreme point of $\mathcal{R}^*_{JD,s}$ can be computed as follows

$$\tilde{R}_{i_j} = f(\{i_1, \ldots, i_j\}) - f(\{i_1, \ldots, i_{j-1}\}).$$

Furthermore, the extreme points of $\mathcal{R}^o_{JD}$ can be enumerated over all the orderings of $\mathcal{K}$. Each ordering of $\mathcal{K}$ is analyzed in the same manner, hence for notational simplicity we only consider the natural ordering $i_j = j$ in the following proof.

By construction,

$$\tilde{R}_j = \min\left\{C - \sum_{\ell\in\mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell|\mathbf{X}_\mathcal{K}),\ \ I\left(\mathbf{X}_1^j; \hat{\mathbf{Y}}_\mathcal{L}|\mathbf{X}_{j+1}^K\right)\right\} - \min\left\{C - \sum_{\ell\in\mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell|\mathbf{X}_\mathcal{K}),\ \ I\left(\mathbf{X}_1^{j-1}; \hat{\mathbf{Y}}_\mathcal{L}|\mathbf{X}_j^K\right)\right\}.$$

Note that $I\left(\mathbf{X}_1^j; \hat{\mathbf{Y}}_\mathcal{L}|\mathbf{X}_{j+1}^K\right) \ge I\left(\mathbf{X}_1^{j-1}; \hat{\mathbf{Y}}_\mathcal{L}|\mathbf{X}_j^K\right)$. Then, it suffices to check the following two cases:

- Case 1: $C - \sum_{\ell\in\mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell|\mathbf{X}_\mathcal{K}) \ge I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{L}\right)$. In this case the resulting extreme point $\mathbf{r}^1_{JD} =$

www.manaraa.com

$(\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_K, C)$ satisfies

$$\begin{cases} \tilde{R}_j = I\left(\mathbf{X}_j; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{j+1}^K\right), \text{ for } j = 1, 2, \ldots, K-1, \\ \tilde{R}_K = I\left(\mathbf{X}_K; \hat{\mathbf{Y}}_{\mathcal{L}}\right), \\ C \geq I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{\mathcal{L}}\right) + \sum_{\ell \in \mathcal{L}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}\right), \end{cases}$$

Following the Markov chain

$$\hat{\mathbf{Y}}_i \leftrightarrow \mathbf{Y}_i \leftrightarrow \mathbf{X}_{\mathcal{K}} \leftrightarrow \mathbf{Y}_j \leftrightarrow \hat{\mathbf{Y}}_j, \quad \forall\, i \neq j$$

it can be shown that

$$\sum_{\ell \in \mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}) + I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{\mathcal{L}}\right) = I(\mathbf{Y}_{\mathcal{L}}; \hat{\mathbf{Y}}_{\mathcal{L}}) \leq C.$$

Clearly, $\mathbf{r}_{JD}^1$ belongs to the polyhedron $\mathcal{P}_{GSD,s}^*$ with successive decoding, since it can be achieved by the decoding order of $\hat{\mathbf{Y}}_{\mathcal{L}} \to \mathbf{X}_K \to \cdots \to \mathbf{X}_1$. Thus, $\mathbf{r}_{JD}^1$ is dominated by a point in $\mathcal{P}_{GSD,s}^*$.

- Case 2: Consider that

$$I\left(\mathbf{X}_1^{j-1}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_j^K\right) \leq C - \sum_{\ell \in \mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}) \leq I\left(\mathbf{X}_1^j; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{j+1}^K\right)$$

for some $1 \leq j < K$. The resulting extreme point $\mathbf{r}_{JD}^2 = (\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_K, C)$ satisfies

$$\begin{cases} \tilde{R}_i = I\left(\mathbf{X}_i; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{i+1}^K\right), & \text{for } i < j, \\ \tilde{R}_i = \left[C - \sum_{\ell \in \mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}) - I\left(\mathbf{X}_1^{j-1}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_i^K\right)\right]^+, & \text{for } i = j, \\ \tilde{R}_i = 0, & \text{for } i > j, \\ C \leq I\left(\mathbf{X}_1^j; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{j+1}^K\right) + \sum_{\ell \in \mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}). \end{cases}$$

where $[\cdot]^+$ means $\max\{\cdot, 0\}$. Note that users with index $i > j$ are inactive, and are essentially removed from the network. Now consider generalized successive decoding with the following two different decoding orders:

(i) Decoding order 1 satisfies

$$\mathbf{X}_{j+1} \to \ldots \to \mathbf{X}_K \to \hat{\mathbf{Y}}_{\mathcal{L}} \to \mathbf{X}_j \to \ldots \to \mathbf{X}_1.$$

The resulting extreme point $\mathbf{r}_{GSD}^{(1)} = (R_1^{(1)}, \ldots, R_K^{(1)}, C^{(1)})$ of $\mathcal{P}_{GSD,s}^*$ satisfies

$$\begin{cases} R_i^{(1)} = I\left(\mathbf{X}_i; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{i+1}^K\right), & \text{for } i \leq j, \\ R_i^{(1)} = 0, & \text{for } i > j, \\ C^{(1)} = I\left(\mathbf{Y}_{\mathcal{L}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{j+1}^K\right). \end{cases}$$

www.manaraa.com

where $C^{(1)}$ represents the required fronthaul capacity in order to achieve the above rate tuple $(R_1^{(1)},\ldots,R_K^{(1)})$ with decoding order 1.

(ii) Decoding order 2 is

$$\mathbf{X}_j \to \ldots \to \mathbf{X}_K \to \hat{\mathbf{Y}}_{\mathcal{L}} \to \mathbf{X}_{j+1} \to \ldots \to \mathbf{X}_1.$$

The resulting extreme point $\mathbf{r}_{GSD}^{(2)} = (R_1^{(2)},\ldots,R_K^{(2)},C^{(2)})$ of $\mathcal{P}_{GSD,s}^*$ satisfies

$$\begin{cases} R_i^{(2)} = I\left(\mathbf{X}_i;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{i+1}^K\right), & \text{for } i<j,\\ R_i^{(2)} = 0, & \text{for } i\geq j,\\ C^{(2)} = I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right). \end{cases}$$

where $C^{(1)}$ represents the required fronthaul capacity in order to achieve the above rate tuple $(R_1^{(2)},\ldots,R_K^{(2)})$ with decoding order 2. Observe that the rate tuples $(R_1^{(1)},\ldots,R_K^{(1)})$ and $(R_1^{(2)},\ldots,R_K^{(2)})$ given by above two decoding orders different at only jth component, where $R_j^{(1)} = I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)$ and $R_j^{(2)}=0$ and $R_i^{(1)} = R_i^{(2)} = \tilde{R}_i$ for all $i<j$. Now choose a parameter θ such that

$$\theta = \frac{C - \sum_{\ell\in\mathcal{L}} I(\mathbf{Y}_\ell;\hat{\mathbf{Y}}_\ell|\mathbf{X}_{\mathcal{K}}) - I\left(\mathbf{X}_1^{j-1};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right)}{I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)}. \tag{A.4}$$

Following the Markov chain $\mathbf{X}_{\mathcal{K}} \leftrightarrow \mathbf{Y}_{\mathcal{L}} \leftrightarrow \hat{\mathbf{Y}}_{\mathcal{L}}$, we have the following identity,

$$1-\theta = \frac{I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right) - C}{I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)}.$$

