UNIVERSITY OF CALIFORNIA SAN DIEGO

Access Point Placement in Massive Multiple-Input-Multiple-Output Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Communication Theory and Systems)

by

Govind Ravikumar Gopal

Committee in charge:

Professor Bhaskar D. Rao, Chair
Professor Florian Meyer
Professor Lawrence K. Saul
Professor Kenneth A. Zeger
Professor Xinyu Zhang

2023

The dissertation of Govind Ravikumar Gopal is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents
whose unconditional love and support over the years
made this dissertation possible.

EPIGRAPH

Humor is the universal solvent against the abrasive elements of life.

—Alan Simpson

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Preparing for the PhD degree for the last nearly seven years has been a long and arduous task. As I finish up my degree and write this dissertation, I can think of the ups and downs I have faced during this journey. However, I do feel an immense sense of accomplishment. This has only been possible with the help and support of a lot of people and it is only right that I acknowledge them here.

Perhaps, my first introduction to the world of communication was in my undergraduate studies at the National Institute of Technology Karnataka, Surathkal, India. I had a great group of teachers including Prof. Arulalan Rajan, Prof. Raghavendra Bobbi, Prof. Deepu Vijayasenan, and Prof. Narasimhadhan. Prof. Narasimhadhan is instrumental in starting my journey down the area of wireless communications. When I was searching for internship opportunities for the summer following my second year, I asked Prof. Narasimhadhan to recommend me to someone in his alma mater. He chose Prof. Chandra R. Murthy at the Department of Electrical Communication Engineering at the Indian Institute of Science (IISc), Bangalore, India. Prof. Murthy, without hesitation, took me as an intern that year and for the next two summers as well. He taught me not only about the current research problems in wireless communications that his group are interested in, but how to be a good researcher. Through the three internship experiences under him, I learned to document my findings well and my writing skills were refined - he alone is responsible for my learning LaTeX, essential for all the journal and conference paper writings! I strongly believe that the recommendation letter by Prof. Murthy played a significant part in my admission to UCSD. I cannot thank Prof. Murthy enough for the support and mentorship during my undergraduate studies.

When I arrived at UCSD to begin my MS degree, I knew that I wanted to pursue PhD after my MS, but was unsure whether I would find the right professor here or at another university. Having been Prof. Murthy's advisor during his time as a PhD student at UCSD, I approached (in the first week of my classes) Prof. Bhaskar D. Rao. He was quick to point out that I should spend my first quarter adjusting to the classes and then only think about doing research. He was also right about Information Theory being a difficult course! After taking his Array Processing course two quarters later, I started as a researcher in wireless systems with him through the summer and ultimately, he accepted me as a PhD student. Over the last 40 years in UCSD, Prof. Rao has imparted immense wisdom and knowledge to all those who have had the honor of being his student as well as anyone who has interacted with him. Two of his virtues that have stood out to me are patience and generosity. From his busy schedule, he always would take the time to meet and discuss my work. Even when I would go wrong either in my analysis or my simulations, Prof. Rao would patiently point me in the correct path. Words just cannot describe my deep gratitude to him.

Of course, other than my advisor, I am lucky to have had excellent lab mates, Rohan, Aditya, Kuan-Lin, and Hitesh. Although, Maher, Jing, Yacong, David, and Elina had graduated, I am sure they had an impact on my work, especially Elina, whose work on placement I continued. Prof. Piya Pal and her students, Pulak, Mehmet, Sina, Robin, Jiawen, and Parth were also very supportive during my time at UCSD. We would often go for coffee breaks at Sunshine Market together. Another collaborator and friend with whom I have had long discussions both technical and personal has been Gabriel from the National Institute of Information and Communications Technology (NICT) in Japan. He has also inspired me to

think differently to obtain solutions to problems and improve my writing skills.

I am also very grateful and lucky to have made very good connections through the three internships during my PhD tenure. My first opportunity was provided by Mike and his manager Dragan at Nokia Bell Labs in the summer of 2020. Although I could not physically go to the prestigious location where the transistor was invented, Mike and I would have frequent meetings to keep updated with the work. My second internship opportunity in the summer of 2021 was provided to me at the Next Generation and Standards Group at Intel by Daewon, my manager, and Debdeep and Toufiq, my mentors. They made me feel very welcome to their group and gave me a very good introduction to the world of wireless cellular standards. My third and final internship in the summer of 2022 was on-site at the Nokia Standards location in Naperville, Illinois under Rapeepat, Mark, and Amitava. All the people I met there made me feel very welcome into the group - Jie, Bill, Jun, Serdar, Qiping, Preetish, Yunchao, Hyunsu, Richa, and Ryan. Of course, I could not have completed the internship without the support of my fellow intern and gossip buddy, Shuvam. My stay and well-being at Naperville was due to Gitana opening her house to me and sharing her dog Toby with whom I think I have made a life-long connection!

My labmates Rohan and Aditya have also been very close friends and have supported me throughout my PhD, along with Rohan's wife Sukanya, Anwesan, Ajay, Sashank and his wife Medha, Tharun, Pranav, Sheel, Nadim, and Bharat. All the weekly dinner meetings, all preceded in the last year by games of ultimate frisbee were so much enjoyable. My time in San Diego would not have been pleasurable without all the restaurants that we would try. So, I have to thank all the restaurants in San Diego county that have kept my tummy more than full all these years. To get to these restaurants and to commute to campus and most importantly, my favorite grocery store Ralph's at La Jolla Village Square, the city transit agency, the San Diego Metropolitan Transit System (MTS) was the lifeline (in addition to my friends' cars). Of course, it did help that UCSD would pay them to make our bus and trolley journeys free. However, throughout my years of riding MTS buses, I got to know some of the drivers with whom I have made lifelong friendships. Karl, Janith, Chris, Antonio, Steve, Vernie, Wade, Aaron, and others have made me feel so loved and the time that we have shared has been so wonderful.

Finally, the most important people I have to thank have been my parents, Ajitha and Ravikumar. Their love and support have been non-stop. Even when they were unsure about my graduate studies in the US, they were more than ready to take a loan to fund my initial quarters at UCSD and even to this day still ask if I need more money. Without them, my accomplishments would not have been possible and I would not be where I am. My uncle, Krishnakumar and my grandmothers were always a strong source of support to me as well. A call to them would always help me relax. I am lucky to have such a family!

Chapter 2, in part, is a reprint with permission of the material as it appears in the papers: Govind Ravikumar Gopal, Elina Nayebi, Gabriel Porto Villardi, and Bhaskar D. Rao, "Modified vector quantization for small-cell access point placement with inter-cell interference," in *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6387–6401, Aug. 2022, and Govind Ravikumar Gopal, Gabriel Porto Villardi, and Bhaskar D. Rao, "Is vector quantization good enough for access point placement?" in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, November 2021. The dissertation author was the primary

investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-2124929, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

Chapter 3, in part, is a reprint with permission of the material as it appears in the papers: Govind Ravikumar Gopal, Bhaskar D. Rao, and Gabriel Porto Villardi, "Access point placement for hybrid UAV-terrestrial small-cell networks," in *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1826–1841, July 2021, and Govind Ravikumar Gopal, Bhaskar D. Rao, and Gabriel Porto Villardi, "Load balancing in small-cell access point placement," in *2023 IEEE 97th Vehicular Technology Conference (VTC-Spring)*, June 2023. The dissertation author was the primary investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-1617365 and Grant CCF-2124929, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

Chapter 4, in part, is a reprint with permission of the material as it appears in the paper: Govind Ravikumar Gopal and Bhaskar D. Rao, "Vector quantization methods for access point placement in cell-free massive MIMO systems," which is under review in *IEEE Transactions on Wireless Communications*. The dissertation author was the primary investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-2124929 and Grant CCF-2225617, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

Chapter 5, in part, is a reprint with permission of the material as it appears in the paper: Govind Ravikumar Gopal and Bhaskar D. Rao, "Throughput oriented access point deployment in cell-free massive MIMO systems," which is submitted to *IEEE Transactions on Wireless Communications*. The dissertation author was the primary investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-2124929 and Grant CCF-2225617, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

VITA

2016	B.Tech. in Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, India
2018	M.S. in Electrical Engineering (Communication Theory and Systems), University of California San Diego, La Jolla, United States
2017-2023	Graduate Research Assistant, University of California San Diego, La Jolla, United States
2023	Ph.D. in Electrical Engineering (Communication Theory and Systems), University of California San Diego, La Jolla, United States

PUBLICATIONS

G. R. Gopal and B. D. Rao, "Throughput and delay driven access point placement," in *Proc. 2019 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 1010–1014.

G. R. Gopal, B. D. Rao, and G. P. Villardi, "Access point placement for hybrid UAV-terrestrial small-cell networks," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1826–1841, July 2021.

G. R. Gopal, G. P. Villardi, and B. D. Rao, "Is vector quantization good enough for access point placement?" in *Proc. 2021 55th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2021, pp. 1210-1214.

G. R. Gopal, E. Nayebi, G. P. Villardi, and B. D. Rao, "Modified vector quantization for small-cell access point placement with inter-cell interference," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6387–6401, Aug. 2022.

G. R. Gopal, B. D. Rao, and G. P. Villardi, "Load balancing in small-cell access point placement," in *Proc. 2023 IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, June 2023.

G. R. Gopal and B. D. Rao, "Vector quantization methods for access point placement in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, (under review).

G. R. Gopal and B. D. Rao, "Throughput oriented access point deployment in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, (submitted).

ABSTRACT OF THE DISSERTATION

Access Point Placement in Massive Multiple-Input-Multiple-Output Systems

by

Govind Ravikumar Gopal

Doctor of Philosophy in Electrical Engineering
(Communication Theory and Systems)

University of California San Diego, 2023

Professor Bhaskar D. Rao, Chair

The past decade has witnessed the surge of wireless communication technologies, one of which is massive multiple-input-multiple-output (MIMO) which forms a prominent part of 5th generation (5G) and future generations of wireless networks. Among distributed massive MIMO networks, where there are a very large number of antennas physically spread out across a geographical area, one of the degrees of freedom available is access point (AP) placement. In both the traditional small-cell and the more recent cell-free networks, AP placement can be a key enabler of higher throughput or spectral efficiency. Thus, in this dissertation, we study AP placement in both small-cell and cell-free regimes by examining the throughput maximization problem. With the vector quantization (VQ) approach showing similarities to AP placement, we study and utilize it as a starting point for our framework on AP placement formulations.

First, the AP placement problem in the small-cell scenario that addresses signal-to-interference-plus-noise ratio (SINR) maximization (as opposed to signal-to-noise ratio (SNR) maximization alone) is studied since inter-cell interference (ICI) is a prevalent feature of small-cell systems. By first establishing that the Lloyd algorithm from VQ that utilizes the squared Euclidean distance as the distortion measure forms a good enough solution for AP placement, the distortion is modified to include inter-cell interference

(ICI) and a Lloyd-type algorithm is proposed to solve for the AP locations.

Second, the placement solution is extended to hybrid networks consisting of both terrestrial (position-fixed) and unmanned aerial vehicle (UAV) enabled (position-flexible) APs. In this scenario, a Lloyd-type algorithm is used to solve for the positions of the UAV-APs while still accounting for ICI. Additionally, a solution to initialize AP locations for the Lloyd and Lloyd-type algorithms is also discussed. As a further extension, we consider load balancing, which is necessary due to the varied user-AP access resulting from unequal cell occupancies. To address this, two Lloyd-type algorithms are proposed to make the cell occupancies more equal and hence enabling fairness in user-AP access.

Third, the throughput problems for AP placement in cell-free networks are studied, namely the sum rate and minimum rate maximization problems. Their solution structures are analyzed and simple supporting examples are provided. The VQ framework is motivated as a practical approach for these problems. Three VQ-based techniques, namely the Lloyd algorithm, the tree-structured VQ (TSVQ), and probability density function optimized VQ (PDFVQ), each with its own advantages for cell-free AP placement, are outlined and their performances compared.

Finally, in addition to the AP placement and to make the cell-free system more viable, two important and necessary features, namely user position determination and limited cooperation, are added. A multi-step AP deployment process is designed and demonstrated incorporating these features.

Chapter 1

Introduction

1.1 Massive Multiple-Input-Multiple-Output (MIMO) Systems

The concept of massive multiple-input-multiple-output (MIMO) [3] has emerged in recent decades as a strong solution for 5G wireless communication systems [4–6], and for the upcoming 6G networks. By having a large number of antennas, such systems enable higher spectral and energy efficiencies, and reduced interference due to increased diversity [7, 8]. When all the antennas are at the same location, such systems are called colocated antenna systems. On the other hand, when the antennas are all at different locations, such systems are said to be distributed. Fig. 1.1 shows examples of both colocated and distributed cellular systems.

Distributed antenna systems (DASs), especially in the form of distributed MIMO [9–15], give rise to even higher average rates over co-located MIMO systems [10, 16–18]. In general, distributed massive MIMO can either be cooperative or non-cooperative. Small-cell networks, the traditional model where each AP serves one user alone in its cell and do not share information with other APs, are non-cooperative systems. On the contrary, the newer cell-free networks [19, 20], where all APs serve all users and exchange information among themselves, form cooperative systems. Fig. shows the comparison between small-cell and cell-free massive MIMO systems.

(a) Colocated (b) Distributed

Figure 1.1: Colocated versus distributed MIMO systems [1].

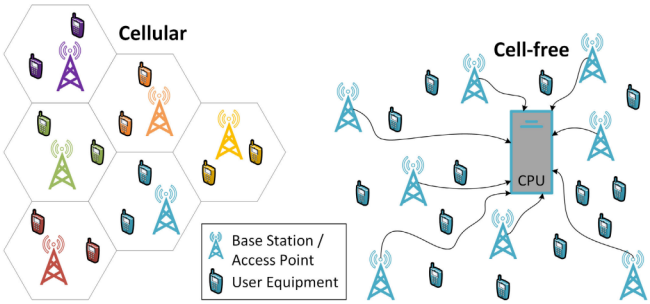

Figure 1.2: Small-cell and cell-free massive MIMO systems [2]. For cooperation, all APs in the cell-free system are connected to a central processing unit (CPU) or network controller (NC).

Although cooperation between the distributed antenna elements or APs in the cell-free approach assists in mitigating interference between users and further increases spectral efficiency over non-cooperative massive MIMO systems, cell-free systems come with a major disadvantage. The required user-related information exchange between APs and the network controller (NC) or the central processing unit, where all or most processing is performed, occupies a significant portion of the usually limited back-haul capacity of wireless systems [9, 21, 22]. These computational and processing requirements are expected to grow with the anticipated network densification in Beyond 5G/6G wireless systems [23]. As a result, despite the benefits of cell-free massive MIMO systems, near future deployments of 5G (3GPP Rel-15 and Rel-16 [24]) and WiFi 6 (IEEE 802.11ax [25]) will still be based on the concept of small-cells, possibly leaving cell-free approaches to posterior deployments of the technology. Nevertheless, a multitude of problems has been explored based on the cell-free network. Prior works investigate topics such as power optimization and energy efficiency [20, 26–28], rate maximization [29, 30], clustering (user- and cell-centric) [31–35], limited fronthaul [36], pilot assignment [37], reconfigurable intelligent surfaces [38], and federated learning [39].

1.2 Access Point (AP) Placement

The focus in this work is a degree of freedom in MIMO system design, namely AP placement. To contextualize the discussion on AP placement, consider a large gathering such as a sporting event, where sections in the stadium see a different number and arrangement of spectators depending on the crowd on the day of the event. To avoid service interruption, more APs should be placed where the number of spectators is larger, and vice versa, leading to the concept of smart stadiums. Additionally, flexible AP deployment is of utmost importance in the infrequent emergency and disaster relief situations, where deployments should be tailored to the time-specific coverage and service requirements, therefore following the dynamics of the emergency event [40]. Thus, AP placement aims to answer the main question: *How do we place APs optimally for a given user distribution?* The optimality can be in terms of a multitude of performance indicators, such as throughput, resource allocation, and fairness.

Previous work in massive MIMO systems, both in the small-cell as well as cell-free regimes, have not focused largely on AP placement. Most works assume some random placement of the APs or a uniform

deployment and unfortunately, the throughput gains that can be obtained from placement have been ignored. Perhaps, one of the initial works on AP or antenna placement has been [13], where the antenna location design problem has been compared to vector quantization (VQ). More recent papers by other research groups have also employed VQ and the Lloyd algorithm as a solution for their optimization problems. However, the discussion on the applicability of VQ and its connection to the placement problem has been limited. In this dissertation, we will thus discuss the VQ framework and we begin by providing a brief review of it.

1.3 A Primer on Vector Quantization (VQ)

VQ [41] is a coding scheme developed as an extension on scalar quantizers for multi-dimensional vectors. In this framework, the data to be quantized, which typically has a probabilistic density (called the source distribution), is quantized to represent it as a set of vectors (called the codebook). It has been extensively used in applications such as image and video compression, clustering, and pattern recognition.

In VQ, the random vector to be quantized is $\mathbf{x} \in \mathbb{R}^p$, where p is the dimension, and the two main steps to be designed are the encoding and decoding steps. The encoder $\mathcal{E}$ splits the source domain under consideration into N regions (called Voronoi regions, each corresponding to a bit sequence of length $\log_2 N$) and assigns a region $\mathcal{R}$ to the input vector $\mathbf{x}$. The encoder performs the following mapping

$$\mathcal{E} : \mathbb{R}^p \rightarrow \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N\}. \tag{1.1}$$

The decoder $\mathcal{D}$ then assigns to each region $\mathcal{R}_n$, where $n = 1, 2, \ldots, N$, a codepoint $\hat{\mathbf{x}}_n$, and performs the mapping

$$\mathcal{D} : \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N\} \rightarrow \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N\}. \tag{1.2}$$

The set of codepoints $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N\}$ is collectively the codebook. Thus, the quantizer $\mathcal{Q}$ assigns for every input $\mathbf{x}$, one of N codepoints, and is given as

$$\mathcal{Q}(\mathbf{x}) = \mathcal{D}(\mathcal{E}(\mathbf{x})) = \hat{\mathbf{x}}_{\mathcal{E}(\mathbf{x})}, \tag{1.3}$$

where $\hat{\mathbf{x}}_{\mathcal{E}(\mathbf{x})}$ specifies that the output codepoint is a function of the input vector and for simplicity in notation, we assume that $\mathcal{E}(\mathbf{x})$ denotes the index of the region that it specifies. The encoder $\mathcal{E}$ assigns to the input $\mathbf{x}$, the region $\mathcal{R}_n$ whose corresponding codepoint $\hat{\mathbf{x}}_n$ is closest to it, defined in terms of a distortion function d between the input vector and a codepoint. The codepoint corresponding to the region can formally be written as

$$\hat{\mathbf{x}}_{\mathcal{E}(\mathbf{x})} = \arg\min_{\hat{\mathbf{x}}_n} d(\mathbf{x}, \hat{\mathbf{x}}_n). \tag{1.4}$$

Taking the average of the distortion function over the distribution of the input vector, the VQ optimization problem is

$$\arg\min_{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N} \mathbb{E}_{\mathbf{x}} \left\{ d(\mathbf{x}, \hat{\mathbf{x}}_{\mathcal{E}(\mathbf{x})}) \right\}. \tag{1.5}$$

To solve the optimization of (1.5), the goal is to find the optimal encoder and decoder jointly, which is difficult. Hence, it is split into two tasks, which are to find a optimal encoder given a fixed decoder and a

optimal decoder given a fixed encoder, and form the two necessary conditions for quantizer optimality. The main methodology then is to alternate between these two tasks in order to converge to a reasonable solution. Accordingly, finding the best encoder given the decoder involves determining the best regions given fixed codepoints. This leads to the *Nearest Neighbor Condition (NNC)*, defined as

$$\mathcal{R}_n = \{\mathbf{x} : d(\mathbf{x}, \hat{\mathbf{x}}_n) \leq d(\mathbf{x}, \hat{\mathbf{x}}_l), \forall l \neq n\}. \tag{1.6}$$

Next, finding the best decoder given the encoder involves determining the best codepoints given the regions. This is the *Centroid Condition (CC)*, given by

$$\hat{\mathbf{x}}_n = \text{Cent}\{\mathbf{x}|\mathbf{x} \in \mathcal{R}_n\}, \tag{1.7}$$

where the centroid Cent[1] of region R_n gives the codepoint $\hat{\mathbf{x}}_n$ for the region. Alternating between the NNC and CC steps until convergence is reached yields the optimal codebook and the algorithm is called the Lloyd algorithm.

1.4 Dissertation Contributions and Organization

It is quite evident that APs should be placed in an area of interest such that the performance (throughput, quality of service, etc.) of all the users in that area are improved, instead of a uniform or random placement (as considered by the majority of popular MIMO works, e.g., [3]). Compared to traditional networks consisting of fixed (in position) base stations (BS), new-age APs could consist, for example, of unmanned aerial vehicles (UAVs) fitted with AP capabilities, and even APs in buildings that can move around on ceiling rails, both of whose flexibilities in positions allow for increased network performance. Solutions to the AP placement problem for both small-cell and cell-free systems in the context of throughput optimality, along with the suitability of the VQ framework and the application of VQ-based methods to solve the same have not been addressed in the literature.

Thus, to start the discussion, AP placement in the small-cell scenario is first considered in Chapter 2. Inter-cell interference (ICI) is a fundamental part of small-cell networks, which for the purposes of AP placement has not been covered in many works. While the standard VQ approach, namely the Lloyd algorithm, has been applied to small-cell AP placement, either a single-cell consideration has been applied or the interference component has been neglected completely. Noting still that the Lloyd algorithm is useful and convenient for signal-to-noise ratio (SNR) improvement, the distortion function used is modified to include the important ICI quantity. The optimization problem in this case now focuses on the signal-to-interference-plus-noise ratio (SINR) as opposed to SNR. Consequently, we come up with two Lloyd-type algorithms, namely the Interference Lloyd algorithm and the Inter-AP Lloyd algorithm that improve the minimum rate of the system compared to when the popular Lloyd algorithm is used.

[1]The centroid is defined [41] as

$$\text{Cent}\{\mathbf{x}|\mathbf{x} \in \mathcal{R}_n\} = \arg\min_{\hat{\mathbf{x}}_n} \mathbb{E}\{d(\mathbf{x}, \hat{\mathbf{x}}_n)|\mathbf{x} \in \mathcal{R}_n\}.$$

Second, we expand the discussion on small-cell AP placement along two fronts in Chapter 3. The first front considers a situation where there both position-fixed (terrestrial) and position-flexible (UAV-enabled) APs exist. We again follow throughput optimality and the inclusion of ICI in the objective function. However, as an alternative to considering SINR above, we include ICI in a metric called signal-to-generated-interference-plus-noise ratio (SGINR). A Lloyd-type algorithm that considers this metric and works on the network consisting of both terrestrial APs and UAV-APs (a hybrid network) is designed for throughput-optimal AP placement. This algorithm is termed Hybrid AP placement algorithm (HAPPA). Further, a common problem encountered in the implementation of Lloyd and Lloyd-type algorithms is the initialization of the AP locations, which ultimately determine the local (or global) optimum to which the AP positions converge. The popular k-means++ method that alleviates this problem to a certain degree is still prone to inconsistent results when the placement algorithms are run multiple times. Hence, utilizing bit allocation from VQ, we design another initialization procedure to improve this inconsistency. On the second front, we address load balancing among the APs for fairness in user access. When the Lloyd algorithm determines the AP positions and associates users to the APs, the resulting unequal occupancies of the cells causes users in some cells to have a more frequent access to their APs than users in other cells. We incorporate tackling this fairness into the Lloyd algorithm through two methods. The first method modifies the distortion function of the Lloyd algorithm by the use of weights that are proportional to the cell occupancies and is called the Occupancy Weighted Lloyd Algorithm (OWLA). The second method involves adding an additional step that prioritizes the re-assignment of users and adopts a distance threshold to cap the resulting throughput loss and is termed the Cell Equalized Lloyd Algorithm-α (CELA-α). Both algorithms show significant gains in enabling fairness in user-AP access while showing minimum throughput losses.

Chapter 4 switches the discussion to AP placement in cell-free networks. Throughput-optimality is considered in the form of two main optimization problems, namely the sum rate and minimum rate optimization problems. With some simple examples, these problems are shown to be challenging to solve. Hence, we again resort to utilizing VQ as a practical approach. Apart from the standard approach, i.e., the Lloyd algorithm, two other VQ-based approaches, namely the tree structured VQ (TSVQ) and probability density function (PDF) optimized VQ (PDFVQ) algorithms, are motivated and described to be used in cell-free AP placement. Each of these methods has their own advantages and offer rate performances. Of the three algorithm, PDFVQ enables a good trade-off between sum and minimum rates in cell-free systems.

The cell-free AP placement paradigm continues in Chapter 5, however, the work described here serves to outline a practical approach to AP deployment. In previous works concerning cell-free AP placement, the user positions or densities are assumed to be known. Further, the concept of limited cooperation has not be integrated into the placement procedure. Thus, to make the cell-free system more viable, we develop a multi-step AP deployment procedure that addresses these two main challenges. Starting from an existing deployment, clustering of the APs for limited cooperation is performed, for which agglomeration clustering is selected as the technique. Then, the user positions can be determined through multilateration by information exchange among within-cluster APs. Utilizing the user locations, a suitable AP placement algorithm can be executed to calculate the new AP locations. Finally, based on the throughput problem of

interest, that is either sum or minimum rate maximization, the fine-tuning of the individual AP locations can be performed by local max-sum gradient and local majorization-minimization, and local max-min gradient methods, respectively.

Finally, it is to be noted that each of the chapters are written to be mostly self-contained. However, it is recommended that Chapters 2 and 3 are read in order, and Chapters 4 and 5 are also read in order since they pertain to small-cell and cell-free AP placement, respectively. Chapter 6 provides concluding remarks for this dissertation.

Chapter 2

Small-Cell AP Placement

In this chapter, we explore the small-cell uplink access point (AP) placement problem in the context of throughput optimality and provide solutions while taking into consideration inter-cell interference (ICI). First, we briefly review the vector quantization (VQ) approach and related single user throughput-optimal formulations for AP placement. Then, we investigate the small-cell case with multiple users and expose the limitations of mean squared error based VQ for solving this problem. While the Lloyd algorithm from the VQ approach is found not to strictly solve the small-cell case, based on the tractability and quality of the resulting AP placement, we deem it suitable as a simple and appropriate framework to solve more complicated problems. Accordingly, to minimize ICI and consequently enhance achievable throughput, we design two Lloyd-type algorithms, namely the Interference Lloyd algorithm and the Inter-AP Lloyd algorithm, both of which incorporate ICI in their distortion functions. Simulation results show that both of the proposed algorithms provide superior 95%-likely rate (the best rate among the worst 5% of the users in the network) over the traditional Lloyd algorithm and the Inter-AP Lloyd algorithm yields a significant increase of up to 36.34% in achievable rate over the Lloyd algorithm.

2.1 Introduction

Due to the large backhaul processing requirements characteristic of cooperative cell-free systems, the non-cooperative small-cell systems will still be useful for the current and future deployments of 5G and Beyond 5G networks. Hence, controlling inter-cell interference (ICI) continues to be a major system design problem, which will actually assume much greater significance with the expected network densification of Beyond 5G systems. Currently, ICI is dealt with in the standards by advanced scheduling techniques such as basic service set (BSS) coloring [25] and dynamic time division duplex (D-TDD) [42]. Also, controlling ICI is of great importance to public protection and disaster relief (PPDR) wireless networks occupying the 700 MHz (and below) frequency bands due to their desirable propagation characteristics and higher signal penetration capabilities, which can cause severe service outages to adjacent emergency networks even in not-so-dense deployments [40, 43–45].

Within the small-cell paradigm, and in recent times, the AP (or antenna) placement problem has attracted a great deal of attention [13–15, 46], however, optimizing the AP or antenna locations by maximizing a signal-to-noise ratio (SNR) objective function alone has traditionally been the standard approach. The authors of [13] consider a distributed antenna system (DAS) and optimize the cell averaged ergodic capacity based only on SNR and neglect ICI. Using the square distance criterion, they notice similarities with codebook design in VQ, which enables the utilization of the well-known (for ease of implementation) Lloyd algorithm to solve the antenna placement problem. In [15], the average achievable per-user rate of uniformly distributed users is optimized in order to find the radius of a circular antenna array; however, due to the adoption of a single-cell model, no ICI is considered. Circular antenna array deployments based on average rate optimization are also considered in [14] based on one-cell and two-cell models, with the latter model accounting for leakage interference alone. Additionally, the authors of [46] simulated an indoor wireless environment where they generated a 10-fold improvement in the distributed system capacity over the co-located one. Further, placing APs in accordance with the user densities generated a significant increase (40% over uniform AP placement) in system capacity. The authors of [47] and the subsequent works by their group [48–51] have discussed heterogeneous wireless sensor network deployment as a source coding problem. In these works, the optimal deployment is solved while studying limited communication range and optimal total power consumption to place both APs and fusion centers, but without addressing ICI. Recently, unmanned aerial vehicles (UAVs) equipped with base stations have also been considered in the context of AP placement [52–59].

In the abovementioned works, the suitability of the Lloyd algorithm from VQ to throughput-optimal AP placement has not been investigated. VQ considers only a single user, and the objective function is averaged over the position of this user. This approach, however, does not conform with the small-cell scenario where there are multiple users, one from each cell, communicating with the serving AP in its cell. Further, ICI has been neglected, therefore leading to AP placements that yield sub-optimal throughput. Hence, in this chapter, we devise a non-cooperative small-cell system based on the Lloyd algorithm, which we show can solve for near-optimal AP locations, in terms of the fundamental performance measure of throughput, while considering ICI.

Contributions

To the best of our knowledge, solutions to the AP placement problem based on the Lloyd algorithm and that are derived from a detailed analysis of throughput optimality, while incorporating ICI, have not been provided in the literature. Hence, in this chapter on small-cell AP placement, our contributions are as follows.

- We first formulate various single user AP placement problems for throughput optimality in terms of rate, SNR, and a higher exponent for the user-AP distance (as opposed to squared distance). We explore the relationship of the Lloyd algorithm from VQ to these problems. We then study the multiple user case and address the small-cell AP placement problem. Although our analysis determines that the application of the Lloyd algorithm to small-cell AP placement is not ideal, we find that the Lloyd

algorithm, apart from being easy to implement, is quite effective in solving the placement problem as a baseline algorithm and yields near-optimal AP locations.

- We present two methods to incorporate ICI into the optimization function of the Lloyd algorithm. Consequently, the distortion function of the Lloyd algorithm is modified and two Lloyd-type algorithms for AP placement that are aware of ICI and, as a result, maximize achievable per-user signal-to-interference-plus-noise ratio (SINR), are proposed, namely the Interference Lloyd algorithm and Inter-AP Lloyd algorithm.

2.2 System Model

We use the small-cell model detailed in [60] and [61, Ch. 4], which is reproduced here for completeness. Now, consider a geographical area where K single-antenna users are distributed, according to some probability density function (pdf) $f_{\mathbf{P}}(\mathbf{p})$, where $\mathbf{p} \in \mathbb{R}^2$ is the random vector denoting the position of a user. There are M single-antenna APs that serve the users in this area. The location of an AP is denoted by $\mathbf{q} \in \mathbb{R}^2$. All APs are connected via error-free backhaul links to the network controller[1] (NC), so that it knows the positions of the APs and their respective users. For simplicity, a narrowband flat-fading channel is considered. With $m = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, K$, the channel coefficient between the m^{th} AP and k^{th} user is

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}, \tag{2.1}$$

where β_{mk} and $h_{mk} \sim \mathcal{CN}(0, 1)$ are the large-scale and small-scale fading coefficients, respectively. h_{mk} is assumed to remain constant during a coherent interval and change independently in the next, and is independent of β_{mk}. The large-scale fading coefficients are modeled as

$$\beta_{mk} = \begin{cases} c_0, & ||\mathbf{p}_k - \mathbf{q}_m|| \leq r_0, \\ \frac{c_1 z_{mk}}{||\mathbf{p}_k - \mathbf{q}_m||^{\gamma}}, & ||\mathbf{p}_k - \mathbf{q}_m|| > r_0, \end{cases} \tag{2.2}$$

where $\mathbf{p}_k$ and $\mathbf{q}_m$ represent the locations of the k^{th} user and m^{th} AP, respectively. Here, γ is the pathloss exponent, z_{mk} is the log-normal shadow fading coefficient, and c_0, c_1, and r_0 are constants. These coefficients can also be estimated by either ray-tracing [62] or data-driven [63] approaches.

The uplink transmission model used in this chapter schedules users in a round robin fashion with their serving APs using time-division multiple access (TDMA). Thus, each AP serves only one user in a time slot. In the small-cell setup, each of the M cells corresponds to each of the M APs, and pursuant with the uplink model, the user in each cell communicating with its associated AP causes interference to all other APs. Now, letting k_m denote a user in the cell associated with AP m, the received signal y_m at this AP is

$$y_m = \sum_{m'=1}^{M} \sqrt{\rho_r} g_{mk_{m'}} s_{k_{m'}} + w_m, \tag{2.3}$$

[1]The NC is where the proposed placement algorithms to be described in detail in the remainder of this chapter will be loaded and executed.

where ρ_r is the uplink transmit power, s_{k_m} is the data symbol with $\mathbb{E}\{|s_{k_m}|^2\} = 1$ (unit power), and $w_m \sim \mathcal{CN}(0,1)$ is the additive noise. A matched filter (MF) employed at the AP m estimates the data symbol s_{k_m} of user k_m as

$$\begin{aligned}\hat{s}_{k_m} &= \frac{g^*_{mk_m}}{|g_{mk_m}|} y_m, \\ &= \underbrace{\sqrt{\rho_r}|g_{mk_m}|s_{k_m}}_{T_{\text{des}}\text{: desired term}} + \underbrace{\sum_{\substack{m'=1 \\ m'\neq m}}^{M} \sqrt{\rho_r}\frac{g^*_{mk_m}}{|g_{mk_m}|} g_{mk_{m'}} s_{k_{m'}}}_{T_{\text{int}}\text{: interference term}} + v_m,\end{aligned} \tag{2.4}$$

where $v_m \sim \mathcal{CN}(0,1)$. Considering T_{int} in (3.4) as noise, the signal-to-interference-plus-noise ratio (SINR) achieved by user k_m at AP m is derived to be

$$\phi_{k_m} = \frac{\rho_r \beta_{mk_m}|h_{mk_m}|^2}{1 + \rho_r \sum\limits_{\substack{m'=1 \\ m'\neq m}}^{M} \beta_{mk_{m'}}|h_{mk_{m'}}|^2}. \tag{2.5}$$

2.3 VQ and the Lloyd Algorithm for AP Placement

Since VQ has been introduced in Chapter 1, we show in this section how the Lloyd algorithm us currently used in its basic form, to solve for AP placement. Note that Section 2.4 will investigate the suitability of the Lloyd algorithm to obtain AP locations.

If the VQ approach were to be used to solve for small-cell AP placement, then the random vector to be quantized is the 2-D position $\mathbf{p}$ of a *single* user. The Voronoi regions are the cells $\mathcal{C}_m$ and the codepoints are the AP locations $\mathbf{q}_m$, where $m = 1, 2, \ldots, M$. The optimization problem in (1.5) can be written by using similar notations and taking the average over the user positions, as follows

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\min}\ \mathbb{E}_{\mathbf{p}}\left\{d(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})})\right\}. \tag{2.6}$$

It is worth reiterating that $\mathcal{E}(\mathbf{p})$ indexes the nearest AP that the user at $\mathbf{p}$ associates to. The objective function in (2.6) can be written as

$$\begin{aligned}J_{\text{VQ}} &= \mathbb{E}_{\mathbf{p}}\left\{d(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})})\right\}, \\ &= \int\limits_{\mathbf{p}\in\mathbb{R}^2} d(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})}) f_{\mathbf{P}}(\mathbf{p})\mathrm{d}\mathbf{p}, \\ &= \sum_{m=1}^{M} \left[\int\limits_{\mathbf{p}\in\mathcal{C}_m} d(\mathbf{p}, \mathbf{q}_m) f_{\mathbf{P}}(\mathbf{p}|\mathbf{p}\in\mathcal{C}_m)\mathrm{d}\mathbf{p}\right] \Pr(\mathbf{p}\in\mathcal{C}_m), \\ &= \sum_{m=1}^{M} S_m \Pr(\mathbf{p}\in\mathcal{C}_m),\end{aligned} \tag{2.7}$$

where the penultimate step arises by splitting the integral in the previous step into the cells (Voronoi regions)

with their respective codepoints and the quantity S_m is defined as

$$S_m = \int\limits_{\mathbf{p}\in\mathcal{C}_m} d(\mathbf{p},\mathbf{q}_m) f_{\mathbf{P}}(\mathbf{p}|\mathbf{p}\in\mathcal{C}_m)\mathrm{d}\mathbf{p}. \tag{2.8}$$

To solve for the optimal AP locations, the most often used distortion function is the squared Euclidean distance

$$d_{\mathrm{SE}}(\mathbf{p},\mathbf{q}_{\mathcal{E}(\mathbf{p})}) = \left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2, \tag{2.9}$$

and the objective function in (2.7) then becomes the mean squared error (MSE). In this chapter, we retain the name 'Lloyd algorithm' for the algorithm that solves (2.7) using d_{SE}, the steps of which are provided in Algorithm 12. For algorithms that use all other distortion functions, we will use the name 'Lloyd-type algorithm'. Note that when the Lloyd algorithm is implemented, we use the K realization of users at positions

Algorithm 1 Lloyd Algorithm With Squared Error Distortion

1: Initialize random AP locations $\mathbf{q}_1^{(0)},\mathbf{q}_2^{(0)},\ldots,\mathbf{q}_M^{(0)}$.

2: Use the NNC to determine the cells $\mathcal{C}_1^{(i+1)},\mathcal{C}_2^{(i+1)},\ldots,\mathcal{C}_M^{(i+1)}$ such that

$$\mathcal{C}_m^{(i+1)} = \left\{\mathbf{p}_k : d_{\mathrm{SE}}\left(\mathbf{p}_k,\mathbf{q}_m^{(i)}\right) \le d_{\mathrm{SE}}\left(\mathbf{p}_k,\mathbf{q}_l^{(i)}\right), \forall l\neq m\right\}.$$

3: Use the CC to determine the AP locations $\mathbf{q}_1^{(i+1)},\mathbf{q}_2^{(i+1)},\ldots,\mathbf{q}_M^{(i+1)}$ such that

$$\mathbf{q}_m^{(i+1)} = \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|}\sum_{\mathbf{p}_k\in\mathcal{C}_m^{(i+1)}}\mathbf{p}_k.$$

4: Repeat from step 2 until convergence (MSE falls below a threshold).

$\mathbf{p}_k$, $k=1,2,\ldots,K$, as described in Section 2.2. We will use this notation for all the Lloyd-type algorithms that follow. Also, observe that in the CC step in 12, the centroid is replaced by the expectation which is evaluated by using the sample average over the user positions $\mathbf{p}_k$ present in cell $\mathcal{C}_m$. The calculation of the centroid of the region is applicable only due to the squared error distortion d_{SE} used [41]. Additionally, as opposed to the ℓ_2-norm in (4.17), the ℓ_1-norm can be considered, and the corresponding algorithm is called the *k-medians algorithm*. Note that the ℓ_1 norm-based distortion function is given as

$$d\left(\mathbf{p},\mathbf{q}_m\right) = ||\mathbf{p}-\mathbf{q}_m||_1^2. \tag{2.10}$$

While the NNC step utilizes the above distortion function, the CC step now involves the calculation of the median of the user positions in the cell, defined as

$$\mathbf{q}_m = \mathrm{Med}\left[\mathbf{p}|\mathbf{p}\in\mathcal{C}_m\right], \tag{2.11}$$

where Med[·] denotes the median operation. This is called the geometric median, which is the point that minimizes the sum of Euclidean distances to the user positions. Formally, this point is chosen according to the following optimization problem

$$\mathbf{q}_m = \arg\min_{\mathbf{y}} \sum_{\mathbf{p}\in\mathcal{C}_m} ||\mathbf{p}-\mathbf{y}||_2. \tag{2.12}$$

The solution of this optimization is performed through a form of iterative least-squares solution called Weiszfeld's algorithm [64].

An interesting observation here is that the VQ framework presented above considers only the positions of the users and APs, and hence is independent of both the small-scale fading and shadow fading components of the wireless system since these quantities are not dependent on the user and AP positions. These random quantities thus do not play a role in AP placement using VQ. It is also very important to note here that VQ considers only a single user to be quantized and the average over the distribution of that user is taken. However, this does not conform to our small-cell system model, where M users are each communicating with its serving AP at the same time. Hence, the VQ approach does not strictly solve the small-cell AP placement problem.