Consider the following point: $\mathbf{r}_{GSD}^{\theta} = \theta\mathbf{r}_{GSD}^{(1)} + (1-\theta)\mathbf{r}_{GSD}^{(2)}$, which is in $\mathcal{P}_{GSD,s}^*$. The corresponding sum fronthaul requirement is given by

$$\begin{aligned}
\theta C^{(1)} + (1-\theta)C^{(2)} &= \theta I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right) + (1-\theta)I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right)\\
&= C\cdot\frac{I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right) - I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right)}{I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)} + \frac{I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)}{I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)}\\
&\quad\times\left[I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right) - I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_1^K\right) - I\left(\mathbf{X}_1^{j-1};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right)\right]\\
&\overset{(c)}{=} C\cdot\frac{I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right) - I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right)}{I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)} + \frac{I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)}{I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)}\\
&\quad\times\left[I\left(\mathbf{X}_1^{j-1},\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right) - I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_1^K\right) - I\left(\mathbf{X}_1^{j-1};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right)\right]\\
&\overset{(d)}{\leq} C\cdot\frac{I\left(\mathbf{X}_j,\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right) - I\left(\mathbf{Y}_{\mathcal{L}};\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_j^K\right)}{I\left(\mathbf{X}_j;\hat{\mathbf{Y}}_{\mathcal{L}}|\mathbf{X}_{j+1}^K\right)}\\
&= C,
\end{aligned} \tag{A.5}$$

www.manaraa.com

where the equality (c) follows from the fact that $I\left(\mathbf{X}_1^{j-1}, \mathbf{Y}_{\mathcal{L}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_j^K\right) = I\left(\mathbf{Y}_{\mathcal{L}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_j^K\right)$ due to Markov chain $\mathbf{X}_{\mathcal{K}} \leftrightarrow \mathbf{Y}_{\mathcal{L}} \leftrightarrow \hat{\mathbf{Y}}_{\mathcal{L}}$, and inequality (d) follows from the fact that $I\left(\mathbf{Y}_{\mathcal{L}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{j+1}^K\right) \leq I\left(\mathbf{X}_j, \mathbf{Y}_{\mathcal{L}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{j+1}^K\right)$. Thus, we have that $\mathbf{r}_{JD}^2$ is dominated by some point lying on line segment between $\mathbf{r}_{GSD}^{(1)}$ and $\mathbf{r}_{GSD}^{(2)}$, which lies in $\mathcal{P}_{GSD,s}^*$.

Therefore, for every extreme point $(\tilde{R}_1, \ldots, \tilde{R}_K)$ of $\mathcal{R}_{JD}^o$, the point $(\tilde{R}_1, \ldots, \tilde{R}_K, C)$ lies in $\mathcal{P}_{GSD,s}^*$. This completes the proof.

www.manaraa.com

Appendix B

Submodular Functions

In this appendix, we review some basic results in submodular optimization used proving Theorem 2.1 and Theorem 2.2. We tailor our statements toward submodularity and supermodularity, which are used in the proofs.

We begin with the definition of submodular function.

Definition B.0.1. Let $\mathcal{D} = \{1, \ldots, n\}$ be a finite set. A set function $f : 2^{\mathcal{D}} \to \mathbb{R}$ is submodular if for all $\mathcal{S}, \mathcal{T} \subseteq \mathcal{D}$,

$$f(\mathcal{S}) + f(\mathcal{T}) \geq f(\mathcal{S} \cup \mathcal{T}) + f(\mathcal{S} \cap \mathcal{T}). \tag{B.1}$$

Definition B.0.2. Let $\mathcal{E} = \{1, \ldots, m\}$ be a finite set. A set function $g : 2^{\mathcal{E}} \to \mathbb{R}$ is supermodular if for all $\mathcal{S}, \mathcal{T} \subseteq \mathcal{E}$,

$$g(\mathcal{S}) + g(\mathcal{T}) \leq g(\mathcal{S} \cup \mathcal{T}) + g(\mathcal{S} \cap \mathcal{T}). \tag{B.2}$$

If the function f is submodular, we call a polyhedron defined by

$$\mathcal{P}(f) = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_{i \in \mathcal{S}} x_i \leq f(\mathcal{S}), \ \ \forall\, \mathcal{S} \subseteq \mathcal{D} \right\} \tag{B.3}$$

the submodular polyhedron associated with the submodular function f. Similarly, we define the supermodular polyhedron $\mathcal{P}(g)$ to be the set of $(x_1, \ldots, x_n) \in \mathbb{R}^n$ satisfying

$$\sum_{i \in \mathcal{T}} x_i \geq g(\mathcal{T}), \ \ \forall\, \mathcal{T} \subseteq \mathcal{E}. \tag{B.4}$$

We say a point in $\mathcal{P}(f)$ is an extreme point if it cannot be expressed as a convex combination of the other two points in $\mathcal{P}(f)$.

One important property of submodular polyhedron is that all the extreme points can be enumerated through solving a linear optimization. The following proposition provides an algorithm that enumerates the extreme points of $\mathcal{P}(f)$.

Proposition B.1 ([82] [83]) *For a linear ordering $i_1 \prec i_2 \prec \cdots \prec i_n$ of the elements in $\mathcal{D}$, Algorithm B.1 returns an extreme point $(v_1, \ldots, v_n)$ of $\mathcal{P}(f)$. Moreover, all extreme points of $\mathcal{P}(f)$ can be enumerated by considering all linear orderings of the elements of $\mathcal{D}$.*

www.manaraa.com

Algorithm B.1 Greedy Algorithm for Submodular Polyhedron

1: **comment**: Returns extreme point $(v_1, \ldots, v_n)$ of $\mathcal{P}(f)$ with the ordering $\prec$.
2: **for** $j = 1, \ldots, n$ **do**
3: Set $v_j = f(\{i_1, i_2, \ldots, i_j\}) - f(\{i_1, i_2, \ldots, i_{j-1}\})$.
4: **end for**
5: **return** $(v_1, \ldots, v_n)$

Proposition B.1 is the key tool we employ to prove Theorem 2.1 and Theorem 2.2. In order to apply this proposition, we require the following lemmas,

Lemma B.1 *For any joint distribution* $\prod_{k=1}^{K} p(\mathbf{x}_k) \prod_{\ell=1}^{L} p\left(\mathbf{y}_\ell | \mathbf{x}_1^K\right) \prod_{\ell=1}^{L} p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$ *and fixed* $C \in \mathbb{R}$*, the set function* $f : 2^{\mathcal{K}} \to \mathbb{R}$ *defined as follows*

$$f(\mathcal{T}) := \min \left\{ C - \sum_{\ell \in \mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}), \;\; I\left(\mathbf{X}_{\mathcal{T}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{\mathcal{T}^c}\right) \right\}$$

is submodular.

Proof. Define a set function $f'(\mathcal{T}) = I\left(\mathbf{X}_{\mathcal{T}}; \hat{\mathbf{Y}}_{\mathcal{L}} | \mathbf{X}_{\mathcal{T}^c}\right)$. By definition, it can be verified that function f' is submodular [84]. Under fixed sum fronthaul capacity C and conditional distribution $\prod_{\ell=1}^{L} p_{\hat{\mathbf{Y}}_\ell | \mathbf{Y}_\ell}$, the expression $C - \sum_{\ell \in \mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}})$ is a constant. Let $C' = C - \sum_{\ell \in \mathcal{L}} I(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}})$. Now the problem reduces to show that $f(\mathcal{T}) = \min\{C', f'(\mathcal{T})\}$ is submodular.

Next, observe that f' is monotonically increasing, i.e. if $\mathcal{S} \subset \mathcal{T}$, then $f'(\mathcal{S}) \leq f'(\mathcal{T})$. Thus, fixing $\mathcal{S}, \mathcal{T} \subseteq \mathcal{K}$, we can assume without loss of generality that

$$f'(\mathcal{S} \cap \mathcal{T}) \leq f'(\mathcal{S}) \leq f'(\mathcal{T}) \leq f'(\mathcal{S} \cup \mathcal{T})$$

If $C' \leq f'(\mathcal{S} \cap \mathcal{T})$, then $f(\mathcal{S}) = f(\mathcal{T}) = f(\mathcal{S} \cap \mathcal{T}) = f(\mathcal{T}) \leq f'(\mathcal{S} \cup \mathcal{T}) = C'$. Clearly, f is then submodular. On the other hand, if $C' \geq f'(\mathcal{S} \cup \mathcal{T})$, then $f(\mathcal{S}) = f'(\mathcal{S})$, $f(\mathcal{T}) = f'(\mathcal{T})$, $f(\mathcal{S} \cap \mathcal{T}) = f'(\mathcal{S} \cap \mathcal{T})$, and $f(\mathcal{S} \cup \mathcal{T}) = f'(\mathcal{S} \cup \mathcal{T})$, f is also submodular. Thus, it suffices to check the following three cases:

- Case 1: $f'(\mathcal{S} \cap \mathcal{T}) \leq C' \leq f'(\mathcal{S}) \leq f'(\mathcal{T}) \leq f'(\mathcal{S} \cup \mathcal{T})$.