2.4 Throughput Formulations and Solutions Without Inter-Cell Interference

In this section, we will describe throughput optimization via various formulations, such as average rate and SNR, and provide solutions to obtain optimal AP locations. We start by considering the single user scenario inherent to VQ and expand to a more realistic one in which multiple users are present. We also illustrate, by formulation only, the case where ICI is present. In summary, we argue how the Lloyd algorithm, despite its simplicity, is suitable for small-cell AP placement.

2.4.1 Single User Case

Rate

The single user case is the simplest case wherein a user at location $\mathbf{p}$ alone is considered. Recall that $\mathbf{p}$ is a random vector with pdf $f_{\mathbf{P}}(\mathbf{p})$. We start our analysis with per-user rate, which is the common measure of interest, achieved by a user at $\mathbf{p}$ with its nearest AP at $\mathbf{q}_{\mathcal{E}(\mathbf{p})}$, as per the VQ principles discussed above. We also approximate the large-scale fading coefficients, given in (3.2), by

$$\beta_{\mathcal{E}(\mathbf{p})} \approx \frac{c_1 z_{\mathcal{E}(\mathbf{p})}}{\left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}, \tag{2.13}$$

since r_0 is much smaller than the dimensions of the area under consideration. Note that the second subscript has been dropped for the ensuing analyses, since we consider a single user. Let us define the average rate, utilizing the per-user SNR $\psi_{k_{\mathcal{E}(\mathbf{p})}}$ (obtained from (3.5) by neglecting ICI and replacing m with $\mathcal{E}(\mathbf{p})$) as follows

$$\overline{r}(\underline{\mathbf{q}}) = \mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\log\left(1 + \psi_{k_{\mathcal{E}(\mathbf{p})}}\right)\right\}, \tag{2.14}$$

where we average over the user position $\mathbf{p}$, the random quantities $h_{\mathcal{E}(\mathbf{p})}$ and $z_{\mathcal{E}(\mathbf{p})}$, $\mathcal{A} = \{h_{\mathcal{E}(\mathbf{p})}, z_{\mathcal{E}(\mathbf{p})}\}$ for brevity, and we use the notation $\underline{\mathbf{q}} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M\}$ to show that the average rate is a function of the M AP locations alone. Similar to VQ in the previous section, we average out the small-scale and shadow

fading components defined in $\mathcal{A}$ since they are position independent and do not contribute to the optimal placement of APs. Assuming high SNR ($\psi_{k_{\mathcal{E}(\mathbf{p})}} \gg 1$), we can write (2.14) as

$$\overline{r}(\underline{\mathbf{q}}) = \mathbb{E}_{\mathcal{A},\mathbf{p}} \left\{ \log \left(\frac{\rho_r c_1 |h_{\mathcal{E}(\mathbf{p})}|^2 z_{\mathcal{E}(\mathbf{p})}}{\left(||\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}|| \right)^{\frac{\gamma}{2}}} \right) \right\}, \tag{2.15}$$

and we wish to perform the optimization

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\max} \ \overline{r}(\underline{\mathbf{q}}). \tag{2.16}$$

After averaging and removing the terms that are not involved in the optimization in (2.15), we obtain

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\min} \ \mathbb{E}_{\mathbf{p}} \left\{ \log \left(||\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}||^2 + \epsilon \right) \right\}, \tag{2.17}$$

where we have added a constant $\epsilon > 0$ (typically very small) to prevent the logarithm from approaching negative infinity if the user position $\mathbf{p}$ were to overlap with the position of the nearest AP $\mathbf{q}_{\mathcal{E}(\mathbf{p})}$. Note that ϵ could correspond to the pathloss at a reference distance or even height of the AP. The objective function to be optimized above is concave as a result of which the Majorization-Minimization (MM) technique [65] can be used to acquire a solution to the centroid computation (CC) step. Although the distortion function in (2.17) is the logarithm of the squared Euclidean distance, the NNC step here remains the same as in the Lloyd algorithm since both $\log(\cdot)$ and ϵ can be ignored when comparing two distortion functions. The MM technique upperbounds the objective function by a surrogate function and minimizes the surrogate through an iterative method. Solving the objective function in (2.17) using the MM method results in an iterative solution with the following two update equations

$$\begin{aligned} \mathbf{q}_m^{(j+1)} &= \frac{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m} w_k^{(j)} \mathbf{p}_k}{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m} w_k^{(j)}}, \\ w_k^{(j+1)} &= \frac{1}{||\mathbf{q}_m^{(j+1)} - \mathbf{p}_k||^2 + \epsilon}, \ \forall \mathbf{p}_k \in \mathcal{C}_m, \end{aligned} \tag{2.18}$$

where j denotes the MM iteration index.

In summary, to solve for the AP locations, we can now formulate a Lloyd-type algorithm with the NNC step remaining the same as that in the Lloyd algorithm, i.e., with d_{SE}, and the CC step replaced by the above iterative solution of (5.20). We call this Lloyd-type algorithm as the *MM-Lloyd algorithm*. The proof of (5.20) is left to Appendix 5.A and the algorithm is provided in Algorithm 2.

SNR

If throughput is measured solely by the SNR averaged over the user location $\mathbf{p}$, then we can show that the simple case of SNR maximization is equivalent to the VQ optimization problem given in (2.7). Let us write the average achievable SNR as

$$\overline{\psi}(\underline{\mathbf{q}}) = \mathbb{E}_{\mathcal{A},\mathbf{p}} \left\{ \frac{\rho_r c_1 |h_{\mathcal{E}(\mathbf{p})}|^2 z_{\mathcal{E}(\mathbf{p})}}{||\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}||^{\gamma}} \right\}, \tag{2.19}$$

Algorithm 2 MM-Lloyd Algorithm

1: Initialize random AP locations $\mathbf{q}_1^{(0)}, \mathbf{q}_2^{(0)}, \ldots, \mathbf{q}_M^{(0)}$.

2: Use the NNC to determine the cells $\mathcal{C}_1^{(i+1)}, \mathcal{C}_2^{(i+1)}, \ldots, \mathcal{C}_M^{(i+1)}$ such that

$$\mathcal{C}_m^{(i+1)} = \left\{ \mathbf{p}_k : d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m \right\}.$$

3: Use MM iterations to determine the AP locations $\mathbf{q}_1^{(i+1)}, \mathbf{q}_2^{(i+1)}, \ldots, \mathbf{q}_M^{(i+1)}$ with the update equations

$$\mathbf{q}_m^{(j+1)} = \frac{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} w_k^{(j)} \mathbf{p}_k}{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} w_k^{(j)}},$$

$$w_k^{(j+1)} = \frac{1}{||\mathbf{q}_m^{(j+1)} - \mathbf{p}_k||^2 + \epsilon}, \ \forall \mathbf{p}_k \in \mathcal{C}_m^{(i+1)},$$

where $\mathbf{q}_m^{(i+1)} = \mathbf{q}_m^{(j+1)}$ after convergence.

4: Repeat from step 2 until convergence.

which is lower bounded by applying Jensen's inequality as

$$\overline{\psi}(\underline{\mathbf{q}}) \geq \mathbb{E}_{\mathcal{A}} \left\{ \frac{\rho_r c_1 |h_{\mathcal{E}(\mathbf{p})}|^2 z_{\mathcal{E}(\mathbf{p})}}{\left(\mathbb{E}_{\mathbf{p}} \left\{ \left|\left| \mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})} \right|\right|^2 \right\} \right)^{\frac{\gamma}{2}}} \right\}, \tag{2.20}$$

with $\mathcal{A}$ defined as before. Maximizing $\overline{\psi}(\underline{\mathbf{q}})$ to obtain the AP locations is the same as minimizing the term in the denominator, leading to the same objective function (2.7) in VQ. The optimization problem is

$$\underset{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M}{\arg\min} \ \mathbb{E}_{\mathbf{p}} \left\{ \left|\left| \mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})} \right|\right|^2 \right\}. \tag{2.21}$$

As before, this is solved using the Lloyd algorithm with d_{SE} (Algorithm 12). For consistency in future discussions, we introduce the notation $d(\mathbf{p}, \underline{\mathbf{q}})$ as a general form of distortion measure with $\underline{\mathbf{q}} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M\}$. Hence, the squared error distortion function in (4.17) is written in the general form as

$$d_{\mathrm{SE}}\left(\mathbf{p}, \underline{\mathbf{q}}\right) = \left|\left| \mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})} \right|\right|^2. \tag{2.22}$$

Higher Exponent for User-AP Distance

The objective function in the Lloyd algorithm is proportional to the square of the user-AP distance while that in the MM-Lloyd algorithm is proportional to the logarithm of the squared distance. This means that the MM-Lloyd algorithm disproportionately considers the contribution of users, as the logarithm suppresses the larger distances inherent to users at the cell borders, in comparison to the Lloyd algorithm. To overcome this effect, we can design another optimization function that exponentially scales up large distances relative to the Lloyd algorithm by raising the distance to a higher power. The topic of optimal quantizer design for higher powers of distance has been studied in [48]. This higher exponent $\chi > 2$ also characterizes

higher frequency (e.g., mmWave) communications. The optimization problem can then be represented as

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\min}\ \mathbb{E}_{\mathbf{p}}\left\{\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\chi}\right\}, \tag{2.23}$$

where χ is the power. This optimization problem can be solved by using a Lloyd-type algorithm that uses the distortion function

$$d_{\chi}\left(\mathbf{p},\underline{\mathbf{q}}\right)=\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\chi}. \tag{2.24}$$

While the NNC step uses d_{χ}, the CC step utilizes the steepest descent method, with the update equation

$$\mathbf{q}_m^{(j+1)}=\mathbf{q}_m^{(j)}-\delta\frac{\partial}{\partial\mathbf{q}_m^{(j)}}\left\{\int\limits_{\mathbf{p}\in\mathcal{C}_m} d_{\chi}\left(\mathbf{p},\mathbf{q}_m^{(j)}\right)f_{\mathbf{P}}(\mathbf{p})\mathrm{d}\mathbf{p}\right\}, \tag{2.25}$$

for all m, where j is the iteration index, δ is the step size, and the gradient expression is given by

$$\frac{\partial}{\partial\mathbf{q}_m}\left\{\int\limits_{\mathbf{p}\in\mathcal{C}_m} d_{\chi}\left(\mathbf{p},\mathbf{q}_m\right)f_{\mathbf{P}}(\mathbf{p})\mathrm{d}\mathbf{p}\right\}=\frac{\chi}{|\mathcal{C}_m|}\sum_{\mathbf{p}_k\in\mathcal{C}_m}\left(\mathbf{q}_m-\mathbf{p}_k\right)\left|\left|\mathbf{p}_k-\mathbf{q}_m\right|\right|^{\chi-2}. \tag{2.26}$$

This Lloyd-type algorithm is called the *Lloyd-χ algorithm* and the proof of the above result for gradient can be found in Appendix 2.B. The algorithm is given in Algorithm 3.

Algorithm 3 Lloyd-χ Algorithm

1: Initialize random AP locations $\mathbf{q}_1^{(0)},\mathbf{q}_2^{(0)},\ldots,\mathbf{q}_M^{(0)}$.

2: Use the NNC to determine the cells $\mathcal{C}_1^{(i+1)},\mathcal{C}_2^{(i+1)},\ldots,\mathcal{C}_M^{(i+1)}$ such that

$$\mathcal{C}_m^{(i+1)}=\left\{\mathbf{p}_k:d_{\chi}\left(\mathbf{p}_k,\mathbf{q}_m^{(i)}\right)\leq d_{\chi}\left(\mathbf{p}_k,\mathbf{q}_l^{(i)}\right),\forall l\neq m\right\}.$$

3: Use the steepest descent method to determine the AP locations $\mathbf{q}_1^{(i+1)},\mathbf{q}_2^{(i+1)},\ldots,\mathbf{q}_M^{(i+1)}$ with the update equation

$$\mathbf{q}_m^{(j+1)}=\mathbf{q}_m^{(j)}-\frac{\delta\chi}{\left|\mathcal{C}_m^{(i+1)}\right|}\sum_{\mathbf{p}_k\in\mathcal{C}_m^{(i+1)}}\left(\mathbf{q}_m^{(j)}-\mathbf{p}_k\right)\left|\left|\mathbf{p}_k-\mathbf{q}_m^{(j)}\right|\right|^{\chi-2},$$

where $\mathbf{q}_m^{(i+1)}=\mathbf{q}_m^{(j+1)}$ after convergence.

4: Repeat from step 2 until convergence.

The above formulations were developed by assuming a single user located at $\mathbf{p}$. However, in practice and according to the system model, M APs serve M users at the same time. Hence, we now consider the case where M users are picked from the distribution.

2.4.2 Multiple User Case

Random User Selection

If M users are selected independently from the overall distribution $f_{\mathbf{P}}(\mathbf{p})$, then the distribution of these users are i.i.d. Let $\underline{\mathbf{p}}\triangleq\{\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_M\}$ be the set of locations of the M users. If we assume that the

users do not interact with each other[2], then the objective function can be the sum of distortions incurred by each user with its closest AP, i.e., the optimization is of the form

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\min}\ \mathbb{E}_{\underline{\mathbf{p}}}\left\{\sum_{m=1}^{M} d(\mathbf{p}_m, \mathbf{q}_{\mathcal{E}(\mathbf{p}_m)})\right\} = \underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\min}\ M \cdot \mathbb{E}_{\mathbf{p}}\left\{d(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})})\right\}, \tag{2.27}$$

where $\mathbf{q}_{\mathcal{E}(\mathbf{p})}$ defined as before and the simplification arises from the fact that each user is i.i.d. The final objective function thus is essentially the same as the single user case.

The above model is applicable in the following scenario. First, since the selection does not limit one user per small cell, the cells must be capable of dealing with more than one user with no multiple access interference. Secondly, since there is no ICI considered, each small cell must be assigned orthogonal resources. This leads to an interesting resource allocation problem which we do not pursue here.

Random Selection of One User Per Cell without ICI

The formulation described above considers M users at a time, but fails to follow the system model as each user is not necessarily picked from the Voronoi region or cell in which its serving AP is present. Under this model, assuming again that the users at $\underline{\mathbf{p}} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M\}$ do not interact with one another, the objective function to minimize would be the sum of the average distortion in each cell, i.e., the optimization is

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\min}\ \mathbb{E}_{\underline{\mathbf{p}}}\left\{\sum_{m=1}^{M} d(\mathbf{p}_m, \mathbf{q}_m)\right\}, \tag{2.28}$$

with the joint distribution of the user positions as

$$f_{\underline{\mathbf{P}}}(\underline{\mathbf{p}}) = \prod_{m=1}^{M} f_{\mathbf{P}_m}(\mathbf{p}_m | \mathbf{p}_m \in \mathcal{C}_m). \tag{2.29}$$

The above objective function can be simplified as

$$\begin{aligned}\sum_{m=1}^{M} \mathbb{E}_{\underline{\mathbf{p}}}\left\{d(\mathbf{p}_m, \mathbf{q}_m)\right\} &= \sum_{m=1}^{M} \int_{\mathbf{p}\in\mathcal{C}_m} d(\mathbf{p}, \mathbf{q}_m) f_{\mathbf{P}}(\mathbf{p}|\mathbf{p}\in\mathcal{C}_m)\mathrm{d}\mathbf{p}, \\ &= \sum_{m=1}^{M} S_m,\end{aligned} \tag{2.30}$$

where S_m is from (2.8). It is worth noting here the difference between this objective function and that of the Lloyd algorithm in (2.7) where each term S_m is weighted by the probability that the user is present in the cell $\Pr(\mathbf{p} \in \mathcal{C}_m)$. The solution to the above objective function is then a Lloyd-type algorithm with the CC step unchanged, but with the NNC step using weighted distortion functions, with the weights being the inverse of the proportion of users present in the cell. More specifically, the squared error distortion $d_{\mathrm{SE}}(\mathbf{p}, \mathbf{q}_m)$ is pre-multiplied with a weight $w_m = 1/\Pr(\mathbf{p} \in \mathcal{C}_m) = K/N_m$, where N_m is the number of users in $\mathcal{C}_m$. The NNC step is

$$\mathcal{C}_m = \left\{\mathbf{p} : w_m d_{\mathrm{SE}}(\mathbf{p}, \mathbf{q}_m) \le w_l d_{\mathrm{SE}}(\mathbf{p}, \mathbf{q}_l), \forall l \ne m\right\}. \tag{2.31}$$

[2]This implies that a user does not influence the AP selection of any other user. In other words, the distortion function between a user at $\mathbf{p}_m$ and its closest AP $\mathbf{q}_{\mathcal{E}(\mathbf{p}_m)}$ is independent of the positions of all other users $\mathbf{p}_{m'}$, where $m' \ne m$. It is worth noting that multiple users can select the same AP as its closest one.

We call this algorithm as the *weighted MSE (WMSE) Lloyd algorithm*. Note that weighted distortion functions have been studied in [47], although the authors have considered weights that remain constant. The weights in the WMSE Lloyd algorithm, on the other hand, are learnt in every iteration. The proof of the above solution is provided in Appendix 2.C and the algorithm is outlined in Algorithm 4.

Algorithm 4 WMSE Lloyd Algorithm

1: Initialize random AP locations $\mathbf{q}_1^{(0)}, \mathbf{q}_2^{(0)}, \ldots, \mathbf{q}_M^{(0)}$.

2: Use the NNC to determine the cells $\mathcal{C}_1^{(i+1)}, \mathcal{C}_2^{(i+1)}, \ldots, \mathcal{C}_M^{(i+1)}$ such that

$$\mathcal{C}_m^{(i+1)} = \left\{ \mathbf{p}_k : w_m d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq w_j d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m \right\}.$$

3: Use the CC to determine the AP locations $\mathbf{q}_1^{(i+1)}, \mathbf{q}_2^{(i+1)}, \ldots, \mathbf{q}_M^{(i+1)}$ such that

$$\mathbf{q}_m^{(i+1)} = \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \mathbf{p}_k.$$

4: Repeat from step 2 until convergence.

Random Selection of One User Per Cell with ICI

In all the above formulations, we have considered only SNR and the fact that users do not interact with one another. However, under the effects of ICI, users do interact with one another in the form of providing interfering signals at the APs which are serving the other users. Thus, the distortion function between a user and its serving AP would be a function of all other users as well and under a similar fashion as in (2.28), we can write the objective function as

$$\sum_{m=1}^{M} \mathbb{E}_{\underline{\mathbf{p}}} \left\{ d(\mathbf{p}_m, \mathbf{q}_m, \underline{\mathbf{p}}'_m) \right\} = \sum_{m=1}^{M} \int_{\mathbf{p}_1 \in \mathcal{C}_1} \cdots \int_{\mathbf{p}_M \in \mathcal{C}_M} d(\mathbf{p}_m, \mathbf{q}_m, \underline{\mathbf{p}}'_m) f_{\underline{\mathbf{P}}}(\underline{\mathbf{p}}) \mathrm{d}\underline{\mathbf{p}}, \tag{2.32}$$

where the (general) distortion function uses the term $\underline{\mathbf{p}}'_m$ which denotes the set of user positions other than the user at $\mathbf{p}_m$ and $f_{\underline{\mathbf{P}}}(\underline{\mathbf{p}})$ is as in (2.29). It is clear from the objective function in (2.32), due to the dependency of the distortion function on the interfering users, the joint distribution $f_{\underline{\mathbf{P}}}(\underline{\mathbf{p}})$ cannot be simplified to consider each cell $\mathcal{C}_m$ independently as in (2.30). This makes the said objective function difficult and intractable, and hence cannot be readily solved.

To deal with ICI in a tractable manner, we adopt a slightly different approach based on the following considerations. Based on results obtained so far, VQ provides a good framework to solve throughput optimization problems by Lloyd-type algorithms, although without ICI. We have also seen that the optimization in (2.32), which considers ICI, is difficult to solve and to derive an AP placement algorithm. Further, numerical simulations (shown in Experiment 1 of Section 2.7.3) show that the average achievable rate is very similar, whether the Lloyd or Lloyd-type algorithms described in this section are used. Motivated by these three facts, in the next section, we show how the Lloyd algorithm can be modified to account for ICI in AP placement. It is worth noting here that we could implement power control along with AP placement, i.e.,

optimizing uplink power with per-user power constraints jointly with the AP locations, in order to increment rate. Our focus in this work, however, is to solely investigate the appropriateness of the VQ approach to small-cell AP placement and the necessary modifications to add ICI to the VQ optimization framework such that a Lloyd-type algorithm can be used to solve the problem, both of which have not been performed in past literature. Thus, we will continue to assume the same uplink power for all users and address power control in future work.

2.5 Throughput Formulations Accounting for Inter-Cell Interference

To account for ICI in the VQ framework, we develop two distortion functions, namely the interference and inter-AP distortion functions.

2.5.1 Interference Distortion Measure

From (2.21), it is clear that the Lloyd algorithm maximizes only the desired signal component. In addition, we are now required to minimize the interference term. To construct a distortion function that considers both the desired and interference signals, we consider the achievable per-user rate, as considered in Section 2.4.1, but using the SINR expression from (3.5). Formally, the rate maximization problem is

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\max}\ \mathbb{E}_{\mathcal{A},\mathcal{B},\underline{\mathbf{p}}}\left\{\log\left(1+\phi_{k_{\mathcal{E}(\mathbf{p})}}\right)\right\}, \tag{2.33}$$

where set $\mathcal{A}=\{h_{\mathcal{E}(\mathbf{p})}, z_{\mathcal{E}(\mathbf{p})}\}$ defined as before and set $\mathcal{B}=\{h_{m'}, z_{m'} : m'\neq\mathcal{E}(\mathbf{p})\}$ consists of the small-scale and shadow fading quantities for all interfering cells $\mathcal{C}_{m'}$, $m'\neq\mathcal{E}(\mathbf{p})$. For notational simplicity, the SINR $\phi_{k_{\mathcal{E}(\mathbf{p})}}$ above can be rewritten using T_{SNR} for the desired signal power in the numerator and T_{ICI} for the interference signal power in the denominator as follows

$$\phi_{k_{\mathcal{E}(\mathbf{p})}}=\frac{T_{\text{SNR}}}{1+T_{\text{ICI}}}, \tag{2.34}$$

where $T_{\text{SNR}}=\rho_r\beta_{\mathcal{E}(\mathbf{p})}|h_{\mathcal{E}(\mathbf{p})}|^2$ and $T_{\text{ICI}}=\rho_r\sum_{m'\neq m}\beta_{m'}|h_{m'}|^2$. To recapitulate the notation, we use a single subscript for simplicity and while $h_{\mathcal{E}(\mathbf{p})}$ and $\beta_{\mathcal{E}(\mathbf{p})}$ are the small-scale and large-scale fading coefficients, respectively, for the user at $\mathbf{p}$ to the serving cell, $h_{m'}$ and $\beta_{m'}$ correspond to the same quantities for the same user, but to the non-serving AP m'. Approximating the rate with high SINR ($\phi_{k_{\mathcal{E}(\mathbf{p})}}\gg 1$) and $T_{\text{ICI}}\gg 1$, and simplifying, we get

$$\log\phi_{k_{\mathcal{E}(\mathbf{p})}}\approx\log T_{\text{SNR}}+\log\frac{1}{T_{\text{ICI}}}. \tag{2.35}$$

It is worth nothing here that the log-sum inequality could be applied to separate the second term above as the sum of inverses of the individual ICI terms. Further, considering the above sum of logarithm terms, it is clear that the MM technique can be applied. However, finding a surrogate function in this case is not as straightforward as in the solution to the MM-Lloyd algorithm discussed in Section 2.4.1. We believe that

the insight obtained from (2.35) is sufficient to generate a solution for AP placement. To simplify further, we negate the quantity in (2.35) and approximate using the relation $\log x < x$ which yields

$$-\log \phi_{k_{\mathcal{E}(\mathbf{p})}} < \frac{1}{T_{\mathrm{SNR}}} + T_{\mathrm{ICI}}. \tag{2.36}$$

We have now expressed the negative rate as the sum of the powers of the inverse of the desired and interference terms. Therefore, to maximize rate or equivalently, minimize the negative of the rate, we need to maximize SNR and minimize ICI, corresponding to the first and second terms in (2.36), respectively. The equation presented also reveals the structure of the distortion function that we will use.

Accordingly, in line with the objective function for a Lloyd-type algorithm in (2.7), we average (2.36) over the user positions and the random quantities defined in (2.33) above, to obtain

$$\mathbb{E}_{\mathcal{A},\mathcal{B},\underline{\mathbf{p}}}\left\{\frac{1}{T_{\mathrm{SNR}}} + T_{\mathrm{ICI}}\right\} = \mathbb{E}_{\mathbf{p}}\left\{\mathbb{E}_{\mathcal{A},\mathcal{B},\underline{\mathbf{p}}'}\left\{\frac{1}{T_{\mathrm{SNR}}} + T_{\mathrm{ICI}}\right\}\right\}, \tag{2.37}$$

where we denote $\underline{\mathbf{p}}'$ as the set of positions of the interfering users with respect to the user at $\mathbf{p}$. We have assumed in (2.37) that the served user position $\mathbf{p}$ is independent from the interfering user positions $\underline{\mathbf{p}}'$. We will also assume that as in (2.29) that the distribution of users in each interfering cell is independent and we can then write the joint distribution of users as

$$\begin{aligned} f_{\underline{\mathbf{P}}}(\underline{\mathbf{p}}) &= f_{\mathbf{P},\underline{\mathbf{P}}'}(\mathbf{p},\underline{\mathbf{p}}') = f_{\mathbf{P}}(\mathbf{p}) f_{\underline{\mathbf{P}}'}(\underline{\mathbf{p}}'), \\ &= f_{\mathbf{P}}(\mathbf{p}) \prod_{m' \neq \mathcal{E}(\mathbf{p})} f_{\mathbf{P}_{m'}}(\mathbf{p}_{m'} | \mathbf{p}_{m'} \in \mathcal{C}_{m'}), \end{aligned} \tag{2.38}$$

where $f_{\underline{\mathbf{P}}'}(\underline{\mathbf{p}}')$ is the joint distribution of the locations of all the interfering users and $f_{\mathbf{P}_{m'}}(\mathbf{p}_{m'}|\mathbf{p}_{m'} \in \mathcal{C}_{m'})$ is the distribution of the user in cell $\mathcal{C}_{m'}$. Consequently, by carrying out the expectations in (2.37) over $\mathcal{A}$, $\mathcal{B}$, and $\underline{\mathbf{p}}'$, we can write the distortion function as

$$d_{\mathrm{IF}}\left(\mathbf{p},\underline{\mathbf{q}}\right) = \kappa_1 \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma} + \kappa_2 \sum_{m' \neq \mathcal{E}(\mathbf{p})} \int \cdots \int \frac{1}{\left|\left|\mathbf{p}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}} f_{\underline{\mathbf{P}}'}(\underline{\mathbf{p}}') \mathrm{d}\underline{\mathbf{p}}', \tag{2.39}$$

where $\kappa_1 = \mathbb{E}_{\mathcal{A}}\{1/\rho_r c_1 z_{\mathcal{E}(\mathbf{p})} |h_{\mathcal{E}(\mathbf{p})}|^2\}$, $\kappa_2 = \mathbb{E}_{\mathcal{B}}\{\rho_r c_1 z_{m'} |h_{m'}|^2\}$, and the integration limits have been omitted for notation simplicity. This is the *interference distortion function* denoted by d_{IF} and the corresponding Lloyd-type algorithm is called the *Interference Lloyd algorithm*. Further simplification using (2.38) leads to a simpler distortion measure

$$d_{\mathrm{IF}}\left(\mathbf{p},\underline{\mathbf{q}}\right) = \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma} + \kappa \sum_{m' \neq \mathcal{E}(\mathbf{p})} \int_{\mathbf{p}_{m'} \in \mathcal{C}_{m'}} \frac{1}{\left|\left|\mathbf{p}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}} f_{\mathbf{P}_{m'}}(\mathbf{p}_{m'} | \mathbf{p}_{m'} \in \mathcal{C}_{m'}) \mathrm{d}\mathbf{p}_{m'}, \tag{2.40}$$

where $\kappa \triangleq \kappa_2/\kappa_1$. We call $\kappa \geq 0$ as the *trade-off factor* and it determines the trade-off between desired signal and ICI power. κ can be varied to determine the importance of ICI power over desired signal power.

To solve for the AP locations, the Interference Lloyd algorithm retains the NNC step and the steepest descent method is to be used for the CC step (update equation given in (2.25) above), both steps utilizing d_{IF}. For the sake of implementation, the integral in d_{IF} from (2.40) is numerically approximated using the sample average over a large number of realizations of the user locations, and is written as

$$d_{\mathrm{IF}}\left(\mathbf{p},\mathbf{q}_m\right) = \left|\left|\mathbf{p} - \mathbf{q}_m\right|\right|^{\gamma} + \kappa \sum_{m' \neq m} \frac{1}{|\mathcal{C}_{m'}|} \sum_{\mathbf{p}_{k_{m'}} \in \mathcal{C}_{m'}} \frac{1}{\left|\left|\mathbf{p}_{k_{m'}} - \mathbf{q}_m\right|\right|^{\gamma}}, \tag{2.41}$$

where $\mathbf{p}_{k_{m'}}$ represents the k^{th} realization of the user position in cell $\mathcal{C}_{m'}$. The gradient function in this update equation is

$$\begin{aligned}\frac{\partial}{\partial \mathbf{q}_m}\left\{\int\limits_{\mathbf{p}\in\mathcal{C}_m} d_{\text{IF}}\left(\mathbf{p},\mathbf{q}_m\right) f_{\mathbf{P}}(\mathbf{p})\mathrm{d}\mathbf{p}\right\} &= \frac{\gamma}{|\mathcal{C}_m|}\sum_{\mathbf{p}_k\in\mathcal{C}_m}\left(\mathbf{q}_m-\mathbf{p}_k\right)||\mathbf{p}_k-\mathbf{q}_m||^{\gamma-2} \\ &\quad+\kappa\sum_{m'\neq m}\frac{\gamma}{|\mathcal{C}_{m'}|}\sum_{\mathbf{p}_{k_{m'}}\in\mathcal{C}_{m'}}\frac{\left(\mathbf{p}_{k_{m'}}-\mathbf{q}_m\right)}{\left|\left|\mathbf{p}_{k_{m'}}-\mathbf{q}_m\right|\right|^{\gamma+2}}.\end{aligned} \tag{2.42}$$

The proof of this result is given in Appendix 2.E and the steps for this Lloyd-type algorithm are provided in Algorithm 5.

Algorithm 5 Interference Lloyd Algorithm

1: Initialize random AP locations $\mathbf{q}_1^{(0)},\mathbf{q}_2^{(0)},\ldots,\mathbf{q}_M^{(0)}$.

2: Use the NNC to determine the cells $\mathcal{C}_1^{(i+1)},\mathcal{C}_2^{(i+1)},\ldots,\mathcal{C}_M^{(i+1)}$ such that

$$\mathcal{C}_m^{(i+1)}=\left\{\mathbf{p}_k:d_{\text{IF}}\left(\mathbf{p}_k,\mathbf{q}_m^{(i)}\right)\leq d_{\text{IF}}\left(\mathbf{p}_k,\mathbf{q}_l^{(i)}\right),\forall l\neq m\right\}.$$

3: Use the steepest descent method to determine the AP locations $\mathbf{q}_1^{(i+1)},\mathbf{q}_2^{(i+1)},\ldots,\mathbf{q}_M^{(i+1)}$ with the update equation

$$\begin{aligned}\mathbf{q}_m^{(j+1)}=\mathbf{q}_m^{(j)}-\delta\Bigg(&\frac{\gamma}{\left|\mathcal{C}_m^{(i+1)}\right|}\sum_{\mathbf{p}_k\in\mathcal{C}_m^{(i+1)}}\left(\mathbf{q}_m^{(j)}-\mathbf{p}_k\right)\left|\left|\mathbf{p}_k-\mathbf{q}_m^{(j)}\right|\right|^{\gamma-2} \\ &+\kappa\sum_{m'\neq m}\frac{\gamma}{\left|\mathcal{C}_{m'}^{(i+1)}\right|}\sum_{\mathbf{p}_{k_{m'}}\in\mathcal{C}_{m'}^{(i+1)}}\frac{\left(\mathbf{p}_{k_{m'}}-\mathbf{q}_m^{(j)}\right)}{\left|\left|\mathbf{p}_{k_{m'}}-\mathbf{q}_m^{(j)}\right|\right|^{\gamma+2}}\Bigg),\end{aligned}$$

which, after convergence, $\mathbf{q}_m^{(i+1)}=\mathbf{q}_m^{(j+1)}$.

4: Repeat from step 2 until convergence.

2.5.2 Inter-AP Distortion Measure

Here, we develop an alternate distortion function that also accounts for ICI. Consider the interference distortion function d_{IF} in (2.40). Each of the ICI terms in the summation in d_{IF} can be approximated as follows

$$\mathbb{E}_{\mathbf{p}_{m'}}\left\{\frac{1}{\left|\left|\mathbf{p}_{m'}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}\right\}\approx\frac{1}{\left|\left|\mathbf{q}_{m'}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}, \tag{2.43}$$

the justification of which is provided in Appendix 2.D. Substituting (2.43), we can simplify (2.40) as

$$d_{\text{IA}}\left(\mathbf{p},\underline{\mathbf{q}}\right)=\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}+\kappa\sum_{m'\neq\mathcal{E}(\mathbf{p})}\frac{1}{\left|\left|\mathbf{q}_{m'}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}. \tag{2.44}$$

We call d_{IA} the *inter-AP distortion measure* as the ICI term now involves the distances between the interfering APs and AP indexed by $\mathcal{E}(\mathbf{p})$. The corresponding Lloyd-type algorithm is called the *Inter-AP Lloyd algorithm*.

The solution of the optimization problem using d_{IA} is similar to that of the Interference Lloyd algorithm. For the steepest descent method, the gradient corresponding to d_{IA} is given as

$$\begin{aligned}\frac{\partial}{\partial \mathbf{q}_m}\left\{\int\limits_{\mathbf{p}\in\mathcal{C}_m} d_{\mathrm{IA}}\left(\mathbf{p},\mathbf{q}_m\right) f_{\mathbf{P}}(\mathbf{p})\mathrm{d}\mathbf{p}\right\} &= \frac{\gamma}{|\mathcal{C}_m|}\sum_{\mathbf{p}_k\in\mathcal{C}_m}\left(\mathbf{q}_m-\mathbf{p}_k\right)||\mathbf{p}_k-\mathbf{q}_m||^{\gamma-2} \\ &\quad+\kappa\gamma\sum_{m'\neq m}\frac{\mathbf{q}_{m'}-\mathbf{q}_m}{||\mathbf{q}_{m'}-\mathbf{q}_m||^{\gamma+2}}.\end{aligned} \tag{2.45}$$

The proof of this result is omitted as it is similar in calculation to the gradient of the interference distortion function in (2.42) and the Inter-AP Lloyd algorithm is given in Algorithm 6. We also summarize the various formulations and solutions discussed in this section and the prior section in Table 2.1.

Algorithm 6 Inter-AP Lloyd Algorithm

1: Initialize random AP locations $\mathbf{q}_1^{(0)},\mathbf{q}_2^{(0)},\ldots,\mathbf{q}_M^{(0)}$.

2: Use the NNC to determine the cells $\mathcal{C}_1^{(i+1)},\mathcal{C}_2^{(i+1)},\ldots,\mathcal{C}_M^{(i+1)}$ such that

$$\mathcal{C}_m^{(i+1)}=\left\{\mathbf{p}_k:d_{\mathrm{IA}}\left(\mathbf{p}_k,\mathbf{q}_m^{(i)}\right)\leq d_{\mathrm{IA}}\left(\mathbf{p}_k,\mathbf{q}_l^{(i)}\right),\forall l\neq m\right\}.$$

3: Use the steepest descent method to determine the AP locations $\mathbf{q}_1^{(i+1)},\mathbf{q}_2^{(i+1)},\ldots,\mathbf{q}_M^{(i+1)}$ with the update equation

$$\mathbf{q}_m^{(j+1)}=\mathbf{q}_m^{(j)}-\delta\left(\frac{\gamma}{\left|\mathcal{C}_m^{(i+1)}\right|}\sum_{\mathbf{p}_k\in\mathcal{C}_m^{(i+1)}}\left(\mathbf{q}_m^{(j)}-\mathbf{p}_k\right)\left|\left|\mathbf{p}_k-\mathbf{q}_m^{(j)}\right|\right|^{\gamma-2}+\kappa\sum_{m'\neq m}\frac{\mathbf{q}_{m'}^{(j)}-\mathbf{q}_m^{(j)}}{\left|\left|\mathbf{q}_{m'}^{(j)}-\mathbf{q}_m^{(j)}\right|\right|^{\gamma+2}}\right),$$

which, after convergence, $\mathbf{q}_m^{(i+1)}=\mathbf{q}_m^{(j+1)}$.

4: Repeat from step 2 until convergence.

Among the distortion functions discussed above, it is evident that the MSE distortion d_{SE} has the lowest complexity. On observing the expressions for the interference d_{IF} and inter-AP d_{IA} distortions, we find that in the former, the summation for each interfering cell is over all of the users in that cell while in the latter, the net summation is only over interfering cells. Hence, d_{IA} has lower implementation complexity than d_{IF}. We will also see in a later section that user association with d_{IA} is relatively much simpler.

2.6 Cell Association Strategies

In the previous sections, we have addressed the problem of how to place APs based on the user locations. For completeness, we now aim at answering the following two questions on cell association: *When a new user enters the system, to which cell should it associate to? What metric should be used?* In this section, we elaborate on these two issues in the context of Lloyd and Lloyd-type algorithms. Accordingly, consider a user at location $\mathbf{p}_{\mathrm{new}}$ that has entered the area after AP placement has already occurred and will associate to the AP at $\mathbf{q}_{m_{\mathrm{new}}}$.

Table 2.1: Summary of Throughput Formulations and Solutions

Number of Users	Formulation	Solution Algorithm
Single user	Rate	MM-Lloyd
	SNR	Lloyd
	Higher exponent for user-AP distance	Lloyd-χ
Multiple users	Random user selection	Lloyd
	Random selection of one user per cell without ICI	WMSE Lloyd
	Random selection of one user per cell with ICI	Interference and Inter-AP Lloyd

For the Lloyd and Lloyd-type algorithms developed in this paper, the user would associate to the AP that yields the lowest distortion value. This is a straightforward implementation of the NNC for each algorithm. Formally, if d represents any of the distortion functions, $\mathbf{q}_{m_{\text{new}}}$ is determined as

$$\mathbf{q}_{m_{\text{new}}} = \{\mathbf{q}_m : d(\mathbf{p}_{\text{new}}, \mathbf{q}_m) \leq d(\mathbf{p}_{\text{new}}, \mathbf{q}_l), \forall l \neq m\}. \tag{2.46}$$

It is worth pointing out that since the distortion function in the Interference Lloyd algorithm involves summing over all users in other (interfering) cells, the complexity of such a calculation cannot be overlooked. Instead, a cell association procedure (2.46) based on the simpler distortion measures of the Lloyd or the Inter-AP Lloyd algorithm can be undertaken as a low-complexity alternative. Note that the distortion function in the latter involves only the knowledge of the interfering APs positions. This is of greater practical value as opposed to knowing the positions of all interfering users in the Interference Lloyd algorithm. In summary, the Inter-AP Lloyd algorithm not only offers lower implementation complexity and thus a simpler cell association strategy, but is also of more practical value compared to the Interference Lloyd algorithm.

2.7 Simulation Methodology and Results

2.7.1 Simulation Parameters

A geographical area of dimensions 2 km$\times$2 km is considered, consisting of $M = 8$ APs and $K = 2000$ users, and one randomly selected user in each cell communicates with its associated AP. The pathloss model in (3.2) is used with $\gamma = 2$, shadow fading z_{mk} ignored as it is averaged out in Sections 2.4 and 2.5, $c_0 = 75.86$ and $c_1 = 7.59 \times 10^{-7}$ as in [61, eq. (4.36), eq. (4.37)] according to the COST 231 Hata propagation model, and $r_0 = 0.001$ km. Also, the value of the trade-off factor is chosen to be $\kappa = 5 \times 10^8$ and the step-size for the gradient descent is $\delta = 5 \times 10^{-5}$ for the Lloyd-χ algorithm and $\delta = 0.5$ for the ICI-aware Lloyd-type algorithms. Moreover, the uplink transmit power is $\rho_r = 200$ mW and the user distribution is a Gaussian mixture model (GMM) of the form

$$f_{\mathbf{P}}(\mathbf{p}) = \sum_{l=1}^{L} p_l \mathcal{N}\left(\mathbf{p}|\boldsymbol{\mu}_l, \sigma_l^2 \mathbf{I}\right), \tag{2.47}$$

where $\mathbf{I}$ is the identity matrix and L is the number of mixture components, called *groups* henceforth. For group l, p_l is the mixture component weight, $\boldsymbol{\mu}_l$ is the mean, and σ_l is the standard deviation. We set a user configuration with the parameters $L = 3$, $\boldsymbol{\mu}_1 = [0.5, -0.5]^T$, $\boldsymbol{\mu}_2 = [0, 0.5]^T$, $\boldsymbol{\mu}_3 = [-0.5, 0]^T$, $\sigma_1 = \sigma_2 = \sigma_3 = 100$, $p_1 = 0.6$, and $p_2 = p_3 = 0.2$.