  By definition of function f, we have

  $$f(\mathcal{S}) + f(\mathcal{T}) \geq C' + f'(\mathcal{S} \cap \mathcal{T}) = f(\mathcal{S} \cup \mathcal{T}) + f(\mathcal{S} \cap \mathcal{T}).$$

- Case 2: $f'(\mathcal{S} \cap \mathcal{T}) \leq f'(\mathcal{S}) \leq C' \leq f'(\mathcal{T}) \leq f'(\mathcal{S} \cup \mathcal{T})$.

  Since f' is monotonically increasing, we have

  $$\begin{aligned} f(\mathcal{S}) + f(\mathcal{T}) = f'(\mathcal{S}) + C' &\geq f'(\mathcal{S} \cap \mathcal{T}) + f(\mathcal{S} \cup \mathcal{T}) \\ &= f(\mathcal{S} \cap \mathcal{T}) + f(\mathcal{S} \cup \mathcal{T}). \end{aligned}$$

- Case 3: $f'(\mathcal{S} \cap \mathcal{T}) \leq f'(\mathcal{S}) \leq f'(\mathcal{T}) \leq C' \leq f'(\mathcal{S} \cup \mathcal{T})$.

www.manaraa.com

In this case, the submodularity of f' and the fact of $f' \leq f$ imply that

$$\begin{aligned} f(\mathcal{S}) + f(\mathcal{T}) = f'(\mathcal{S}) + f'(\mathcal{T}) &\geq f'(\mathcal{S} \cap \mathcal{T}) + f'(\mathcal{S} \cup \mathcal{T}) \\ &\geq f(\mathcal{S} \cap \mathcal{T}) + f(\mathcal{S} \cup \mathcal{T}). \end{aligned}$$

Hence, $f = \min\{C', f'\}$ is submodular, which completes the proof of Lemma B.1. □

Lemma B.2 *For any joint distribution* $\prod_{k=1}^{K} p\left(\mathbf{x}_k\right) \prod_{\ell=1}^{L} p\left(\mathbf{y}_\ell | \mathbf{x}_1^K\right) \prod_{\ell=1}^{L} p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$ *and fixed* $R \in \mathbb{R}$, *define the set function* $g : 2^{\mathcal{L}} \to \mathbb{R}$ *as:*

$$g\left(\mathcal{S}\right) := R + \sum_{\ell \in \mathcal{S}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c}\right),$$

and the corresponding non-negative set function $g^+ : 2^{\mathcal{L}} \to \mathbb{R}_+$ *as* $g^+ = \max\{g, 0\}$. *The functions* g *and* g^+ *are supermodular.*

Proof. We first prove that the set function $g'\left(\mathcal{T}\right) = I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{T}\right)$ is submodular. To this end, we evaluate

$$\begin{aligned} g'\left(\mathcal{T} \cap \mathcal{S}\right) + g'\left(\mathcal{T} \cup \mathcal{S}\right) &= I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{T} \cup \mathcal{S}}\right) + I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{T} \cap \mathcal{S}}\right) \\ &= I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{S}, \hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}}\right) + I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{T} \cap \mathcal{S}}\right) \\ &= g'\left(\mathcal{S}\right) + g'\left(\mathcal{T}\right) + I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_\mathcal{S}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_{\mathcal{T} \cap \mathcal{S}}\right) \end{aligned}$$

Furthermore,

$$\begin{aligned} & I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_\mathcal{S}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_{\mathcal{T} \cap \mathcal{S}}\right) \\ = \; & h\left(\hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_\mathcal{S}\right) - h\left(\hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_\mathcal{S}, \mathbf{X}_\mathcal{K}\right) - h\left(\hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_{\mathcal{T} \cap \mathcal{S}}\right) + h\left(\hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_{\mathcal{T} \cap \mathcal{S}}, \mathbf{X}_\mathcal{K}\right) \\ = \; & h\left(\hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_\mathcal{S}\right) - h\left(\hat{\mathbf{Y}}_{\mathcal{S}^c \cap \mathcal{T}} | \hat{\mathbf{Y}}_{\mathcal{S} \cap \mathrm{T}}\right) \leq 0. \end{aligned}$$

Therefore, $g'\left(\mathcal{T} \cap \mathcal{S}\right) + g'\left(\mathcal{T} \cup \mathcal{S}\right) \leq g'\left(\mathcal{S}\right) + g'\left(\mathcal{T}\right)$, which proves that g' is submodular.
In the following, we prove that g is supermodular. Evaluate $g(\mathcal{S}) + g(\mathcal{T})$ as

$$\begin{aligned} & g(\mathcal{S}) + g(\mathcal{T}) \\ = \; & 2R + \sum_{\ell \in \mathcal{S}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) + \sum_{\ell \in \mathcal{T}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{T}^c}\right) \\ \overset{(e)}{\leq} \; & 2R + \sum_{\ell \in \mathcal{S} \cup \mathcal{T}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) + \sum_{\ell \in \mathcal{S} \cap \mathcal{T}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{(\mathcal{S} \cap \mathcal{T})^c}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{(\mathcal{S} \cup \mathcal{T})^c}\right) \\ = \; & g(\mathcal{S} \cap \mathcal{T}) + g(\mathcal{S} \cup \mathcal{T}). \end{aligned}$$

where inequality (e) follows from the fact that $g'\left(\mathcal{T}\right) = I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{T}\right)$ is a submodular function.

Therefore, we show that g is supermodular. Following the result of [53, Lemma 6], it can be shown that $g^+ = \max\{g, 0\}$ is also supermodular. □

www.manaraa.com

Appendix C

Optimality of Successive Decoding for Maximizing Sum Rate

Similar to the proof of Theorem 2.1, Theorem 2.2 can also be proven using submodular optimization. In the following, we consider the region $(R, C_1, \ldots, C_L)$, and prove that joint decoding and successive decoding achieve the same maximum rate using the properties of submodular optimization.

Definition C.0.3. Define $\mathcal{P}_{JD}^s$ to be the closure of the convex hull of all $(R, C_1, \ldots, C_L)$ satisfying

$$R < \sum_{\ell \in \mathcal{S}} \left[C_\ell - I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) \right] + I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c}\right), \quad \forall\, \mathcal{S} \subseteq \mathcal{L}, \tag{C.1}$$

for some product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$.

Definition C.0.4. Define $\mathcal{P}_{SD}^s$ to be the closure of the convex hull all $(R, C_1, \ldots, C_L)$ satisfying

$$\begin{cases} R < I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_\mathcal{L}\right), \\ \sum\limits_{\ell \in \mathcal{S}} C_\ell > I\left(\mathbf{Y}_\mathcal{S}; \hat{\mathbf{Y}}_\mathcal{S} | \hat{\mathbf{Y}}_{\mathcal{S}^c}\right), \quad \forall\, \mathcal{S} \subseteq \mathcal{L} \end{cases} \tag{C.2}$$

for some product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$.

Note that $\mathcal{P}_{JD}^s$ represents the sum-rate and fronthaul-capacity region of joint decoding. All the partial sums over $\mathcal{S}$ in (C.1) can be strictly attained with equality depending on the values of the fronthaul capacities C_ℓ for $\ell = 1, \ldots, L$ and the sum rate R. Similarly, $\mathcal{P}_{SD}^s$ corresponds to the region of successive decoding. For fixed product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$, we say a point $(R, C_1, \ldots, C_L)$ is dominated by a point $(R', C_1', \ldots, C_L')$ in $\mathcal{P}_{SD}^s$ if $C_\ell' \leq C_\ell$ for $\ell = 1, \ldots, L$ and $R' \geq R$.