2.7.2 Performance Measures

We use the per-user achievable rate of user k_m, which is calculated using SINR ϕ_{k_m} from (3.5). As given in [61, Ch. 4], we can also write the achievable rate as

$$R_{k_m} = \mathbb{E}\left\{\log_2\left(1 + \phi_{k_m}\right)\right\} = \frac{1}{\ln 2} e^{\mu_k} \mathrm{Ei}\left(\mu_k\right), \tag{2.48}$$

where $\mu_k = (1 + \rho_r \sum_{m' \neq m} \beta_{mk_{m'}})/\rho_r \beta_{mk_m}$ and $\mathrm{Ei}(x) = \int_x^\infty (e^{-t}/t)\mathrm{d}t$ is the exponential integral.

For each of the proposed algorithms and the benchmark Lloyd algorithm, the maximum iteration number is set at 50. Each of the above performance measures is calculated through Monte Carlo simulations with $10,000$ iterations, choosing a set of users randomly for transmission each time. Cumulative distribution function (CDF) plots are generated for each measure, though normalized by the largest value so as to focus on the relative performance of the considered algorithms. For comparison, we utilize the 95%-likely metric that represents the best rate of the worst 5% of the users (users closer to cell borders). We denote this by $R_{k_m}^{5\%}$. To quantify the improvement in relative performance of the proposed algorithms over the Lloyd algorithm, we use the following measure expressed as percentage

$$\text{Improvement Ratio} = \frac{R_{k_m}^{5\%,\text{Proposed}} - R_{k_m}^{5\%,\text{Lloyd}}}{R_{k_m}^{5\%,\text{Lloyd}}} \times 100. \tag{2.49}$$

All algorithms are initialized with the same initial AP locations for unbiased comparison.

2.7.3 Numerical Results

Experiment 1. We compare the throughput performances of the proposed Lloyd-type algorithms in Section 2.4 with the baseline Lloyd algorithm. For the Lloyd-χ algorithm, we use $\chi = 4$ and we note that the rate calculations still use the exponent $\gamma = 2$. The AP locations resultant from the algorithms are shown in Fig. 2.1. Relative to the AP positions of the Lloyd algorithm which are shown as blue circles, the APs in both the MM-Lloyd and WMSE Lloyd algorithms are placed closer to the GMM centers. For the MM-Lloyd algorithm, this can be explained by the logarithm in its objective function which suppresses the effect of users which are at large distances (e.g., cell periphery users away from the GMM center) from the AP position during the placement process. This, in turn, causes the APs to position themselves closer to the GMM centers where the majority of the users at smaller distances are present. The WMSE Lloyd algorithm works in a different manner as the objective function in (2.30) is not weighted by the cell probabilities $\Pr(\mathbf{p} \in \mathcal{C}_m)$ as in (2.7) of the Lloyd algorithm. This allows cells in the WMSE Lloyd algorithm to have a larger number of users than the Lloyd algorithm. On the other hand, the objective function of the Lloyd-χ algorithm amplifies the contribution of the users at large distances and results in the APs moving away from

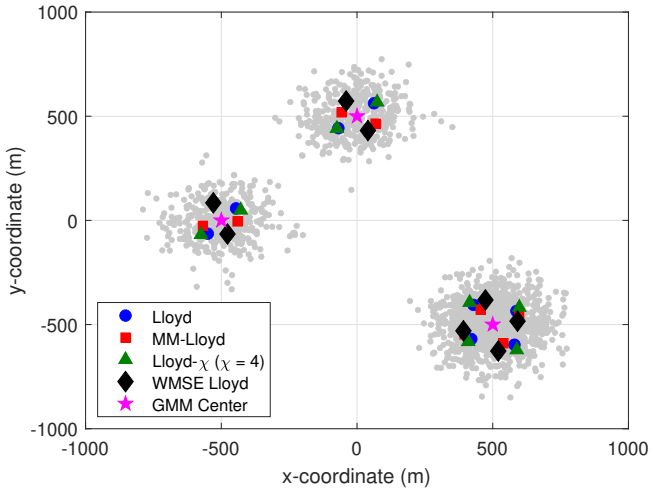

Figure 2.1: AP locations after convergence of the Lloyd, MM-Lloyd, Lloyd-χ ($\chi = 4$), and WMSE Lloyd algorithms with $M = 8$.

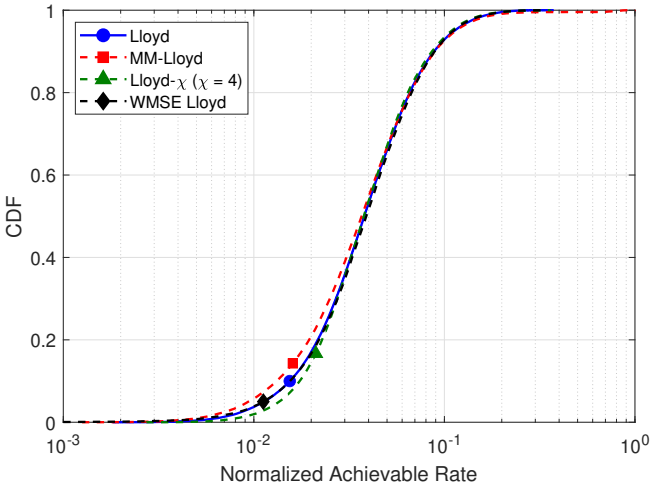

Figure 2.2: CDF plots of per-user achievable rate for the Lloyd, MM-Lloyd, Lloyd-χ ($\chi = 4$), and WMSE Lloyd algorithms with $M = 8$.

the GMM center. The effects of these placements are observed in their achievable rate plots in Fig. 2.2. For both the MM-Lloyd and WMSE Lloyd algorithms, we observe that due to their AP positions, the lower rate suffers a reduction in comparison to the Lloyd algorithm. Nevertheless, note that there are more users achieving higher rates (right side of the CDF plot), particularly for the MM-Lloyd algorithm. The average rate values, however, are higher than that of the Lloyd algorithm, up to about 4%, as shown in Table 2.2. On the other hand, the opposite of these effects are observed for the Lloyd-χ algorithm, with higher low rate values and lower average rate (only 0.65% lower) than the Lloyd algorithm. These effects were also found, through further experiments, to increase as the power χ increases.

Experiment 2. Here, our simulations show throughput performances for the proposed ICI-aware Lloyd-type algorithms and the Lloyd algorithm, as well as their respective AP placements for comparison.

Table 2.2: Percentage Improvements in Average Achievable Rates for the Lloyd-Type Algorithms of Section 2.4

Algorithm	Average Achievable Rate
MM-Lloyd	4.26%
Lloyd-χ ($\chi = 4$)	-0.64%
WMSE Lloyd	1.14%

The AP locations obtained after the algorithms converge are shown in Fig. 2.3. AP locations for the Lloyd algorithm are shown as circles around the GMM center, which in turn are shown by stars. Compared to these positions, we can observe that the AP locations for both the Lloyd-type algorithms are situated further away from the GMM centers. For the Interference Lloyd algorithm, the AP positions denoted by the squares are the farthest. This is due to the interference term in its distortion function that forces neighboring cells apart. This effect is different (smaller) for the Inter-AP Lloyd algorithm due to the inter-AP distances term in its distortion function in contrast to the interference term in the Interference Lloyd algorithm.

In Fig. 2.4, we show the CDFs of the achievable rate obtained per user for each of the considered algorithms. The horizontal line at the 5^{th} percentile shows the 95%-likely rate and we compare the values where it intersects the throughput curves. It is clear that accounting for ICI during the AP placement procedure yields a superior performance to both Lloyd-type algorithms in comparison to the Lloyd algorithm in terms of the 95%-likely rate. It is to be noted that the average rate performances of both the proposed algorithms are lower than that of the Lloyd algorithm, however, the magnitude of increase in the 95%-likely rates overshadows the decrease in average rates. This occurrence is due to the fact that the original average rate maximization problem in (2.33) has been transformed into maximization of its lower bound in (2.36). In practice, Fig. 2.4 shows us that the worst 5% of the users, usually the ones located closer to the cell borders and thus more susceptible to the deleterious effects of ICI, will have an uplink performance boost when APs are placed according to the proposed algorithms. The percentage of improvements are quantified in Table 2.3 from where we can confirm a very significant rate enhancement of up to 36.34% in the 95%-likely achievable rate, in comparison to the Lloyd algorithm. Also, from the same table, we can quantify that the Inter-AP Lloyd algorithm, despite its significantly lower computational complexity, performs slightly to moderately better than the Interference Lloyd algorithm, giving an approximately 3% improvement in the 95%-likely achievable rate. It is worth pointing out that in our experiments, lower κ values resulted in less improvements as the Lloyd-type algorithms approached the results of the Lloyd algorithm. Higher κ values resulted in convergence issues during the AP placement process. Many iterations of the algorithms were performed with other GMM configurations and κ values. Similar performance trends were observed for various standard deviations of the GMMs. Thus, the choice of κ is an important part of the AP placement process and depends primarily on the area under consideration and the pathloss model. Finally, it is important to notice that although we have focused on the worst 5% of the users, the Inter-AP Lloyd algorithm actually boosts the performances of the worst (nearly) 25% of the users. The performance loss of the best users, as seen in the CDF plot, is justifiable due to the fact that users closer to the cell center

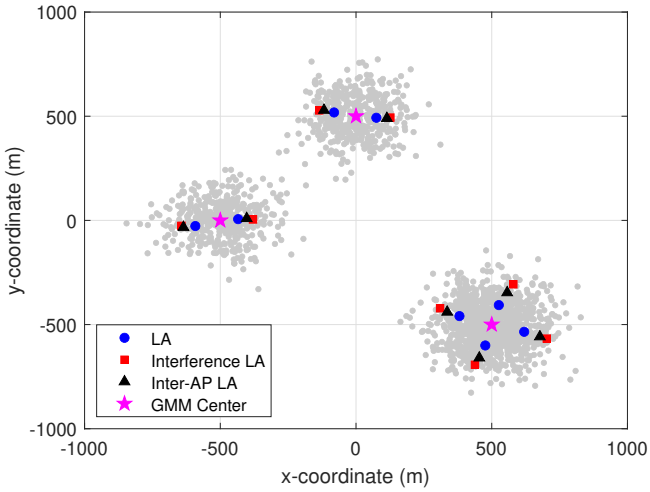

Figure 2.3: AP locations after convergence of the Lloyd and ICI-aware Lloyd-type algorithms with $\kappa = 5 \times 10^8$ and $M = 8$.

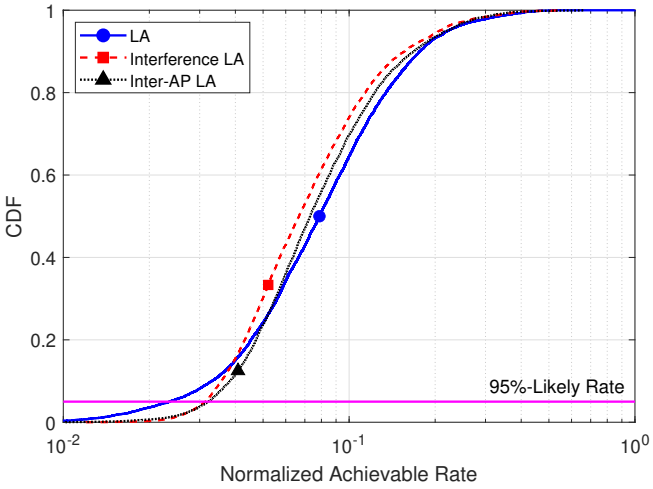

Figure 2.4: CDF plots of per-user achievable rate for the Lloyd and ICI-aware Lloyd-type algorithms with $\kappa = 5 \times 10^8$ and $M = 8$.

Table 2.3: Percentage Improvements in Average and 95%-Likely Achievable Rates for the ICI-Aware Lloyd-Type Algorithms

Algorithm	Average Rate	95%-Likely Rate
Interference Lloyd	−10.94%	33.37%
Inter-AP Lloyd	−4.35%	36.34%

tend to benefit from large SINR values that already suffice to provide them with more than their throughput requirements.

2.8 Conclusion

In this chapter, we have addressed the access point (AP) placement problem in the small-cell uplink paradigm under the criteria of throughput, while considering inter-cell interference (ICI). After reviewing vector quantization (VQ), we explored related throughput formulations in the single user case and subsequently, the multiple user case corresponding to the considered small-cell model. Without ICI, we showed that the simple Lloyd algorithm performed similarly to the aforementioned formulations (only up to a 4% difference) and could be a baseline algorithm to solve more complex problems. Accordingly, we accounted for ICI in the optimization function of the Lloyd algorithm and mathematically arrived at two distinct distortion functions. Correspondingly, we proposed two Lloyd-type algorithms, namely the Interference Lloyd algorithm and the Inter-AP Lloyd algorithm. Both algorithms yield significant improvement to achievable rates, giving up to a marked 36.34% increase in the 95%-likely rate over the benchmark Lloyd algorithm. The Inter-AP Lloyd algorithm achieves throughput gains coupled with lower complexity and simpler user association over the Interference Lloyd algorithm. Finally, cell association strategies were outlined for all algorithms for completeness.

Chapter 2, in part, is a reprint with permission of the material as it appears in the papers: Govind Ravikumar Gopal, Elina Nayebi, Gabriel Porto Villardi, and Bhaskar D. Rao, "Modified vector quantization for small-cell access point placement with inter-cell interference," in *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6387–6401, Aug. 2022, and Govind Ravikumar Gopal, Gabriel Porto Villardi, and Bhaskar D. Rao, "Is vector quantization good enough for access point placement?" in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, November 2021. The dissertation author was the primary investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-2124929, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

Appendices

2.A Proof of Solution for MM-Lloyd Algorithm

The expectation in the objective function in (2.17) can be replaced by the sample average using the user realizations at $\mathbf{p}_k$ and written as

$$J = \sum_{\mathbf{p}_k \in \mathcal{C}_m} \log\left(||\mathbf{q}_m - \mathbf{p}_k||^2 + \epsilon\right), \tag{2.50}$$

where $\mathbf{q}_{\mathcal{E}(\mathbf{p})}$ is replaced by $\mathbf{q}_m$ and the average is taken over all the users in cell $\mathcal{C}_m$ as the update steps correspond to the CC step of the Lloyd-type algorithm. Following the MM literature, a concave function can be upper bounded by its first-order Taylor expansion [66]

$$h(z) \leq h'(z_l)(z - z_l) + h(z_l), \tag{2.51}$$

where $h(\cdot)$ is concave on $\mathbb{R}^+$, z is the variable, z_l is the point around which the expansion is carried out, and $h'(\cdot)$ is the first derivative. In (2.50), we can take $h(z_k) = \log(z_k)$ and $z_k = ||\mathbf{q}_m - \mathbf{p}_k||^2 + \epsilon$, where we note that z_k is scalar. Thus, using the upper bound (5.33) in (2.50), the objective function is

$$J_1 = \sum_{\mathbf{p}_k \in \mathcal{C}_m} h(z_k) \leq \sum_{\mathbf{p}_k \in \mathcal{C}_m} [h'(z_{k,l})(z_k - z_{k,l}) + h(z_{k,l})]. \tag{2.52}$$

Removing the terms that are not involved in the optimization, we have

$$\arg\min_{\mathbf{q}_m} \sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k z_k = \arg\min_{\mathbf{q}_m} \sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k \left(||\mathbf{q}_m - \mathbf{p}_k||^2 + \epsilon\right), \tag{2.53}$$

where the weight w_k is defined as

$$\begin{aligned} w_k = h'(z_{k,l}) &= \left.\frac{\partial h(z_{k,l})}{\partial z_{k,l}}\right|_{z_{k,l} = ||\mathbf{q}_m - \mathbf{p}_k||^2 + \epsilon}, \\ &= \left.\frac{1}{z_{k,l}}\right|_{z_{k,l} = ||\mathbf{q}_m - \mathbf{p}_k||^2 + \epsilon} = \frac{1}{||\mathbf{q}_m - \mathbf{p}_k||^2 + \epsilon}, \end{aligned} \tag{2.54}$$

which gives the weight update equation. Now, given the weights, the objective function in (2.53) is

$$J_2 = \sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k \left(||\mathbf{q}_m - \mathbf{p}_k||^2 + \epsilon\right). \tag{2.55}$$

Taking the derivative and equating it to 0, i.e., $\partial J_2/\partial \mathbf{q}_m = 0$, gives the update equation for the AP position

$$\mathbf{q}_m = \frac{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m} w_k \mathbf{p}_k}{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m} w_k}. \tag{2.56}$$

2.B Proof of Gradient for Lloyd-χ Algorithm

The gradient of the distortion function $d_\chi(\mathbf{p}, \mathbf{q}_m)$ is calculated as

$$\begin{aligned} \frac{\partial}{\partial \mathbf{q}_m}\left\{ \int\limits_{\mathbf{p}\in\mathcal{C}_m} d_\chi(\mathbf{p}, \mathbf{q}_m) f_{\mathbf{P}}(\mathbf{p}) \mathrm{d}\mathbf{p} \right\} &\overset{(a)}{=} \frac{\partial}{\partial \mathbf{q}_m}\left\{ \frac{1}{|\mathcal{C}_m|} \sum_{\mathbf{p}_k \in \mathcal{C}_m} d_\chi(\mathbf{p}_k, \mathbf{q}_m) \right\}, \\ &= \frac{\partial}{\partial \mathbf{q}_m}\left\{ \frac{1}{|\mathcal{C}_m|} \sum_{\mathbf{p}_k \in \mathcal{C}_m} ||\mathbf{p}_k - \mathbf{q}_m||^\chi \right\}, \\ &\overset{(b)}{=} \frac{\chi}{|\mathcal{C}_m|} \sum_{\mathbf{p}_k \in \mathcal{C}_m} (\mathbf{q}_m - \mathbf{p}_k) ||\mathbf{p}_k - \mathbf{q}_m||^{\chi-2}. \end{aligned} \tag{2.57}$$

where (a) is obtained by replacing the expectation with the sample mean and the factor of 2 is assumed to be absorbed by the step-size δ in (b).