Clearly, $R_{JD,SUM}^* \geq R_{SD,SUM}^*$. To show $R_{JD,SUM}^* = R_{SD,SUM}^*$, it remains to show that $R_{JD,SUM}^* \leq R_{SD,SUM}^*$. For any given product distribution $\prod_{k=1}^K p\left(\mathbf{x}_k\right) \prod_{\ell=1}^L p(\hat{\mathbf{y}}_\ell | \mathbf{y}_\ell)$ and joint decoding sum rate R_{JD}, define $\mathcal{P}_C \subset \mathbb{R}_+^L$ to be the set of $(C_1, \ldots, C_L)$ such that

$$\sum_{\ell \in \mathcal{S}} C_\ell \geq \left[R_{JD} + \sum_{\ell \in \mathcal{S}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c}\right) \right]^+, \tag{C.3}$$

www.manaraa.com

for all $\mathcal{S} \subseteq \mathcal{L}$. Now, to show $R^*_{JD,SUM} \leq R^*_{SD,SUM}$, it suffices to show that each extreme point of $\mathcal{P}_C$ is dominated by a point in $\mathcal{P}^s_{SD}$ that achieves a sum rate greater or equal to the joint decoding sum rate R.

To this end, define a set function $g : 2^{\mathcal{L}} \to \mathbb{R}$ as follows:

$$g(\mathcal{S}) := R_{JD} + \sum_{\ell \in \mathcal{S}} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_{\mathcal{S}^c}\right),$$

for each $\mathcal{S} \subseteq \mathcal{L}$. It can be verified that the function $g^+(\mathcal{S}) = \max\{g(\mathcal{S}), 0\}$ is a supermodular function (see Appendix B, Lemma B.2). By construction, $\mathcal{P}_C$ is equal to the set of $(C_1, R_2, \ldots, C_L)$ satisfying

$$\sum_{\ell \in \mathcal{S}} C_\ell \geq g^+(\mathcal{S}), \quad \forall\, \mathcal{S} \subseteq \mathcal{L}.$$

Following the results in submodular optimization (Appendix B, Proposition B.1), we have that for a linear ordering $i_1 \prec i_2 \prec \cdots \prec i_K$ on the set $\mathcal{K}$, an extreme point of $\mathcal{P}_C$ can be computed as follows

$$\tilde{C}_{i_j} = g^+(\{i_1, \ldots, i_j\}) - g^+(\{i_1, \ldots, i_{j-1}\}).$$

All the $L!$ extreme points of $\mathcal{P}_C$ can be analyzed in the same manner. For notational simplicity we only consider the natural ordering $i_j = j$ in the following proof.

By construction,

$$\tilde{C}_j = \left[R_{JD} + \sum_{\ell=1}^{j} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}^L_{j+1}\right)\right]^+ - \left[R_{JD} + \sum_{\ell=1}^{j-1} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}^L_{j}\right)\right]^+.$$

Let j be the first index for which $g(\{1, \ldots, j\}) > 0$. Then, by construction,

$$\tilde{C}_k = I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}_k | \hat{\mathbf{Y}}^L_{k+1}\right) + I\left(\mathbf{Y}_k; \hat{\mathbf{Y}}_k | \mathbf{X}_\mathcal{K}\right) = I\left(\mathbf{Y}_k; \hat{\mathbf{Y}}_k | \hat{\mathbf{Y}}^L_{k+1}\right)$$

for all $k > j$, where the Markov chain $\hat{\mathbf{Y}}_i \leftrightarrow \mathbf{Y}_i \leftrightarrow \mathbf{X}_\mathcal{K} \leftrightarrow \mathbf{Y}_j \leftrightarrow \hat{\mathbf{Y}}_j$, for $i \neq j$, is utilized in deriving the second equality. Clearly, all the $\tilde{C}_k$'s are in the successive decoding region $\mathcal{P}^s_{SD}$.

Moreover, we have $g(\{1, \ldots, j'\}) \leq 0$ for all $j' < j$. Thus, $\tilde{C}_j$ can be expressed as

$$\begin{aligned} \tilde{C}_j &= R_{JD} + \sum_{\ell=1}^{j} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}^L_{j+1}\right) \\ &= \alpha I\left(\mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}^L_{j+1}\right) \end{aligned}$$

where $\alpha \in [0, 1]$ is defined as

$$\alpha = \frac{R_{JD} + \sum_{\ell=1}^{j} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_\mathcal{K}\right) - I\left(\mathbf{X}_\mathcal{K}; \hat{\mathbf{Y}}^L_{j+1}\right)}{I\left(\mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}^L_{j+1}\right)}.$$

Consider the two following successive decoding schemes:

- Scheme 1: The CP decodes quantization codewords $\hat{\mathbf{Y}}_{j+1}, \ldots, \hat{\mathbf{Y}}_L$ first, then decodes the message

www.manaraa.com

codewords $\mathbf{X}_{\mathcal{K}}$ sequentially. Note that the BSs with index $i \leq j$ are inactive, and are essentially removed from the network. The resulting extreme point $\mathbf{c}^{(1)} = (R_{SD}^{(1)}, C_1^{(1)}, \ldots, C_L^{(1)})$ of $\mathcal{P}_{SD}^s$ satisfies

$$\begin{cases} C_i^{(1)} = 0, & \text{for } i \leq j, \\ C_i^{(1)} = I\left(\mathbf{Y}_i; \hat{\mathbf{Y}}_i | \hat{\mathbf{Y}}_{i+1}^L\right) & \text{for } i > j, \\ R_{SD}^{(1)} = I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{j+1}^L\right). \end{cases}$$

- Scheme 2: The CP decodes quantization codewords $\hat{\mathbf{Y}}_j, \ldots, \hat{\mathbf{Y}}_L$ first, then decodes the message codewords $\mathbf{X}_{\mathcal{K}}$ sequentially. Note that in this scheme, the BSs with index $i < j$ are inactive, and are essentially removed from the network. The resulting extreme point $\mathbf{c}^{(2)} = (R_{SD}^{(2)}, C_1^{(2)}, \ldots, C_L^{(2)})$ of $\mathcal{P}_{SD}^s$ satisfies

$$\begin{cases} C_i^{(2)} = 0, & \text{for } i < j, \\ C_i^{(2)} = I\left(\mathbf{Y}_i; \hat{\mathbf{Y}}_i | \hat{\mathbf{Y}}_{i+1}^L\right) & \text{for } i \geq j, \\ R_{SD}^{(2)} = I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right). \end{cases}$$

Since C_ℓ is defined to be the maximum long-term average throughput of fronthaul link ℓ, the following point: $\mathbf{c}^\alpha = (1-\alpha)\mathbf{c}^{(1)} + \alpha \mathbf{c}^{(2)}$ lies in $\mathcal{P}_{SD}^s$. The corresponding sum rate R_{SD} in $\mathbf{c}^\alpha$ is given by

$$\begin{aligned}
(1-\alpha)R_{SD}^{(1)} + \alpha R_{SD}^{(2)} &= (1-\alpha) I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{j+1}^L\right) + \alpha I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right) \\
&\overset{(f)}{=} \frac{I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right) - R_{JD} - \sum_{\ell=1}^{j-1} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}\right)}{I\left(\mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}_{j+1}^L\right)} \cdot I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{j+1}^L\right) \\
&\quad + \frac{R_{JD} + \sum_{\ell=1}^{j} I\left(\mathbf{Y}_\ell; \hat{\mathbf{Y}}_\ell | \mathbf{X}_{\mathcal{K}}\right) - I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{j+1}^L\right)}{I\left(\mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}_{j+1}^L\right)} \cdot I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right) \\
&\geq \frac{R_{JD} \cdot \left[I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right) - I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{j+1}^L\right)\right] + I\left(\mathbf{Y}_j; \hat{\mathbf{Y}}_j | \mathbf{X}_{\mathcal{K}}\right) \cdot I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right)}{I\left(\mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}_{j+1}^L\right)} \\
&\overset{(g)}{\geq} R_{JD} \cdot \frac{I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right) - I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_{j+1}^L\right) + I\left(\mathbf{Y}_j; \hat{\mathbf{Y}}_j | \mathbf{X}_{\mathcal{K}}\right)}{I\left(\mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}_{j+1}^L\right)} \\
&= R_{JD}, \qquad \text{(C.4)}
\end{aligned}$$

where the equality (f) follows from the fact that $I\left(\mathbf{X}_{\mathcal{K}}, \mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}_{j+1}^L\right) = I\left(\mathbf{Y}_{j+1}; \hat{\mathbf{Y}}_{j+1} | \hat{\mathbf{Y}}_{j+1}^L\right)$, and inequality (g) follows from the fact that $R_{JD} \leq I\left(\mathbf{X}_{\mathcal{K}}; \hat{\mathbf{Y}}_j^L\right)$.

Therefore, for every extreme point $(\tilde{C}_1, \ldots, \tilde{C}_L)$ of $\mathcal{P}_C$, the point $(R_{JD}, \tilde{C}_1, \ldots, \tilde{C}_L)$ is dominated by a point in $\mathcal{P}_{SD}^s$. This proves Theorem 2.2.

www.manaraa.com

Appendix D

Constant-gap Result for Compress-and-Forward with Joint Decoding

The idea of the proof is to compare the achievable rate of compress-and-forward with joint decoding with the following cut-set upper bound [41]

$$\sum_{k\in\mathcal{T}} R_k \leq \min\left\{\sum_{\ell\in\mathcal{S}} C_\ell + \log\frac{\left|\sum_{\ell\in\mathcal{S}^c}\mathbf{H}_{\ell,\mathcal{T}}^\dagger\boldsymbol{\Sigma}_\ell^{-1}\mathbf{H}_{\ell,\mathcal{T}}+\mathbf{K}_{\mathcal{T}}^{-1}\right|}{\left|\mathbf{K}_{\mathcal{T}}^{-1}\right|}\right\} \tag{D.1}$$

for all $\emptyset\subset\mathcal{T}\subseteq\mathcal{K}$ and $\mathcal{S}\subseteq\mathcal{L}$. In the expression of cut-set bound, the first term represents the cut across the fronthaul links in set $\mathcal{S}$, and the second term represents the cut from the users to the BSs in set $\mathcal{S}^c$.

Recall that the rate region for joint decoding (2.19) under Gaussian quantization is the of $(R_1,\cdots,R_K)$ such that

$$\sum_{k\in\mathcal{T}} R_k < \sum_{\ell\in\mathcal{S}}\left[C_\ell - \log\frac{|\boldsymbol{\Sigma}_\ell^{-1}|}{|\boldsymbol{\Sigma}_\ell^{-1}-\mathbf{B}_\ell|}\right] + \log\frac{\left|\sum_{\ell\in\mathcal{S}^c}\mathbf{H}_{\ell,\mathcal{T}}^\dagger\mathbf{B}_\ell\mathbf{H}_{\ell,\mathcal{T}}+\mathbf{K}_{\mathcal{T}}^{-1}\right|}{\left|\mathbf{K}_{\mathcal{T}}^{-1}\right|}$$

for all $\emptyset\subset\mathcal{T}\subseteq\mathcal{K}$ and $\mathcal{S}\subseteq\mathcal{L}$, for some $0\preceq\mathbf{B}_\ell\preceq\boldsymbol{\Sigma}_\ell^{-1}$. We now show that if a rate tuple $(R_1,\cdots,R_K)$ is within the cut-set bound, then $(R_1-\eta,\cdots,R_K-\eta)$ is in the achievable rate region of joint decoding, where

$$|\mathcal{T}|\eta \leq \sum_{\ell\in\mathcal{S}}\log\frac{|\boldsymbol{\Sigma}_\ell^{-1}|}{|\boldsymbol{\Sigma}_\ell^{-1}-\mathbf{B}_\ell|} + \log\frac{\left|\sum_{\ell\in\mathcal{S}^c}\mathbf{H}_{\ell,\mathcal{T}}^\dagger\boldsymbol{\Sigma}_\ell^{-1}\mathbf{H}_{\ell,\mathcal{T}}+\mathbf{K}_{\mathcal{T}}^{-1}\right|}{\left|\sum_{\ell\in\mathcal{S}^c}\mathbf{H}_{\ell,\mathcal{T}}^\dagger\mathbf{B}_\ell\mathbf{H}_{\ell,\mathcal{T}}+\mathbf{K}_{\mathcal{T}}^{-1}\right|}$$

is the gap between the cut-set bound and achievable rate of joint decoding.

Choose quantization noise level to be at the background noise level, i.e. $\mathbf{Q}_\ell=\boldsymbol{\Sigma}_\ell$. Then we have

$$\mathbf{B}_\ell = (\boldsymbol{\Sigma}_\ell+\mathbf{Q}_\ell)^{-1} = \frac{1}{2}\boldsymbol{\Sigma}_\ell^{-1}.$$

www.manaraa.com

Evaluate gap η with the above choice of $\mathbf{B}_\ell$ gives

$$\eta \leq \frac{|\mathcal{S}|}{|\mathcal{T}|} \cdot N + M \leq NL + M,$$

which completes the proof of Proposition 2.3.

www.manaraa.com

Appendix E

Constant-gap Result for the VMAC-WZ scheme

The idea is to choose $q_i = \alpha\sigma_i^2$, $i = 1, 2, \ldots, L$ where $\alpha > 0$ is an appropriately chosen constant, then compare the achievable rate of VMAC-WZ with the following cut-set like sum-capacity upper bound [13]

$$\bar{C} = \min\left\{\log\frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + \mathrm{diag}(\sigma_i^2)\right|}{|\mathrm{diag}(\sigma_i^2)|}, C\right\} \tag{E.1}$$

where the first term is the cut from the users to the BSs, and the second term is the cut across the fronthaul links.

We choose the quantization level α depending on C as follows: When $C \geq \log\frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}+2\mathrm{diag}(\sigma_i^2)\right|}{|\mathrm{diag}(\sigma_i^2)|}$, we choose $\alpha = 1$, i.e., the quantization noise levels are set to be at the background noise levels. Since $\alpha = 1$, it can be verified that

$$I(\mathbf{Y}; \hat{\mathbf{Y}}) = \log\frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + 2\mathrm{diag}(\sigma_i^2)\right|}{|\mathrm{diag}(\sigma_i^2)|}. \tag{E.2}$$

Thus, we have $C \geq I(\mathbf{Y}; \hat{\mathbf{Y}})$. This implies that the sum fronthaul constraint (3.2) is satisfied. Therefore, the sum rate

$$R_{sum} = I(\mathbf{X}; \hat{\mathbf{Y}}) = \log\frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + 2\mathrm{diag}(\sigma_i^2)\right|}{|2\mathrm{diag}(\sigma_i^2)|} \tag{E.3}$$

is achievable. In this case, the gap between $\bar{C}$ and R_{sum} can be bounded by

$$\begin{aligned}\bar{C} - R_{sum} &\leq \log\frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + \mathrm{diag}(\sigma_i^2)\right|}{|\mathrm{diag}(\sigma_i^2)|} - \log\frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + 2\mathrm{diag}(\sigma_i^2)\right|}{|2\mathrm{diag}(\sigma_i^2)|} \\ &< L.\end{aligned}$$