2.C Proof of Solution for WMSE Lloyd Algorithm

Consider the simplified objective function in (2.30), which can be written as

$$\begin{aligned} \sum_{m=1}^{M} S_m &\overset{(a)}{=} \sum_{m=1}^{M} \frac{1}{N_m} \sum_{\mathbf{p}_k \in \mathcal{C}_m} d_{\mathrm{SE}}(\mathbf{p}_k, \mathbf{q}_m), \\ &\overset{(b)}{=} \frac{1}{K} \sum_{m=1}^{M} \frac{1}{\Pr(\mathbf{p} \in \mathcal{C}_m)} \sum_{\mathbf{p}_k \in \mathcal{C}_m} d_{\mathrm{SE}}(\mathbf{p}_k, \mathbf{q}_m), \end{aligned} \tag{2.58}$$

where in (a), we have replaced the integral of S_m (defined in (2.8)) with the sample average over the users present in the cell and N_m represents the number of users in cell $\mathcal{C}_m$, and in (b), we have used $\Pr(\mathbf{p} \in \mathcal{C}_m) = N_m/K$, with K as the total number of users. Comparing (2.58) with the objective function of the Lloyd algorithm J_{VQ} in (2.7), we have

$$\begin{aligned} \sum_{m=1}^{M} S_m \Pr(\mathbf{p} \in \mathcal{C}_m) &= \sum_{m=1}^{M} \frac{1}{N_m} \sum_{\mathbf{p}_k \in \mathcal{C}_m} d_{\mathrm{SE}}(\mathbf{p}_k, \mathbf{q}_m) \times \frac{N_m}{K}, \\ &= \frac{1}{K} \sum_{m=1}^{M} \sum_{\mathbf{p}_k \in \mathcal{C}_m} d_{\mathrm{SE}}(\mathbf{p}_k, \mathbf{q}_m), \end{aligned} \tag{2.59}$$

where we can observe that the objective function in (2.58) is a weighted MSE (WMSE) measure, with the weight related to AP m given by $w_m = 1/\Pr(\mathbf{p} \in \mathcal{C}_m)$. Thus, the NNC step is updated to use a weighted squared error distortion function, i.e., $w_m d_{\mathrm{SE}}(\mathbf{p}_k, \mathbf{q}_m)$. The CC step however remains independent of the

weights. This can be proven by taking the derivative of the objective function in (2.58) with respect to the AP location $\mathbf{q}_m$, which gives the AP location as

$$\mathbf{q}_m = \frac{1}{|\mathcal{C}_m|} \sum_{\mathbf{p}_k \in \mathcal{C}_m} \mathbf{p}_k. \tag{2.60}$$

2.D Justification of (2.43)

Consider the term in the denominator of (2.43)

$$||\mathbf{p}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}||^2 = ||\underbrace{\mathbf{p}_{m'} - \mathbf{q}_{m'}}_{\mathbf{y}} + \underbrace{\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}}_{\mathbf{x}}||^2. \tag{2.61}$$

Let us assume that the distance between the interfering user and its serving AP, denoted by $\mathbf{y}$, is always smaller than the distance of that same AP from the nearest AP, denoted by $\mathbf{x}$, which means

$$||\mathbf{p}_{m'} - \mathbf{q}_{m'}|| \leq \left|\left|\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right| \Rightarrow ||\mathbf{y}|| \leq ||\mathbf{x}||. \tag{2.62}$$

The above inequality always holds true when the interfering user is on the near side of the interfering AP with respect to the serving AP and does not always hold true when the interfering user is on the far side of the interfering AP with respect to the serving AP. We note that the importance of the second scenario is reduced when the interfering cells are farther away. Further, even among the interfering cells that are near (the neighboring cells), the proportion of users within such cells that does not satisfy the inequality above is small. Thus, we make the assumption that (2.62) is satisfied for all users. Given that m' indexes the interfering cells, we can classify these cells into cells that are the immediate neighbors of cell $\mathcal{C}_{\mathcal{E}(\mathbf{p})}$, denoted by $\mathcal{IN}(\mathcal{E}(\mathbf{p}))$ and those that are not, and are thus farther away. Hence, the two cases are

$$\begin{aligned} &||\mathbf{x}|| \geq ||\mathbf{y}||, \quad \forall m' \in \mathcal{IN}(\mathcal{E}(\mathbf{p})), \quad m' \neq \mathcal{E}(\mathbf{p}), \\ &||\mathbf{x}|| \gg ||\mathbf{y}||, \quad \forall m' \notin \mathcal{IN}(\mathcal{E}(\mathbf{p})), \quad m' \neq \mathcal{E}(\mathbf{p}). \end{aligned} \tag{2.63}$$

However, to simplify, we make the optimistic assumption that $||\mathbf{x}|| \gg ||\mathbf{y}||$ holds true for all $m' \neq \mathcal{E}(\mathbf{p})$. This gives

$$\begin{aligned} ||\mathbf{x}+\mathbf{y}||^2 &= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 + 2||\mathbf{x}||||\mathbf{y}||\cos\theta, \\ &= ||\mathbf{x}||^2 \left(1 + \frac{||\mathbf{y}||^2}{||\mathbf{x}||^2} + \frac{||\mathbf{y}||\cos\theta}{||\mathbf{x}||}\right), \\ &\approx ||\mathbf{x}||^2. \end{aligned} \tag{2.64}$$

Note that this relation holds true even when γ assumes values other than $\gamma = 2$. Thus, from (3.43), we have

$$\mathbb{E}_{\mathbf{p}_{m'}} \left\{ \frac{1}{\left|\left|\mathbf{p}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}} \right\} \approx \frac{1}{\left|\left|\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}. \tag{2.65}$$

2.E Proof of Gradient for Interference Lloyd Algorithm

The gradient is calculated using the distortion function as

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{q}_m}\left\{\int\limits_{\mathbf{p}\in\mathcal{C}_m} d_{\mathrm{IF}}\left(\mathbf{p},\mathbf{q}_m\right) f_{\mathbf{P}}(\mathbf{p})\mathrm{d}\mathbf{p}\right\} \\
&\qquad = \frac{\partial}{\partial \mathbf{q}_m}\left\{\frac{1}{|\mathcal{C}_m|}\sum_{\mathbf{p}_k\in\mathcal{C}_m} d_{\mathrm{IF}}\left(\mathbf{p}_k,\mathbf{q}_m\right)\right\}, \\
&\qquad = \frac{\partial}{\partial \mathbf{q}_m}\left\{\frac{1}{|\mathcal{C}_m|}\sum_{\mathbf{p}_k\in\mathcal{C}_m} ||\mathbf{p}_k-\mathbf{q}_m||^{\gamma} + \kappa\sum_{m'\neq m}\frac{1}{|\mathcal{C}_{m'}|}\sum_{\mathbf{p}_{k_{m'}}\in\mathcal{C}_{m'}}\frac{1}{\left|\left|\mathbf{p}_{k_{m'}}-\mathbf{q}_m\right|\right|^{\gamma}}\right\}, \\
&\qquad = \frac{\gamma}{|\mathcal{C}_m|}\sum_{\mathbf{p}_k\in\mathcal{C}_m}\left(\mathbf{q}_m-\mathbf{p}_k\right)||\mathbf{p}_k-\mathbf{q}_m||^{\gamma-2} + \kappa\sum_{m'\neq m}\frac{\gamma}{|\mathcal{C}_{m'}|}\sum_{\mathbf{p}_{k_{m'}}\in\mathcal{C}_{m'}}\frac{\left(\mathbf{p}_{k_{m'}}-\mathbf{q}_m\right)}{\left|\left|\mathbf{p}_{k_{m'}}-\mathbf{q}_m\right|\right|^{\gamma+2}},
\end{aligned} \tag{2.66}$$

where the factor of 2 is assumed to be absorbed by the step-size δ as in Appendix 2.B.

Chapter 3

Small-Cell AP Placement with AP Hybridity and with Load Balancing

In this chapter, we consider two additional problem frameworks to throughput oriented small-cell AP placement.

First, we address the problem of access point (AP) placement in small-cell networks with partial infrastructure flexibility, i.e., a novel class of problem in Beyond 5G, resultant from the utilization of unmanned aerial vehicles (UAVs) with AP functionalities (UAV-APs), to aid fixed wireless networks in coping with momentary peak-capacity requirements. We use the signal-to-generated-interference-plus-noise ratio (SGINR) metric as an alternative to the traditional signal-to-interference-plus-noise ratio (SINR) to quantify the effects of inter-cell interference (ICI) on the per-user capacity. From average SGINR, we derive the ICI-aware distortion measure leading to the Inter-AP Lloyd algorithm to obtain throughput-optimal AP placement for a fully flexible infrastructure. We then impose a hybridity constraint to the AP placement problem which turns a fraction of the network into a fixed infrastructure composed of terrestrial APs (T-APs) while the remainder is constituted by UAV-APs with flexibility in position. This newly formulated AP placement problem is solved by the proposed Lloyd-type algorithm called Hybrid AP Placement Algorithm (HAPPA). Furthermore, we present an initialization method for the Lloyd and Lloyd-type algorithms for Gaussian mixture models (GMMs) that offers an AP allocation leading to a higher rate compared to the k-means++ initialization. Finally, computer simulations show that the Inter-AP Lloyd algorithm can improve the performance of the worst users by up to 42.75% in achievable rate, assuming a fully flexible network. By using HAPPA on hybrid networks, we achieve improvements of up to 71.92% in sum rate over the fixed network and close the performance gap with fully flexible networks down to 2.02%, when an equal number of UAV-APs and T-APs is used. Further, our proposed initialization scheme always results in a balanced AP allocation, which means a more even distribution of users per AP, whereas the k-means++ scheme results in unbalanced allocations at least 30% of the time, resulting in a worse minimum rate.

Second, we address the uplink small-cell access point (AP) placement problem for optimal through-

put, while considering load balancing (LB) among the APs. To consider LB and consequently incorporate fairness in user spectral access, i.e., the frequency of user-to-AP communications, we modify the Lloyd algorithm from vector quantization so that delays incurred by the existence of a large number of users in a cell are accounted for in the AP placement process. Accordingly, we present two methods, the first of which involves the incorporation of weights proportional to the cell occupancy, hence called the Occupancy Weighted Lloyd Algorithm (OWLA). The second method adds a new step to the Lloyd algorithm, which involves re-assigning users from higher to lower occupancy cells, and the adoption of a distance threshold to cap the throughput lost in the assignment process. This formulated Lloyd-type algorithm is called the Cell Equalized Lloyd Algorithm-α (CELA-α) where α is a factor that allows for throughput and spectrum access delay trade-off. Extensive simulations show that both CELA-α and OWLA algorithms provide significant gains, in comparison to the standard Lloyd algorithm, in 95%-likely user spectral access. For the α values considered in this paper, CELA-α achieves gains up to 20.83%, while OWLA yields a gain of 12.5%. Both algorithms incur minimal throughput losses of different degrees, and the choice of using one algorithm over the other for AP placement depends on system LB as well as throughput requirements.

3.1 Introduction

Hybridity

Recently, unmanned aerial vehicles (UAVs) have been conceived as a means of providing wireless connectivity for the Beyond 5G paradigm. Most of the prior works focus on UAV-terrestrial propagation channel models, routing and energy efficiency, and applications of UAVs to ad hoc networks, civil applications, and Internet-of-Things. Cellular-connected UAVs have seen an increased interest [67–69]. Additionally, the mobility inherent to UAVs has made the topic of small-cell AP placement gain even more traction recently. For instance, [59] provides an in-depth tutorial on UAVs, from both the perspective of end-users and APs (UAV-APs). References [38, 52–56, 70–79] (and references therein) provide some examples of UAV placement techniques. The authors in [52] propose the deployment of UAVs at demand hotspots along with an incumbent macro-cell network, by intelligently off-loading sets of users obtained using the K-means algorithm. In [53], mixed integer non-linear optimization with a quadratic constraint is employed to determine the 3-D positions of a single UAV. Placement, however does not aim at throughput maximization and instead attempts to increase coverage to users not previously covered by a terrestrial AP (T-AP). The work in [70] considers the placement of UAV-APs to maximize ground user coverage and alters their altitude to minimize interference to users, however, without considering an existing ground network. Authors in [71] expand the integer non-linear optimization to include a minimum transmit power constraint and separate horizontal and vertical placements. Heuristic successive placement of UAVs with a fixed ground coverage radius in a spiral fashion is performed in [54] to ensure coverage for all users. A cloud radio access network is considered in [72] consisting of terrestrial remote radio heads (RRHs) and UAVs. The positions and contents of the cache-enabled UAVs are optimized, but interference between RRH-user and UAV-user links is not modeled. The work in [73] considers a hybrid architecture consisting of a single ground base station

and UAVs flying cyclically around the cell edge to offload users and improve minimum user throughput. By considering a one-hop or two-hop downlink communication scheme between T-APs, UAV-APs, and users, [55] investigates UAV placement coupled with bandwidth and power allocation in the backhaul and access links, however, leaving out ICI considerations. A spectrum sharing scenario is considered in [74] in which UAVs communicate to secondary receivers while minimizing interference to the primary terrestrial transmitter-receiver pair. Two scenarios are considered to maximize the rate of the secondary receiver - one, where the 3-D placements of static UAVs are optimized, and the other, where the trajectories of mobile UAVs are optimized, both with power control. Authors in [75] also address the ground coverage problem (as in [53]). Unlike the previous works, [56] utilizes a user density-driven 3-D UAV placement to attempt maximum coverage of users with minimum data rate requirements. Also, maximizing throughput (utilizing the SNR alone) for mobile users along with random network coding packet scheduling is jointly considered for UAV-AP placement in [76]. In [77], authors take into account the time varying nature of the users' positions to find the optimal placement and coverage radius of the UAV-APs by using cooperative stochastic approximation for Wi-Fi (IEEE 802.11) networks. [78] explores the 3-D deployment of multiple UAVs having directional antennas to minimize the average transmit power of the users. The work in [79] develops user access-based trajectory design for UAV-APs using a value decomposition reinforcement learning algorithm, trained using meta training. Finally, [38] considers UAV-APs to offload traffic from terrestrial networks. Using a weighted expectation maximization algorithm, the user distribution and traffic demand are estimated so that the UAV-APs can be placed by maximizing the utility of the overloaded AP. However, interference between the terrestrial and aerial communication networks is neglected.

None of the works described above, especially the ones that study multiple UAVs, consider ICI during the placement process. Hence, with the Lloyd algorithm as the basis, [60] optimizes rate using the signal-to-interference-plus-noise ratio (SINR) and proposes an alteration to the mean squared error (MSE) distortion function which accounts for ICI and succeeds in improving the minimum rate of the small-cell system over the traditional Lloyd algorithm. Following this direction, in [80], we derive a Lloyd-type algorithm (called the Inter-AP Lloyd algorithm) with a simpler, yet effective distortion measure to deal with ICI when compared to the distortion function of the algorithm presented in [60]. Note that these discussions were completed in the previous chapter. It is important to note also that all the above papers consider networks with fully flexible infrastructure, in which all APs are position-flexible and can have their locations updated according to the user distribution at a certain time.

In this chapter, we investigate the AP placement problem, with the objective of maximizing throughput in a small-cell system, with prevalent ICI and varying user distributions. More specifically, we investigate the placement of APs susceptible to ICI in networks with partial infrastructure flexibility, interchangeably referred to as hybrid networks, which is composed of fixed-position T-APs and flexible-position UAV-APs. The degree of infrastructure flexibility varies between the two extremes determined by fully flexible networks and fixed networks, consisting of either only UAV-APs or only T-APs, respectively. In practice, hybrid networks can be deployed in a fast and dynamical manner to tackle the random nature of capacity due to user mobility, to compensate an expected increase in throughput requirements of users attending an event,

or even, a sudden reduction in network capacity due to wireless infrastructure being knocked over as a result of natural disasters. From a wireless operator point-of-view, building a fixed infrastructure with high T-AP density, for instance, in a football stadium, will render the network underutilized most of the time. On the other hand, offloading T-APs by deploying UAV-APs on-demand is more economically appealing. Also, wildfires, earthquakes, and the ensuing tsunamis have, in the recent past, disrupted communication services, leaving populations isolated [81, 82]. Complementing the part of a network that has withstood a natural disaster with UAV-APs is speedy and desirable; however, it requires careful placement considerations since ICI has been shown to significantly hinder the coverage area of emergency network services [40].

Load Balancing

Besides throughput, delay in spectrum access is also a relevant system design parameter, especially for the delay sensitive applications expected in the deployment of Beyond 5G networks. In systems where throughput alone is optimized, either by using the Lloyd algorithm for SNR optimization only or the Lloyd-type algorithm by incorporating ICI in the problem formulation, the cells have unequal occupancies, that is to say, unequal number of users, after the algorithm converges, thus yielding placement with sub-optimal fairness in spectral usage. In other words, this results in an unbalanced distribution of users across the cells, which in turn leads to users of cells with lower occupancy having more opportunities to access the spectrum over users of other cells. Naturally, the question that arises is: *How to efficiently perform user-cell association so that users are ensured opportunity to access spectrum without undue delay?* Defining spectrum access delay as the time that a user waits for its opportunity to communicate with its assigned AP, we create a metric called *spectral access fraction* in order to allow us to quantify the access delay of the proposed algorithms. One approach is to equalize the occupancy of each cell by re-assigning users from cells with higher-than-average occupancy (among all cells), to cells with lower-than-average occupancy. The objective of this procedure is balancing the cell loads, following the motivation behind cell breathing [83–86]. Nevertheless, the main drawback of this strategy is that users can be moved to far away cells and therefore suffer a significant reduction in overall throughput. To this end, our work aims to create a desirable and flexible trade-off between throughput reduction and increase in the spectral access fraction. We therefore devise algorithms for non-cooperative small-cell systems that modify the Lloyd algorithm yielding AP placements that maximize throughput while minimizing spectrum access delay, thereby considering load balancing (LB) among the small-cells.

Contributions

Solutions to the AP placement problem for throughput maximization for hybrid UAV-terrestrial small-cell networks while incorporating ICI into the placement process have not been discussed in past literature. Additionally, solutions to the small-cell AP placement problem based on the Lloyd algorithm and that jointly address throughput and spectrum access delay (incorporating LB), have also not been provided in literature. Hence, in this chapter, our contributions are as follows.

- To include ICI in the AP placement process and with the Lloyd algorithm as the basis, we employ an alternate metric to SINR, namely, the signal-to-generated-interference-plus-noise ratio (SGINR).

Optimizing the average SGINR, we derive the distortion measure that accounts for ICI through inter-AP distances, establishing the fact that the optimization of SGINR, like SINR, can lead to the Inter-AP Lloyd algorithm, which is known to perform better than the Lloyd algorithm in terms of the rates of the users most affected by ICI.

- Based on the Inter-AP Lloyd algorithm, we formulate and develop an algorithm for hybrid networks that places UAV-APs in an area occupied by T-APs currently providing sub-optimal throughput performances due to a change in user configuration. This is the Hybrid AP Placement Algorithm (HAPPA) that outperforms the T-APs (fixed network) alone. It also provides a performance close to the ideal one of fully flexible networks, despite the fact that only a fraction of the APs can have their positions adjusted to maximize throughput (minimize ICI).

- For unbiased comparison of the presented algorithms, we develop an initialization scheme for the Gaussian mixture model (GMM) user distribution, based on the bit allocation procedure in vector quantization (VQ). The proposed scheme always offers a balanced allocation of APs that results in each AP serving a relatively similar number of users compared to the k-means++ scheme, which can result in unbalanced allocations. The unbalanced allocations are shown to yield a worse minimum rate over the balanced one.

- To account for LB, the Lloyd algorithm is modified to incorporate weights chosen to prevent users from associating with APs having a large occupancy. This weighted Lloyd algorithm is hereafter referred to as the Occupancy Weighted Lloyd Algorithm (OWLA), and considers LB and throughput altogether.

- An alternate LB procedure also based on the Lloyd algorithm that re-assigns users between cells is proposed. By prioritizing and re-assigning users from higher to lower occupancy cells, the joint effect of throughput and delay is addressed. Moreover, in order to control the trade-off between throughput and spectrum access delay, the distance threshold used incorporates a factor α. The Lloyd-type algorithm created is called the Cell Equalized Lloyd Algorithm-α (CELA-α).

3.2 System Model

The small-cell system model utilized here is similar to the one used in the previous chapter, however, repeated here for completeness. We consider a geographical area with K single-antenna users distributed according to a probability density function (pdf) $f_{\mathbf{P}}(\mathbf{p})$, where $\mathbf{p} \in \mathbb{R}^2$ is the random vector denoting the position of a user. These users are served by M single-antenna APs, designated by the set $\mathcal{M}$ such that $|\mathcal{M}| = M$. The location of an AP is denoted by $\mathbf{q} \in \mathbb{R}^2$. Following the theme of a hybrid network, we divide the set of APs $\mathcal{M}$ into T-APs and UAV-APs, denoted by the sets $\mathcal{M}_f$ and $\mathcal{M}_u$, respectively, such that $|\mathcal{M}_f| = M_f$, $|\mathcal{M}_u| = M_u$, $\mathcal{M}_f \cup \mathcal{M}_u = \mathcal{M}$, and $M_f + M_u = M$. For simplicity, the position-flexible UAV-APs are assumed to have the same transmission parameters as the fixed T-APs. Although 3-D channel models with height as a parameter exist for modeling the channel between users and UAV-APs (e.g., [87]), references [88] and [89] use a fixed-height log-distance pathloss model between UAVs and users. We follow

this approach and use the same pathloss model between the users and the UAV-APs as well as between the users and T-APs. Note that when we consider the load balancing problem, we do not differentiate between these T-APs and UAV-APs; they are collectively referred to as APs. Moreover, a narrowband flat-fading channel is considered and the channel coefficient between AP m and user k is given by

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}, \tag{3.1}$$

where $m = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, K$. Here, β_{mk} and h_{mk} are the large-scale and small-scale fading coefficients, respectively. $h_{mk} \sim \mathcal{CN}(0, 1)$ is assumed to remain constant in a coherent interval and change independently in the next. Further, h_{mk} is independent of other small-scale coefficients as well as of β_{mk}. The large-scale coefficients are modeled as

$$\beta_{mk} = \begin{cases} c_0, & ||\mathbf{p}_k - \mathbf{q}_m|| \leq r_0, \\ \frac{c_1 z_{mk}}{||\mathbf{p}_k - \mathbf{q}_m||^\gamma}, & ||\mathbf{p}_k - \mathbf{q}_m|| > r_0, \end{cases} \tag{3.2}$$

where $\mathbf{p}_k$ and $\mathbf{q}_m$ represent the locations of user k and AP m, respectively. Here, γ is the pathloss exponent, z_{mk} is the log-normal shadow fading coefficient, and c_0, c_1, and r_0 are constants. Note that these coefficients can also be estimated by either ray-tracing [62] or data-driven [63] approaches.

We assume an uplink model where each AP serves a subset of the users (cell $\mathcal{C}_m$ for AP m). Users are scheduled in a round robin fashion according to time-division multiple access (TDMA), and each AP serves only one user in a time slot (thus, there are only M users transmitting at a given time). All APs are in turn connected to a network controller (NC) via high capacity backhaul links and the NC is assumed to have knowledge of the positions of all APs and their respective users. Thus, it is in the NC where the AP placement algorithms will be loaded and executed. In the small-cell set-up, each of the M cells corresponds to each of the M APs, and pursuant with the uplink model, users in their cells communicating with their serving APs cause interference to all other APs. Now, letting k_m denote a user in the cell associated with AP m, the received signal y_m at this AP is

$$y_m = \sum_{m'=1}^{M} \sqrt{\rho_r} g_{mk_{m'}} s_{k_{m'}} + w_m, \tag{3.3}$$

where ρ_r is the uplink transmit power, s_{k_m} is the data symbol with $\mathbb{E}\{|s_{k_m}|^2\} = 1$ (unit power), and $w_m \sim \mathcal{CN}(0, 1)$ is the additive noise. A matched filter (MF) employed at the AP m estimates the data symbol s_{k_m} of user k_m as

$$\begin{aligned} \hat{s}_{k_m} &= \frac{g^*_{mk_m}}{|g_{mk_m}|} y_m, \\ &= \sqrt{\rho_r} |g_{mk_m}| s_{k_m} + \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \sqrt{\rho_r} \frac{g^*_{mk_m}}{|g_{mk_m}|} g_{mk_{m'}} s_{k_{m'}} + v_m, \end{aligned} \tag{3.4}$$

where $v_m \sim \mathcal{CN}(0, 1)$. From this expression, the SINR achieved by user k_m at AP m can be determined as

$$\phi_{k_m} = \frac{\rho_r \beta_{mk_m} |h_{mk_m}|^2}{1 + \rho_r \sum\limits_{\substack{m'=1 \\ m' \neq m}}^{M} \beta_{mk_{m'}} |h_{mk_{m'}}|^2}. \tag{3.5}$$

3.3 Vector Quantization Framework and AP Placement

Although the VQ framework and its connection to small-cell AP placement have been covered in the previous chapter, we provide a brief overview here to make this chapter self-sufficient.

The VQ framework, described in [41], can be applied to the AP placement problem by assuming that the position $\mathbf{p}$ of a single user is the input random vector to be quantized, the Voronoi regions are the cells $\mathcal{C}_m$, and the AP locations $\mathbf{q}_m$ are the output codepoints, for $m = 1, 2, \ldots, M$. Accordingly, the optimization problem to be solved is

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\min}\ \mathbb{E}_{\mathbf{p}}\left\{d(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})})\right\}, \tag{3.6}$$

where d is the distortion function that measures the quantization error and $\mathcal{E}$ refers to the encoder of the quantizer. It is to be noted that although d is defined for a user at $\mathbf{p}$, the objective function averages over the entire user distribution. Also, $\mathbf{q}_{\mathcal{E}(\mathbf{p})}$ is the position of the nearest AP, in terms of d, to the user at $\mathbf{p}$ and this is formally expressed as

$$\mathbf{q}_{\mathcal{E}(\mathbf{p})} = \underset{\mathbf{q}_m}{\arg\min}\ d(\mathbf{p}, \mathbf{q}_m). \tag{3.7}$$

Solution of (3.6) requires design of both encoder and decoder. It involves finding the best encoder given a fixed decoder using the *Nearest Neighbor Condition (NNC)*

$$\mathcal{C}_m = \{\mathbf{p} : d(\mathbf{p}, \mathbf{q}_m) \leq d(\mathbf{p}, \mathbf{q}_l), \forall l \neq m\}, \tag{3.8}$$

and finding the best decoder given a fixed encoder using the *Centroid Condition (CC)*

$$\mathbf{q}_m = \mathrm{Cent}\{\mathbf{p}|\mathbf{p} \in \mathcal{C}_m\}, \tag{3.9}$$

where Cent calculates the centroid of the user positions in cell $\mathcal{C}_m$. To solve, the Lloyd algorithm alternates between the NNC and CC steps until convergence to find the optimal AP locations. The most common distortion measure used is the squared Euclidean distance, denoted by

$$d_{\mathrm{SE}}(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})}) = \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2, \tag{3.10}$$

and the objective function in (3.6) then becomes the mean squared error (MSE). We utilize this distortion measure in the Lloyd algorithm and it is provided in Algorithm 12. For algorithm implementation, we use the K realization of users at positions $\mathbf{p}_k$, $k = 1, 2, \ldots, K$, as described in Section 5.2. We will use this notation for all the algorithms in this chapter. Also, in the CC step, as a result of d_{SE}, the centroid calculation to obtain $\mathbf{q}_m$ is replaced by the expectation which is evaluated by using the sample average over the user positions $\mathbf{p}_k$ present in cell $\mathcal{C}_m$. The overall complexity of the Lloyd algorithm is $\mathcal{O}(KMI_L)$, where $\mathcal{O}$ is the 'Big O' notation and I_L is the number of iterations taken for the algorithm to converge. Finally, the term 'Lloyd algorithm' will henceforth be used to denote the basic Lloyd algorithm above using the squared error distance as the distortion measure to solve for the AP locations. For algorithms that use other distortion measures, i.e., the Inter-AP Lloyd algorithm and the Hybrid AP Placement Algorithm (HAPPA) below, we will use the term 'Lloyd-type algorithm'.

Algorithm 7 Lloyd Algorithm With Squared Error Distortion

1: **Initialize:**

$\mathbf{q}_m^{(0)}$, $\forall m \in \mathcal{M}$

$i \leftarrow 0$

2: **while** total distortion less than threshold **do**

3: update $\mathcal{C}_m^{(i+1)}$, $\forall m \in \mathcal{M}$ using ▷ NNC

$$\mathcal{C}_m^{(i+1)} \leftarrow \left\{\mathbf{p}_k : d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m\right\}$$

4: update $\mathbf{q}_m^{(i+1)}$, $\forall m \in \mathcal{M}$ using ▷ CC

$$\mathbf{q}_m^{(i+1)} \leftarrow \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \mathbf{p}_k$$

5: $i \leftarrow i + 1$

6: **return** $\mathbf{q}_m$, $\forall m \in \mathcal{M}$

Besides the parallels between VQ and the AP placement process discussed at the beginning of this section, we proved mathematically in the previous chapter that the maximization of the average SNR can result in the same optimization problem of the Lloyd algorithm in (3.6) above, that uses d_{SE}. Further, we found that the Lloyd algorithm could be used as a baseline algorithm and hence, we use it to solve for more complex problems, e.g., accounting for ICI, by modifying its distortion measure.

In addition, it is also important to note here that the VQ framework presented above quantifies the distortion d for a single user at position $\mathbf{p}$ and the average over the distribution of user positions is taken. Finally, VQ does not involve the small-scale and shadow fading components. These quantities, according to the system model, are not dependent on either the user or AP positions and thus do not contribute to the placement process. Consequently, for the subsequent SINR-based and SGINR maximization problems that we will discuss, we will average over the abovementioned quantities.

3.4 SGINR-Based AP Placement for Fully Flexible Networks

We have seen in the previous chapter that maximizing the average per-user rate that utilizes the SINR in (3.5) to determine the throughput-optimal AP locations is solved by using a Lloyd-type algorithm. For this purpose, the large-scale fading coefficients given in (3.2) for every user-AP pair (shown below for a user at $\mathbf{p}$ and its nearest AP indexed by $\mathcal{E}(\mathbf{p})$), is approximated as

$$\beta_{\mathcal{E}(\mathbf{p})} \approx \frac{c_1 z_{\mathcal{E}(\mathbf{p})}}{\left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}, \tag{3.11}$$

since r_0 is much smaller than the area dimensions considered. We also note here that for notational simplicity from Section 5.2, the second subscript has been dropped and we will continue this for the ensuing analyses.

Consequently, the SINR-based rate optimization problem for AP placement is

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\max}\ \mathbb{E}_{\mathcal{A},\underline{\mathbf{p}}}\left\{\log\left(1+\phi_{k_{\mathcal{E}(\mathbf{p})}}\right)\right\}, \tag{3.12}$$

where the notation $\underline{\mathbf{p}}$ denotes the set containing the served user at $\mathbf{p}$ which is in cell $\mathcal{C}_{\mathcal{E}(\mathbf{p})}$ and the $M-1$ interfering users from cells $\mathcal{C}_{m'}$, denoted by $\mathbf{p}_{m'}$, where $m' \neq \mathcal{E}(\mathbf{p})$, and $\mathcal{A} = \{h_{\mathcal{E}(\mathbf{p})}, z_{\mathcal{E}(\mathbf{p})}, h_{m'}, z_{m'}\}$. For the user at $\mathbf{p}$, $h_{\mathcal{E}(\mathbf{p})}$ and $z_{\mathcal{E}(\mathbf{p})}$ are the small-scale fading coefficient and shadow fading, respectively, and $h_{m'}$ and $z_{m'}$ correspond to the same quantities, but for the interfering user at $\mathbf{p}_{m'}$. A solution to (3.12) was the Inter-AP Lloyd algorithm. It is worth noting that implementing power control along with AP placement, i.e., jointly optimizing uplink power with per-user power constraints and AP locations, would increment the rate performance. However, in this work, we want to solely investigate the effect of AP placement, with ICI that is prevalent in small-cell systems and that have been ignored by a majority of previous works. For this reason, we will continue to consider that all users transmit at the same power to the APs and leave the power optimization problem to future work.

Accounting for ICI is necessary in a small-cell scenario, and as an alternative to the SINR maximization in (3.12), we maximize the SGINR. In prior literature, the terms signal-to-leakage-plus-noise ratio (SLNR) and SGINR have been introduced. SLNR is considered in the downlink case [90] and quantifies the ICI by the leaked power by an AP to non-served users. SGINR is considered in the uplink [91] and quantifies the interference generated by the served user to all other APs. The concepts of both SLNR and SGINR are the same and are fundamentally different to that of SINR. As we model the uplink scenario in our work, we will use the term SGINR. It is worth noting that for SGINR, the knowledge of the selected users from the interfering cells is not necessary, unlike SINR. Additionally, it has been shown that SGINR has computational advantages over SINR in both downlink [92] and uplink [93] beamforming.

Here, we consider the instantaneous SGINR of user k_m at AP m as

$$\psi_{k_m} = \frac{\rho_r \beta_{mk_m} |h_{mk_m}|^2}{1+\rho_r \sum\limits_{\substack{m'=1 \\ m'\neq m}}^{M} \beta_{m'k_m} |h_{m'k_m}|^2}. \tag{3.14}$$

Note that while the numerator of the above expression represents the desired signal power, the denominator, unlike SINR, denotes the user-generated interference at other APs. Maximizing the SGINR provides a balance between maximizing user power versus generated ICI. Consequently, we have the following optimization problem

$$\underset{\mathbf{q}_m\in\mathcal{M}}{\arg\max}\ \mathbb{E}_{\mathcal{B},\mathbf{p}}\left\{\psi_{k_{\mathcal{E}(\mathbf{p})}}\right\}, \tag{3.15}$$

where $\mathcal{B} = \{h_{\mathcal{E}(\mathbf{p})}, z_{\mathcal{E}(\mathbf{p})}, h_{m'}, z_{m'}\}$ and $\mathbf{p}$ is the position of served user. Again, we use only a single subscript for simplicity, noting that while $h_{\mathcal{E}(\mathbf{p})}$ and $z_{\mathcal{E}(\mathbf{p})}$ are the small-scale fading coefficient and shadow fading, respectively, for the user at $\mathbf{p}$ to the serving cell, $h_{m'}$ and $z_{m'}$ correspond to the same quantities for the same user, but to the non-serving AP m'. Substituting the expression for SGINR in the objective function above and using Jensen's inequality, we have

$$\mathbb{E}_{\mathcal{B},\mathbf{p}}\left\{\psi_{k_{\mathcal{E}(\mathbf{p})}}\right\} \geq \mathbb{E}_{\mathcal{B}}\left\{\frac{\rho_r c_1 |h_{\mathcal{E}(\mathbf{p})}|^2 z_{\mathcal{E}(\mathbf{p})}}{D}\right\}, \tag{3.16}$$

where the denominator D (which we now have to minimize) of the above expression is

$$D = \mathbb{E}_{\mathbf{p}} \left\{ \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma} \left(1 + \rho_r \sum_{\substack{m'=1 \\ m' \neq \mathcal{E}(\mathbf{p})}}^{M} \frac{c_1 |h_{m'}|^2 z_{m'}}{||\mathbf{p} - \mathbf{q}_{m'}||^{\gamma}} \right) \right\}. \tag{3.17}$$

The first term $\left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}$ is the distance term of the pathloss and corresponds to the SNR while the other term is the ICI term. D can further be simplified using the independence between the two terms above to

$$D = \mathbb{E}_{\mathbf{p}} \left\{ \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma} \right\} \times \left(1 + \rho_r \sum_{\substack{m'=1 \\ m' \neq \mathcal{E}(\mathbf{p})}}^{M} c_1 |h_{m'}|^2 z_{m'} \mathbb{E}_{\mathbf{p}} \left\{ \frac{1}{||\mathbf{p} - \mathbf{q}_{m'}||^{\gamma}} \right\} \right). \tag{3.18}$$

We now simplify the inner term using the following approximation

$$\mathbb{E}_{\mathbf{p}} \left\{ \frac{1}{||\mathbf{p} - \mathbf{q}_{m'}||^{\gamma}} \right\} \approx \frac{1}{\left|\left|\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}, \tag{3.19}$$

whose justification is provided in Appendix 3.A, and results in

$$D \approx \mathbb{E}_{\mathbf{p}} \left\{ \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma} \right\} \left(1 + \rho_r \sum_{\substack{m'=1 \\ m' \neq \mathcal{E}(\mathbf{p})}}^{M} \frac{c_1 |h_{m'}|^2 z_{m'}}{\left|\left|\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}} \right). \tag{3.20}$$

We consider two terms in the above expression of D. To minimize D, the first term $\mathbb{E}_{\mathbf{p}}\{||\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}||^{\gamma}\}$ corresponding to the SNR, which mirrors the objective function of the Lloyd algorithm, is to be minimized. Further, the second ICI term contains $1/||\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}||^{\gamma}$, which are the inverses of the inter-AP distances $\forall m' \neq \mathcal{E}(\mathbf{p})$ and need also to be minimized. Since all the abovementioned quantities are positive and have all to be minimized, the distortion function, keeping in line with the VQ optimization in (3.6), can be written as

$$d_{\mathrm{IA}}\left(\mathbf{p}, \underline{\mathbf{q}}\right) = \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma} + \sum_{m' \neq \mathcal{E}(\mathbf{p})} \frac{\kappa_{m'}}{\left|\left|\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}, \tag{3.21}$$

where $\kappa_{m'} \geq 0$, $m' \neq \mathcal{E}(\mathbf{p})$ are the *trade-off factors*. For simplicity, we assume a common $\kappa \geq 0$, and we have

$$d_{\mathrm{IA}}\left(\mathbf{p}, \underline{\mathbf{q}}\right) = \left|\left|\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma} + \kappa \sum_{m' \neq \mathcal{E}(\mathbf{p})} \frac{1}{\left|\left|\mathbf{q}_{m'} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^{\gamma}}, \tag{3.22}$$

The trade-off factor κ is a network design parameter that can be selected to decide the relative importance of the ICI term to the SNR (desired signal) term. This is called the inter-AP distortion measure, denoted by d_{IA}, and coincides with the distortion measure derived from SINR in the previous chapter. This reinforces the fact that the inter-AP distortion function (and correspondingly, the Inter-AP Lloyd algorithm) is helpful in tackling ICI in AP placement. The steps of the Inter-AP Lloyd algorithm were described in the previous chapter for networks with fully flexible infrastructure and hence is omitted here.

3.5 AP Placement in Hybrid UAV-Terrestrial Networks: Problem Formulation and Solution

We now formulate the AP placement problem for hybrid networks. Consider the scenario where M_f T-APs (the fixed network) are optimally placed for the user configuration $f_{\mathbf{P}_1}(\mathbf{p}_1)$. Conceivably, after some time, the user configuration changes, say, to $f_{\mathbf{P}_2}(\mathbf{p}_2)$, and this results in the M_f T-APs being at sub-optimal positions. Thus, the added M_u UAV-APs to the system requires placement such that the performance gap between the scenario where all $M = M_f + M_u$ APs are optimally placed (the fully flexible network) and the scenario where only M_u out of M are movable (the hybrid network) is to be minimized. In other words, the hybrid network consisting of both T-APs and UAV-APs is compared to the ideal case of a fully flexible network that consists of UAV-APs alone. We aim for the performance of the hybrid network to be as close to that of the fully flexible network, which is the benchmark solution. We quantify the performances using the average SGINRs of both networks.

Let $\psi^{\text{flex}}_{k_{\mathcal{E}(\mathbf{p})}}$ be the SGINR achieved by optimal placement of all M APs in a fully flexible network for the new user distribution $f_{\mathbf{P}}(\mathbf{p})$ (subscript '2' removed for simplicity) and $\psi^{\text{hbd}}_{k_{\mathcal{E}(\mathbf{p})}}$ be the SGINR achieved by using the hybrid UAV-terrestrial system where only M_u out of M APs can be optimally placed. Clearly, the average SGINR of the hybrid network is lower than the average SGINR of the fully flexible network. Hence, the average SGINR of the hybrid network is subtracted from that of the fully flexible network in order to define the performance gap. Formally, the objective function to be minimized (the performance gap) can be written as follows

$$J = \mathbb{E}_{\mathcal{B},\mathbf{p}} \left\{ \psi^{\text{flex}}_{k_{\mathcal{E}(\mathbf{p})}} - \psi^{\text{hbd}}_{k_{\mathcal{E}(\mathbf{p})}} \right\}, \tag{3.23}$$

where the set $\mathcal{B}$ is as defined above for (3.15). Alternative to minimizing J, we maximize its negative $J' = -J$ such that

$$J' = \mathbb{E}_{\mathcal{B},\mathbf{p}} \left\{ \psi^{\text{hbd}}_{k_{\mathcal{E}(\mathbf{p})}} \right\} - \mathbb{E}_{\mathcal{B},\mathbf{p}} \left\{ \psi^{\text{flex}}_{k_{\mathcal{E}(\mathbf{p})}} \right\}. \tag{3.24}$$

It is assumed that the optimal AP locations for the fully flexible case is known. Note that the second term in (3.24) is thus independent of the optimal UAV-AP locations of the hybrid network and hence can be neglected. Consequently, maximizing the above lower bound of J' leads to the optimization problem

$$\underset{\mathbf{q}_m \in \mathcal{M}_u}{\arg\max}\, \mathbb{E}_{\mathcal{B},\mathbf{p}} \left\{ \psi^{\text{hbd}}_{k_{\mathcal{E}(\mathbf{p})}} \right\}, \tag{3.25}$$

which is performed only over the M_u UAV-APs.

To perform the optimization in (3.25), we derive a Lloyd-type algorithm which maximizes the average per-user SGINR over the UAV-APs, which we call the *Hybrid AP Placement Algorithm (HAPPA)*. Following the steps of the iterative Lloyd-type algorithm, we apply the nearest neighbor condition (NNC) to all M APs as

$$\mathcal{C}_m = \left\{ \mathbf{p} : d_{\text{IA}}\left(\mathbf{p}, \mathbf{q}_m\right) \leq d_{\text{IA}}\left(\mathbf{p}, \mathbf{q}_l\right), \forall l \neq m \right\}, \tag{3.26}$$

where $d_{\text{IA}}(\cdot)$ is the inter-AP distortion function from (3.22). The centroid condition (CC) step now updates

the M_u UAV-AP locations using the steepest descent method whose update equation is given as

$$\mathbf{q}_m^{(j+1)} = \mathbf{q}_m^{(j)} - \delta \frac{\partial}{\partial \mathbf{q}_m^{(j)}} \int\limits_{\mathcal{C}_m} d_{\mathrm{IA}}\left(\mathbf{p}, \mathbf{q}_m^{(j)}\right) f_{\mathbf{P}}(\mathbf{p}) \mathrm{d}\mathbf{p}, \tag{3.27}$$

for all $m \in \mathcal{M}_u$, where j is the iteration index, δ is the step size, and the gradient expression is given by

$$\frac{\partial}{\partial \mathbf{q}_m} \int\limits_{\mathcal{C}_m} d_{\mathrm{IA}}\left(\mathbf{p}, \mathbf{q}_m\right) f_{\mathbf{P}}(\mathbf{p}) \mathrm{d}\mathbf{p} = \frac{\gamma}{|\mathcal{C}_m|} \sum_{\mathbf{p}_n \in \mathcal{C}_m} \left(\mathbf{q}_m - \mathbf{p}_n\right) ||\mathbf{p} - \mathbf{q}_m||^{\gamma-2} + \kappa \sum_{m' \neq m} \frac{\mathbf{q}_{m'} - \mathbf{q}_m}{||\mathbf{q}_{m'} - \mathbf{q}_m||^{\gamma+2}}, \tag{3.28}$$

The T-APs, however, remain at their original locations as they are fixed in position. The NNC and CC steps are iterated until the overall distortion is less than a threshold or cell assignments no longer change. HAPPA is summarized in Algorithm 8. HAPPA incurs the same overall complexity as the Inter-AP Lloyd algorithm, i.e., $\mathcal{O}(KMI_SI_L)$, where I_S and I_L are the number of iterations taken for the steepest descent method and the Lloyd-type algorithm to converge, respectively. We do point out that inspite of the same complexity order, the number of operations involved in the CC step of HAPPA is less than that of the Inter-AP Lloyd algorithm and depends on the degree of hybridity, i.e., the proportion of UAV-APs to the total number of APs.