When $C < \log\frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}+2\mathrm{diag}(\sigma_i^2)\right|}{|\mathrm{diag}(\sigma_i^2)|}$, we choose α so that $I(\mathbf{Y}; \hat{\mathbf{Y}}) = C$. First, note that for such a

www.manaraa.com

choice of α the sum rate $R_{sum} = I(\mathbf{X};\hat{\mathbf{Y}})$ is achievable. Next, observe that

$$I(\mathbf{Y};\hat{\mathbf{Y}}) = \log \frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + \mathrm{diag}(\sigma_i^2) + \alpha\mathrm{diag}(\sigma_i^2)\right|}{|\alpha\mathrm{diag}(\sigma_i^2)|} \tag{E.4}$$

is a monotonically decreasing function of α. Since $C = I(\mathbf{Y};\hat{\mathbf{Y}}) < \log \frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}+2\mathrm{diag}(\sigma_i^2)\right|}{\left|\mathrm{diag}(\sigma_i^2)\right|}$, we have $\alpha > 1$. Now, we use $C = I(\mathbf{Y};\hat{\mathbf{Y}})$ as an upper bound. The gap between $\bar{C}$ and R_{sum} can now be bounded by

$$\begin{aligned}
\bar{C} - R_{sum} &\leq I(\mathbf{Y};\hat{\mathbf{Y}}) - I(\mathbf{X};\hat{\mathbf{Y}}) \\
&= \log \frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + (1+\alpha)\mathrm{diag}(\sigma_i^2)\right|}{|\alpha\mathrm{diag}(\sigma_i^2)|} - \log \frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + (1+\alpha)\mathrm{diag}(\sigma_i^2)\right|}{|(1+\alpha)\mathrm{diag}(\sigma_i^2)|} \\
&= L\log\left(1+\frac{1}{\alpha}\right) < L
\end{aligned}$$

where the last inequality follows from the fact that $\alpha > 1$.

Combining the two cases, we see that the gap to the sum capacity for the VMAC-WZ scheme with appropriately chosen quantization noise levels (which are proportional to the background noise levels) is always less than 1 bit per user per channel use.

www.manaraa.com

Appendix F

Constant-gap Result for the VMAC-SU scheme

Lemma F.1 *For fixed $\kappa > 1$, suppose that a $n \times n$ matrix $\mathbf{\Psi}$ is κ-strictly diagonally dominant, then*

$$|\mathbf{\Psi}| \geq \left(1 - \frac{1}{\kappa}\right)^n \prod_{i=1}^{n} |\mathbf{\Psi}(i,i)|. \tag{F.1}$$

Proof. The proof follows from the lower bound given in [85], which shows that if $\mathbf{\Psi}$ is strictly diagonally dominant, i.e. $|\mathbf{\Psi}(i,i)| > \sum_{j\neq i}^{n} |\mathbf{\Psi}(i,j)|$ for $i = 1, \ldots, n$, then the determinant of $\mathbf{\Psi}$ can be bounded from below as follows,

$$|\mathbf{\Psi}| \geq \prod_{i=1}^{n} \left(|\mathbf{\Psi}(i,i)| - \sum_{j\neq i}^{n} |\mathbf{\Psi}(i,j)| \right). \tag{F.2}$$

Under the condition that $\mathbf{\Psi}$ is κ-strictly diagonally dominant, i.e. $\sum_{j\neq i}^{n} |\mathbf{\Psi}(i,j)| \leq \frac{|\mathbf{\Psi}(i,i)|}{\kappa}$ we further bound $|\mathbf{\Psi}|$ by

$$|\mathbf{\Psi}| \geq \prod_{i=1}^{n} \left(|\mathbf{\Psi}(i,i)| - \frac{|\mathbf{\Psi}(i,i)|}{\kappa} \right) = \left(1 - \frac{1}{\kappa}\right)^n \prod_{i=1}^{n} |\mathbf{\Psi}(i,i)|,$$

which completes the proof. □

We now prove Theorem 3.3. The proof uses the same technique as in that of Theorem 3.2. We first choose the quantization noise levels $q_i = \alpha\sigma_i^2$, $i = 1, 2, \ldots, L$, where $\alpha > 0$ is a constant depending on C, then compare the achievable rate of the VMAC-SU scheme with the following cut-set like upper bound [13]

$$\bar{C} = \min \left\{ \log \frac{\left| \mathbf{H}_{\mathcal{L}} \mathbf{K}_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger} + \mathrm{diag}(\sigma_i^2) \right|}{|\mathrm{diag}(\sigma_i^2)|}, C \right\}. \tag{F.3}$$

We consider two different cases as follows: when $C \geq \log \frac{\left|\mathrm{diag}\left(\mathbf{H}_{\mathcal{L}} \mathbf{K}_{\mathcal{L}} \mathbf{H}_{\mathcal{L}}^{\dagger}\right) + 2\mathrm{diag}(\sigma_i^2)\right|}{\left|\mathrm{diag}(\sigma_i^2)\right|}$, i.e. the sum fronthaul capacity is large enough to support the choice of $q_i = \sigma_i^2$, we choose $\alpha = 1$. In this case, the

www.manaraa.com

gap between $\bar{C}$ and R_{sum} can be bounded by

$$\begin{aligned}\bar{C} - R_{sum} &\leq \log \frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + \text{diag}(\sigma_i^2)\right|}{|\text{diag}(\sigma_i^2)|} - \log \frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + 2\text{diag}(\sigma_i^2)\right|}{|2\text{diag}(\sigma_i^2)|} \\ &< L.\end{aligned}$$

When $C < \log \frac{\left|\text{diag}\left(\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}\right)+2\text{diag}(\sigma_i^2)\right|}{\left|\text{diag}(\sigma_i^2)\right|}$, we choose α so that $\sum_{i=1}^{L} I(Y_i;\hat{Y}_i) = C$. First, notice that

$$\sum_{i=1}^{L} I(Y_i;\hat{Y}_i) = \log \frac{\left|\text{diag}\left(\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}\right) + (1+\alpha)\text{diag}(\sigma_i^2)\right|}{|\alpha\text{diag}(\sigma_i^2)|}$$

is a monotonically decreasing function of α. Since $C = \sum_{i=1}^{L} I(Y_i;\hat{Y}_i) < \log \frac{\left|\text{diag}\left(\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}\right)+2\text{diag}(\sigma_i^2)\right|}{\left|\text{diag}(\sigma_i^2)\right|}$, we have $\alpha > 1$. Now, we use $C = \sum_{i=1}^{L} I(Y_i;\hat{Y}_i)$ as an upper bound. Let $\mathbf{\Psi} = \mathbf{H}K_X\mathbf{H}^H\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + (1+\alpha)\text{diag}(\sigma_i^2)$ and note that $\mathbf{\Psi}(i,i) \geq 0$. The gap between $\bar{C}$ and R_{sum} is bounded by

$$\begin{aligned}\bar{C} - R_{sum} &\leq \sum_{i=1}^{L} I(Y_i;\hat{Y}_i) - I(\mathbf{X};\widehat{\mathbf{Y}}) \\ &= \log \frac{\left|\text{diag}\left(\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger}\right) + (1+\alpha)\text{diag}(\sigma_i^2)\right|}{|\alpha\text{diag}(\sigma_i^2)|} - \log \frac{\left|\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + (1+\alpha)\text{diag}(\sigma_i^2)\right|}{|(1+\alpha)\text{diag}(\sigma_i^2)|} \\ &= \log \left[\left(1+\frac{1}{\alpha}\right)^{L} \frac{\prod_{i=1}^{L}\mathbf{\Psi}(i,i)}{|\mathbf{\Psi}|}\right].\end{aligned}$$

Since matrix $\mathbf{H}_{\mathcal{L}}\mathbf{K}_{\mathcal{L}}\mathbf{H}_{\mathcal{L}}^{\dagger} + \text{diag}(\sigma_i^2)$ is κ-strictly diagonally dominant, $\mathbf{\Psi}$ is also κ-strictly diagonally dominant. Following the result of Lemma F.1, we further bound the gap as follows,

$$\begin{aligned}\bar{C} - R_{sum} &\leq L\log\left(1+\frac{1}{\alpha}\right) + \sum_{i=1}^{L}\log\frac{\kappa}{\kappa-1} \\ &< L\left(1+\log\frac{\kappa}{\kappa-1}\right),\end{aligned}$$

where the last inequality follows from the fact that $\alpha > 1$.

Combining the two cases, we see that the gap to sum capacity for the VMAC-SU scheme with quantization noise levels proportional to the background noise levels is always less than $1 + \log\frac{\kappa}{\kappa-1}$ per user per channel use.

www.manaraa.com

Appendix G

Convergence of WMMSE-SCA Algorithm

In this appendix, we provide the convergence proof of WMMSE-SCA algorithm. The proof is a direct application of the convergence result of the successive convex approximation algorithm [77]. Let $\mathbf{V} = \text{diag}\left(\{\mathbf{V}_k\}_{k=1}^K\right)$. Define the objective function and fronthaul constraints in problem (4.10) to be

$$f(\mathbf{V},\mathbf{Q}) = \sum_{k=1}^{K} \alpha_k \log\left|\mathbf{I} + \mathbf{V}_k^\dagger \mathbf{H}_{\mathcal{L},k}^\dagger \mathbf{D}_k^{-1} \mathbf{H}_{\mathcal{L},k} \mathbf{V}_k\right|,$$
$$g_\ell(\mathbf{V},\mathbf{Q}) = \log \frac{\left|\sum_{k=1}^K \mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell\right|}{|\mathbf{Q}_\ell|} - C_\ell,$$

where $\mathbf{D}_k = \sum_{j\neq k}^K \mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^\dagger\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}$ for the linear receiver or $\mathbf{D}_k = \sum_{j>k}^K \mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^\dagger\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}$ for the SIC receiver.

At the tth iteration, assume that the output of WMMSE-SCA algorithm is $(\mathbf{V}^t, \mathbf{Q}^t)$. Putting $(\mathbf{V}^t, \mathbf{Q}^t)$ into equations (4.13) and (4.17) gives

$$\mathbf{\Gamma}_\ell^t = \sum_{k=1}^K \mathbf{H}_{\ell,k}\mathbf{V}_k^t(\mathbf{V}_k^t)^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell^t,$$
$$\mathbf{W}_k^t = \mathbf{I} + \mathbf{H}_{\mathcal{L},k}^\dagger(\mathbf{V}_k^t)^\dagger\mathbf{U}_k^t,$$

where

$$\mathbf{U}_k^t = \left(\sum_{j\neq k}\mathbf{H}_{\mathcal{L},j}\mathbf{V}_j^t(\mathbf{V}_j^t)^\dagger\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}^t\right)^{-1}\mathbf{H}_{\mathcal{L},k}\mathbf{V}_k^t.$$

Then the objective function and fronthaul constraints in problem (4.19) can be written as

$$\tilde{f}(\{\mathbf{V},\mathbf{Q}\},\{\mathbf{V}^t,\mathbf{Q}^t\}) = \sum_{k=1}^K \alpha_k\left(\log|\mathbf{W}_k^t| - \text{Tr}\left(\mathbf{W}_k^t\mathbf{E}_k\right)\right) + \rho\sum_{\ell=1}^L\left\|\mathbf{\Gamma}_\ell^t - \mathbf{\Omega}_\ell\right\|_F^2,$$
$$\tilde{g}_\ell(\{\mathbf{V},\mathbf{Q}\},\{\mathbf{V}^t,\mathbf{Q}^t\}) = \log\left|\mathbf{\Gamma}_\ell^t\right| + \text{Tr}\left((\mathbf{\Gamma}_\ell^t)^{-1}\mathbf{\Omega}_\ell\right) - \log|\mathbf{Q}_\ell| - C_\ell - N,$$

www.manaraa.com

where

$$\mathbf{E}_k = \left(\mathbf{I} - (\mathbf{U}_k^t)^\dagger \mathbf{H}_{\mathcal{L},k}\mathbf{V}_k\right)\left(\mathbf{I} - (\mathbf{U}_k^t)^\dagger \mathbf{H}_{\mathcal{L},k}\mathbf{V}_k\right)^\dagger + (\mathbf{U}_k^t)^\dagger \left(\sum_{j\neq k}^{K} \mathbf{H}_{\mathcal{L},j}\mathbf{V}_j\mathbf{V}_j^\dagger\mathbf{H}_{\mathcal{L},j}^\dagger + \mathbf{\Sigma} + \mathbf{Q}\right)\mathbf{U}_k^t,$$

and $\mathbf{\Omega}_\ell = \sum_{k=1}^{K} \mathbf{H}_{\ell,k}\mathbf{V}_k\mathbf{V}_k^\dagger\mathbf{H}_{\ell,k}^\dagger + \mathbf{\Sigma}_\ell + \mathbf{Q}_\ell$.

We now observe that the WMMSE-SCA algorithm is actually a special case of the general successive convex approximation (SCA) method, with $\tilde{f}$ and $\tilde{g}_\ell$ being the convex approximation functions of f and g_ℓ respectively. Furthermore, it is easy to verify that $\tilde{f}$ is strictly convex over $(\mathbf{V}, \mathbf{Q})$. Following the result of [86, Lemma 3.1], it can be shown that $\tilde{f}$ is uniformly strongly convex over $(\mathbf{V}, \mathbf{Q})$. Applying the convergence result of the SCA algorithm [77, Theorem 2], we prove that each of the limit points generated by the proposed WMMSE-SCA algorithm is a stationary point of problem (4.10). This completes the proof of Theorem 4.1.

www.manaraa.com

Bibliography

[1] 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.

[2] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks – A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 2014.

[3] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Netw.*, to apprear 2015. [Online]. Available: http://arxiv.org/pdf/1512.07743v1.pdf

[4] H. Al-Raweshidy and S. Komaki, *Radio over fiber technologies for mobile communications networks*. Artech House, 2002.

[5] D. M. Pozar, *Microwave and RF design of wireless systems*. John Wiley & Sons, Inc., 2000.

[6] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[7] Ericsson AB, "Microwave towards 2020," *Ericsson Technical Report*, Sep. 2015. [Online]. Available: www.ericsson.com/res/docs/2015/microwave-2020-report.pdf

[8] O. Somekh and S. Shamai, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1401–1425, Jul. 2000.

[9] E. Katranaras, M. Imran, C. Tzaras *et al.*, "Uplink capacity of a variable density cellular system with multicell processing," *IEEE Trans. Commun.*, vol. 57, no. 7, pp. 2098–2108, Jul. 2009.

[10] H. Dai and H. V. Poor, "Asymptotic spectral efficiency of multicell MIMO systems with frequency-flat fading," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2976–2988, Nov. 2003.

[11] S. Chatzinotas, M. A. Imran, and R. Hoshyar, "On the multicell processing capacity of the cellular MIMO uplink channel in correlated rayleigh fading environment," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3704–3715, Jul. 2009.

[12] A. Sanderovich, S. Shamai, Y. Steinberg, and G. Kramer, "Communication via decentralized processing," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3008–3023, Jul. 2008.

المنارة للاستشارات
www.manaraa.com

[13] A. Sanderovich, O. Somekh, H. V. Poor, and S. Shamai, "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3457–3478, Aug. 2009.

[14] A. Sanderovich, S. Shamai, and Y. Steinberg, "Distributed MIMO receiver–Achievable rates and upper bounds," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4419–4438, Oct. 2009.

[15] A. Avestimehr, S. Diggavi, and D. Tse, "Wireless network information flow: A deterministic approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1872–1905, Apr. 2011.

[16] S. H. Lim, Y.-H. Kim, A. El Gamal, and S.-Y. Chung, "Noisy network coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 3132–3152, May 2011.

[17] M. H. Yassaee and M. R. Aref, "Slepian–Wolf coding over cooperative relay networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3462–3482, Jun. 2011.

[18] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[19] A. del Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4698–4709, Sep. 2009.

[20] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[21] J. Hoydis, M. Kobayashi, and M. Debbah, "Optimal channel training in uplink network MIMO systems," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2824–2833, Jun. 2011.

[22] P. Marsch and G. Fettweis, "Uplink CoMP under a constrained backhaul and imperfect channel knowledge," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1730–1742, Jun. 2011.

[23] J. Kang, O. Simeone, J. Kang, and S. S. Shitz, "Joint signal and channel state information compression for the backhaul of uplink network MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1555–1567, Mar. 2014.

[24] R. Karasik, O. Simeone, and S. S. Shitz, "Robust uplink communications over fading channels with variable backhaul connectivity," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5788–5799, Nov. 2013.

[25] C. Fan, Y. J. Zhang, and X. Yuan, "Scalable coordinated uplink processing in cloud radio access networks," in *Prof. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 3591–3596.

[26] X. Rao and V. K. Lau, "Distributed fronthaul compression and joint signal recovery in Cloud-RAN," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1056–1065, Feb. 2015.

[27] W. Wang, V. K. Lau, and M. Peng, "Delay-optimal fronthaul allocation via perturbation analysis in cloud radio access networks," in *Proc. IEEE International Conf. Comm. (ICC)*, Jun. 2015, pp. 3999–4004.

[28] E. Heo, O. Simeone, and H. Park, "Optimal fronthaul compression for synchronization in the uplink of cloud radio access networks," Oct. 2015. [Online]. Available: http://arxiv.org/abs/1510.01545

www.manaraa.com

[29] B. Nazer, A. Sanderovich, M. Gastpar, and S. Shamai, "Structured superposition for backhaul constrained cellular uplink," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2009, pp. 1530–1534.

[30] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227–5243, Sep. 2013.

[31] K. N. Pappi, G. K. Karagiannidis, and R. Schober, "How sensitive is compute-and-forward to channel estimation errors?" in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 3110–3114.

[32] C. Feng, D. Silva, and F. R. Kschischang, "Blind compute-and-forward," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2012, pp. 403–407.

[33] J. Hou and G. Kramer, "Short message noisy network coding with a decode-forward option," *IEEE Trans. Inf. Theory*, to appear, 2015. [Online]. Available: http://arxiv.org/abs/1304.1692

[34] L. Zhou and W. Yu, "Uplink multicell processing with limited backhaul via per-base-station successive interference cancellation," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 1981–1993, Oct. 2013.

[35] Y. Zhou, Y. Xu, J. Chen, and W. Yu, "Optimality of Gaussian fronthaul compression for uplink MIMO cloud radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2241–2245.

[36] Y. Zhou and W. Yu, "Approximate bounds for limited backhaul uplink multicell processing with single-user compression," in *Proc. IEEE Canadian Workshop Inf. Theory (CWIT)*, Jun. 2013, pp. 113–116.

[37] Y. Zhou, W. Yu, and D. Toumpakaris, "Uplink multi-cell processing: Approximate sum capacity under a sum backhaul constraint," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2013, pp. 569–573.

[38] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

[39] ——, "Optimized beamforming and backhaul compression for uplink MIMO cloud radio access networks," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2014, pp. 1487–1492.

[40] ——, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE Trans. Signal Process.*, submitted, 2015.

[41] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[42] X. Wu and L.-L. Xie, "On the optimal compressions in the compress-and-forward relay schemes," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2613–2628, May 2013.

[43] C. Tian and J. Chen, "Remote vector Gaussian source coding with decoder side information under mutual information and distortion constraints," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4676–4680, Oct. 2009.

www.manaraa.com

[44] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem [multiterminal source coding]," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996.

[45] Y. Oohama, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2577–2593, Jul. 2005.

[46] J. Wang and J. Chen, "Vector Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5533–5552, Sep. 2014.

[47] E. Ekrem and S. Ulukus, "An outer bound for the vector Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6870–6887, Nov. 2014.

[48] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2000, pp. 368–377.

[49] T. Liu and P. Viswanath, "An extremal inequality motivated by multiterminal information-theoretic problems," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1839–1851, May 2007.

[50] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[51] T. M. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 572–584, Sep. 1979.

[52] A. El Gamal, M. Mohseni, and S. Zahedi, "Bounds on capacity and minimum energy-per-bit for AWGN relay channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1545–1561, Apr. 2006.

[53] T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.

[54] D. N. Tse and S. V. Hanly, "Multiaccess fading channels – Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[55] A. Dembo, T. Cover, and J. Thomas, "Information theoretic inequalities," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1501–1518, Nov. 1991.

[56] D. Palomar and S. Verdu, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 141–154, Jan. 2006.

[57] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[58] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, 2014.

[59] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

www.manaraa.com

[60] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels (corresp.)," *IEEE Trans. Inf. Theory*, vol. 19, no. 3, pp. 357–359, May 1973.

[61] O. Somekh, B. M. Zaidel, and S. Shamai, "Sum rate characterization of joint multiple cell-site processing," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4473–4497, Dec. 2007.

[62] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Local base station cooperation via finite-capacity links for the uplink of linear cellular networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 190–204, Jan. 2009.

[63] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, Feb. 2013.

[64] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, Jan. 2004.

[65] Q. Li, M. Hong, H.-T. Wai, Y.-F. Liu, W.-K. Ma, and Z.-Q. Luo, "Transmit solutions for MIMO wiretap channels using alternating optimization," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1714–1727, Sep. 2013.

[66] M. Hong, Q. Li, and Y.-F. Liu, "Decomposition by successive convex approximation: A unifying approach for linear transceiver design in interfering heterogeneous networks," *IEEE Trans. Wireless Commun.*, to appear, 2015. [Online]. Available: http://arxiv.org/abs/1210.1507

[67] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, no. 3, pp. 127–136, Mar. 2000.

[68] A. Beck and L. Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2037–2060, Apr. 2013.

[69] 3GPP, "Coordinated multi-point operation for LTE physical layer aspects," 3rd Generation Partnership Project (3GPP), TR 36.819, Sep. 2011. [Online]. Available: http://www.qtc.jp/3GPP/Specs/36819-b10.pdf

[70] C. T. K. Ng and H. Huang, "Linear precoding in cooperative MIMO cellular networks with limited coordination clusters," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, Sep. 2010.

[71] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[72] S. S. Christensen, R. Agarwal, E. Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Jul. 2008.

[73] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

www.manaraa.com

[74] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[75] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Jun. 2015. [Online]. Available: http://cvxr.com/cvx/doc/CVX.pdf

[76] D. P. Bertsekas, *Nonlinear programming*, 2nd ed. Athena scientific, 1999.

[77] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Distributed methods for constrained nonconvex multi-agent optimization-part I: Theory," *submitted to IEEE Trans. Signal Process.*, 2014. [Online]. Available: http://arxiv.org/abs/1410.4754

[78] F. A. Potra and S. J. Wright, "Interior-point methods," *J. Comput. Appl. Math.*, vol. 124, no. 1–2, pp. 281–302, Dec. 2000.

[79] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 2, pp. 499–533, 1998.

[80] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," in *Proc. IEEE Inf. Theory and Appl. (ITA) Workshop*, Feb. 2014, pp. 1–6.

[81] S. H. Lim, K. T. Kim, and Y.-H. Kim, "Distributed decode-forward for relay networks," *IEEE Trans. Inf. Theory*, Submitted, 2015. [Online]. Available: http://arxiv.org/pdf/1510.00832v1.pdf

[82] S. Fujishige, *Submodular functions and optimization*, 2nd ed. Elsevier, 2005.

[83] S. Iwata, "Submodular function minimization," *Math. Program.*, vol. 112, no. 1, pp. 45–64, 2008.

[84] X. Zhang, J. Chen, S. B. Wicker, and T. Berger, "Successive coding in multiuser information theory," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2246–2254, 2007.

[85] A. M. Ostrowski, "Note on bounds for determinants with dominant principal diagonal," *Proc. Amer. Math. Soc.*, vol. 3, no. 1, pp. 26–30, 1952.

[86] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, May 2010.

www.manaraa.com