Algorithm 8 Hybrid AP Placement Algorithm (HAPPA)

Initialize:
$\mathbf{q}_m^{(0)}, \forall m \in \mathcal{M}$
$i \leftarrow 0$
while total distortion less than threshold **do**
  update $\mathcal{C}_m^{(i+1)}, \forall m \in \mathcal{M}$ using ▷ NNC

$$\mathcal{C}_m^{(i+1)} \leftarrow \left\{\mathbf{p}_k : d_{\mathrm{IA}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq d_{\mathrm{IA}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m\right\}$$

  if $m \in \mathcal{M}_f$ **then** ▷ CC for T-APs
    $\mathbf{q}_m^{(i+1)} \leftarrow \mathbf{q}_m^{(i)}$
  else if $m \in \mathcal{M}_u$ **then** ▷ CC for UAV-APs
    update $\mathbf{q}_m^{(i+1)}$ using

$$\mathbf{q}_m^{(j+1)} \leftarrow \mathbf{q}_m^{(j)} - \delta \left(\frac{\gamma}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \left(\mathbf{q}_m^{(j)} - \mathbf{p}_k\right) \left|\left|\mathbf{p}_k - \mathbf{q}_m^{(j)}\right|\right|^{\gamma-2} + \kappa \sum_{m' \neq m} \frac{\mathbf{q}_{m'}^{(j)} - \mathbf{q}_m^{(j)}}{\left|\left|\mathbf{q}_{m'}^{(j)} - \mathbf{q}_m^{(j)}\right|\right|^{\gamma+2}} \right)$$

    until convergence
    $\mathbf{q}_m^{(i+1)} \leftarrow \mathbf{q}_m^{(j+1)}$
  $i \leftarrow i + 1$
return $\mathbf{q}_m, \forall m \in \mathcal{M}$

3.6 AP Placement Accounting for Load Balancing

As mentioned in Section 4.1, the Lloyd and Lloyd-type algorithms described previously for throughput optimality result in unequal cell occupancies, where users of cells with a lower occupancy would unfairly get more opportunities to access the spectrum than users of cells with a higher occupancy. Hence, when user delay is measured by the frequency of user-to-AP access (the *spectral access fraction*), these algorithms result in significantly varied spectral access profiles. For delay sensitive applications, LB capabilities are required in order to achieve a certain application-dependent degree of similarity in spectral access profile for all users. This means that we should strive to equalize the occupancy of each cell, the degree of which is determined by the specific application. Below, we outline and motivate the OWLA and CELA-α AP placement algorithms, both generating Lloyd-type algorithms that improve system fairness, with CELA-α having a flexible throughput-delay trade-off adaptable to application requirements.

Algorithm 9 OWLA

1: Initialize random AP locations $\mathbf{q}_m^{(0)}$, $\forall m$.

2: Use the NNC to determine the cells $\mathcal{C}_m^{(i+1)}$, $\forall m$

$$\mathcal{C}_m^{(i+1)}=\left\{\mathbf{p}_k: w_m^{(i)} d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq w_l^{(i)} d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m\right\}.$$

3: Use the CC to determine the AP locations $\mathbf{q}_m^{(i+1)}$, $\forall m$

$$\mathbf{q}_m^{(i+1)} = \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \mathbf{p}_k.$$

4: Repeat from step 2 until convergence.

3.6.1 Occupancy Weighted Lloyd Algorithm (OWLA)

The access rate using per-user SNR $\psi_{k_{\mathcal{E}(\mathbf{p})}}$, which can be obtained by neglecting ICI from SINR $\phi_{k_{\mathcal{E}(\mathbf{p})}}$ in (3.5), is

$$R_{k_{\mathcal{E}(\mathbf{p})}}^{\mathrm{acc}} = \frac{1}{N_{\mathcal{E}(\mathbf{p})}} \mathbb{E}\left\{\log_2\left(1+\psi_{k_{\mathcal{E}(\mathbf{p})}}\right)\right\}, \tag{3.29}$$

where $\mathcal{E}(\mathbf{p})$ indexes the AP closest to user k and $N_{\mathcal{E}(\mathbf{p})}$ is the number of users in cell $\mathcal{C}_{\mathcal{E}(\mathbf{p})}$. Note that the achievable rate $R_{k_{\mathcal{E}(\mathbf{p})}} = \mathbb{E}\left\{\log_2\left(1+\phi_{k_{\mathcal{E}(\mathbf{p})}}\right)\right\}$ does not account for the delay incurred by a user as it waits to transmit to its AP with TDMA scheduling. Therefore, the rate is normalized using the resource sharing factor $1/N_{\mathcal{E}(\mathbf{p})}$. The logarithm in (3.29) can be well fitted with a third degree polynomial $\log_2(1+x) \approx a_1 + a_2 x + a_3 x^2 + a_4 x^3$, and therefore (3.29) can be rewritten as

$$R_{k_{\mathcal{E}(\mathbf{p})}}^{\mathrm{acc}} = \frac{1}{N_{\mathcal{E}(\mathbf{p})}} \mathbb{E}\left\{a_1 + a_2 \psi_{k_{\mathcal{E}(\mathbf{p})}} + a_3 \psi_{k_{\mathcal{E}(\mathbf{p})}}^2 + a_4 \psi_{k_{\mathcal{E}(\mathbf{p})}}^3\right\}. \tag{3.30}$$

Focusing first on the linear term $a_2 x$, we can write

$$\mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\frac{a_2}{N_{\mathcal{E}(\mathbf{p})}}\psi_{k_{\mathcal{E}(\mathbf{p})}}\right\} \geq \mathbb{E}_{\mathcal{A}}\left\{\frac{a_2\rho_r c_1 |h_{\mathcal{E}(\mathbf{p})}|^2 z_{\mathcal{E}(\mathbf{p})}}{\left(\mathbb{E}_{\mathbf{p}}\left\{N_{\mathcal{E}(\mathbf{p})}^{\frac{2}{\gamma}} \left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2\right\}\right)^{\frac{\gamma}{2}}}\right\}, \tag{3.31}$$

using Jensen's inequality, where $z_{\mathcal{E}(\mathbf{p})}$ is the shadow fading coefficient, c_1 is the pathloss constant, γ is the pathloss exponent, $\beta_{\mathcal{E}(\mathbf{p})}$ in $\psi_{k_{\mathcal{E}(\mathbf{p})}}$ approximated as in [80], and $\mathcal{A} = \{h_{\mathcal{E}(\mathbf{p})}, z_{\mathcal{E}(\mathbf{p})}\}$. Note that we wish to maximize (3.30) and correspondingly minimize the denominator in (3.31). Following the VQ framework, the distortion function for the linear term is $d\left(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right) = N_{\mathcal{E}(\mathbf{p})}^{2/\gamma}\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2$. Applying the same technique to all terms in (3.30), the new distortion function is defined by summing the distortion from each term

$$d\left(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right) = \left(N_{\mathcal{E}(\mathbf{p})}^{\frac{2}{\gamma}} + N_{\mathcal{E}(\mathbf{p})}^{\frac{2}{2\gamma}} + N_{\mathcal{E}(\mathbf{p})}^{\frac{2}{3\gamma}}\right)\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2, \tag{3.32}$$

where the squared Euclidean distance measure of the standard Lloyd algorithm has been multiplied with a weight $w_m = N_{\mathcal{E}(\mathbf{p})}^{2/\gamma} + N_{\mathcal{E}(\mathbf{p})}^{2/2\gamma} + N_{\mathcal{E}(\mathbf{p})}^{2/3\gamma}$. Also, note that curve fitting in (3.29) with a higher order polynomial is unnecessary since the growth of the weight w_m diminishes as the polynomial order grows. The Lloyd-type algorithm corresponding to this weighted distortion measure is termed OWLA and is outlined in Algorithm 9.

3.6.2 Cell Equalized Lloyd Algorithm-α (CELA-α)

Here, we start by explaining a simple way to achieve load balancing. The Lloyd algorithm used for throughput oriented AP placement can be modified to equalize the number of users in each cell, which is performed at every iteration of the algorithm. First, a hard criterion can be set which comprises of simply "moving" (re-assigning) the users from cells with excess users to cells that have a user deficiency. Upon completion (convergence), all cells would have the same number of users. This algorithm is outlined below.

Modification to the Lloyd Algorithm

The same iterative procedure as the Lloyd algorithm is to be followed, with the addition of an extra step (the moving step) in the iterative process. We call this step *User Re-distribution for Equalization (URE)* and is added between Step 1 (NNC) and Step 2 (CC) of the Lloyd algorithm. Letting N_m be the number of users in each cell $\mathcal{C}_m$ and given that K is the total number of users in the area under consideration, the target number of users in each cell (for equal number of users) is $N_m = K/M = N$, $\forall m = 1, 2, \ldots, M$. The URE procedure is outlined in Algorithm 10. It is clear from the algorithm that all cells will have N users after convergence.

Cell Equalized Lloyd Algorithm (CELA)

We observe that the Lloyd algorithm and its modified version above constitute two extreme cases. Specifically, in the Lloyd algorithm, the MSE alone is optimized while user access is not considered. On the

Algorithm 10 URE Procedure

1: Find all the cells that have number of users $> N$ and arrange them in descending order. Let the ordered set of cells generated be $\mathcal{C}^G$.

2: Iterate through the cells in $\mathcal{C}^G$ and perform the following process for each cell $\mathcal{C}_g \in \mathcal{C}^G$:

a: Arrange all users associated with $\mathcal{C}_g$ in ascending order of their *second lowest* distortion value: $\mathcal{U}^g$, and iterate through these users.

b: For each such user, if the cell corresponding to the second lowest distortion value has number of users $< N$, then assign user to that cell and move on to the next user. If not, simply move on to the next user.

c: Continue this process until all excess users in $\mathcal{C}_g$ are moved to other cells. If not, use the next (third, fourth, ...) lowest distortion value and repeat from Step 2a.

other hand, in the modified version, users are moved between cells in order for all users to have equal access to the AP by way of all cells having an equal number of users. The main drawback in the latter is that during the process of moving users to cells with fewer users, many cells would have users assigned to them which are very far away, reducing throughput. In other words, equal access comes at the cost of significant user throughput.

CELA is a compromise between the Lloyd algorithm and its modification. In fact, the only difference from the modified Lloyd algorithm is Step 2b of the URE procedure (Algorithm 10), which now involves a soft criterion for moving users, and is given below. It is important to note that in this case, all cells need not have exactly the same number of users N after convergence.

2b: For each user $u \in \mathcal{U}^g$, if the cell corresponding to the second lowest distortion value has number of users $< N$, look at the distance d_u^g between the center of that cell and the current user position. Let $\mathcal{C}_m^{\text{nearest}}$ denote the nearest cell to $\mathcal{C}_m$, $\forall m$.

- If d_u^g is greater than the distance between $\mathcal{C}_g$ and $\mathcal{C}_g^{\text{nearest}}$, then do not move the user, but assume that the user has been moved for counting purposes.
- Otherwise, assign the user to that cell.

Proceed to the next user. If the number of users $\geq N$, simply move on to the next user.

We note that the property of convergence and the complexity of CELA is the same as that of the Lloyd algorithm.

CELA-α

In CELA above, a hard criterion of equal user access was set, resulting in all cells ending up with the same occupancy, however, causing significant throughput loss due to users being re-assigned to far away cells. Further, a distance threshold to prevent such undesirable user re-assignments was introduced, however,

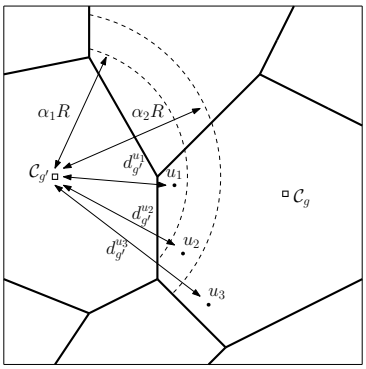

Figure 3.1: The effect of emphasis factor α on distance thresholds in CELA-α. In this case, R is the communications radius for cell $\mathcal{C}'_g$.

having two main drawbacks in its re-assignment procedure. Firstly, only the distortion values are considered to re-assign users. That is, although the cell occupancies are checked prior to user re-assignment, they are not used in the decision of the order in which users are re-assigned to other cells. Thus, there is a need to jointly consider both distortion and cell occupancies in this decision process. Secondly, depending on system requirements, a trade-off between delay and throughput might be necessary. Therefore, we next address the above needs, leading to a more comprehensive algorithm called CELA-α, where α is the trade-off factor allowing flexibility between throughput and delay.

Allowing throughput and delay trade-off. By pre-multiplying the abovementioned cell-specific distance thresholds, i.e., R_m^{th}, with the trade-off factor α, then the new threshold becomes αR_m^{th}. We know that if the distance threshold is set to 0, i.e., $R_m^{\text{th}} = 0$, then the algorithm then becomes the standard Lloyd algorithm. On the other hand, high distance thresholds would enable completely equal user access due to the equal occupancy in each cell, but it would result in reduced throughput owing to the large distances between select users and their APs. Here, α enables us to adjust the threshold between these two extremes. In Fig. 3.1, we illustrate the discussions about CELA-α. Two cells $\mathcal{C}_g$ and $\mathcal{C}_{g'}$ are shown along with three users u_1, u_2, and u_3 in cell $\mathcal{C}_g$. Assume that the three users in $\mathcal{C}_g$ with excess of users are to be moved to $\mathcal{C}_{g'}$ with low occupancy. For simplicity, $R_{g'}^{\text{th}} = R$, the communication radius of the AP in $\mathcal{C}_{g'}$ and the distance between the three users and $\mathcal{C}_{g'}$ are $d_{g'}^{u_1}$, $d_{g'}^{u_2}$, and $d_{g'}^{u_3}$, respectively; α_1 and α_2 represent two trade-off factors. Under α_1, since $d_{g'}^{u_1} < \alpha_1 R < d_{g'}^{u_2} < d_{g'}^{u_3}$, user u_1 will be moved to $\mathcal{C}_{g'}$ while users u_2 and u_3 will remain in $\mathcal{C}_g$. On the other hand, under α_2, we have $d_{g'}^{u_1} < d_{g'}^{u_2} < \alpha_2 R < d_{g'}^{u_3}$, which implies that users u_1 and u_2 will be moved to $\mathcal{C}_{g'}$ while user u_3 will remain in $\mathcal{C}_g$. Hence, it is evident that if the value of the trade-off factor α is increased, more users are re-assigned to $\mathcal{C}_{g'}$ and hence lesser spectrum access delay is obtained in $\mathcal{C}_g$ at the expense of some overall throughput loss.

Addition of cell occupancy to re-assignment. As mentioned above, considering the joint influence of distance and cell occupancy in order to determine the priority with which to re-assign users to other cells would involve updating the algorithm as follows. For each cell with excess users, the distance between the users in the cell and all other APs is multiplied with the occupancy of the corresponding cell and the user with the lowest such value is considered first.

Algorithm 11 CELA-α

1: Initialize random AP locations $\mathbf{q}_m^{(0)}$, $\forall m$.

2: Use the NNC to determine the cells $\mathcal{C}_m^{(i+1)}$, $\forall m$

$$\mathcal{C}_m^{(i+1)} = \left\{ \mathbf{p}_k : d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m \right\}.$$

3: Perform the re-assignment procedure:

3.1: Find all the cells that have number of users $> N$ and arrange them in descending order. Let the ordered set of cells generated be $\mathcal{C}^G$.

3.2: Iterate through the cells in $\mathcal{C}^G$ and perform the following process for each cell $\mathcal{C}_g \in \mathcal{C}^G$:

a: For each user u_g associated with $\mathcal{C}_g$, generate a vector $\mathbf{v}_{u_g}$ containing distances to *all other* APs. Multiply each element of $\mathbf{v}_{u_g}$ with the occupancy of the corresponding cell. Arrange these composite values (product of user-AP distance and cell occupancy) in ascending order within the vector to generate $\overline{\mathbf{v}}_{u_g}$.

b: Take the first element of all vectors $\overline{\mathbf{v}}_{u_g}$, $\forall u_g \in \mathcal{C}_g$ and sort them in a new vector $\mathbf{y}_g$ in ascending order.

c: Iterate through the elements of $\mathbf{y}_g$ and for both corresponding user u_g and cell $\mathcal{C}_{g'}$, the two following conditions have to be met to allow user u_g to be assigned to cell $\mathcal{C}_{g'}$:

- Occupancy of cell $\mathcal{C}_{g'}$, $N_{g'} < N$
- User-AP distance for cell $\mathcal{C}_{g'}$, $d(\mathbf{p}_{u_g}, \mathbf{q}_{g'}) < \alpha R_{g'}^{\mathrm{th}}$

If either condition is not satisfied, u_g is not re-assigned to $\mathcal{C}_{g'}$ and remains in $\mathcal{C}_g$.

d: Once all elements of $\mathbf{y}_g$ are considered, use the next (second, third, ...) element of every vector $\overline{\mathbf{v}}_{u_g}$ of users who have not been re-assigned and repeat from step 3.2-b.

3.3: Repeat from step 3.2-a for the next cell in $\mathcal{C}^G$ until all cells have been considered.

4: Use the CC to determine the AP locations $\mathbf{q}_m^{(i+1)}$, $\forall m$

$$\mathbf{q}_m^{(i+1)} = \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \mathbf{p}_k.$$

5: Repeat from step 2 until convergence.

The complete CELA-α detailing the above two modifications is provided in Algorithm 11. Note that the vector $\overline{\mathbf{v}}_{u_g}$ represents the ordered set of cells which should be followed when re-assigning the user u_g. On the other hand, vector $\mathbf{y}_g$ provides the order in which each user in cell $\mathcal{C}_g$ has to be re-assigned to its respective cell $\mathcal{C}_{g'}$.

3.7 Simulation Methodology and Results

3.7.1 Parameters

We start by defining the parameters for the simulations used for the hybrid UAV-terrestrial small-cell network. The simulation parameters used are listed in Table 3.1. We consider a geographical area of 2 km $\times$ 2 km in which $K = 2000$ users and $M_f = 8$ T-APs are used, unless otherwise stated. For our analysis, by adding $M_u = 2$, 4, 6, and 8 UAV-APs, we vary the total number of APs up to $M = 16$. For the user distribution, we consider the GMM of the form

$$f_{\mathbf{P}}(\mathbf{p}) = \sum_{l=1}^{L} p_l \mathcal{N}\left(\mathbf{p}|\boldsymbol{\mu}_l, \sigma_l^2 \mathbf{I}\right), \tag{3.33}$$

where $\mathbf{I}$ is the identity matrix and L is the number of mixture components, called *groups* henceforth. For group l, p_l is the mixture component weight, $\boldsymbol{\mu}_l$ is the mean, and σ_l is the variance. In line with our problem formulation, we consider two user configurations, namely, GMM-1 $f_{\mathbf{P}_1}(\mathbf{p}_1)$ and GMM-2 $f_{\mathbf{P}_2}(\mathbf{p}_2)$. For simplicity, we assume that the new user configuration GMM-2 has the same parameters as GMM-1, except that $\boldsymbol{\mu}_1 = [0.52, -0.52]^T$ and $\sigma_1 = 200$, i.e., GMM-2 causes users of group 1 to be more dispersed than in GMM-1. We also model a compact version of GMM-2, called GMM-3. GMM-3 has the same parameters as GMM-2, but with means $\boldsymbol{\mu}_1 = [0.27, -0.27]^T$, $\boldsymbol{\mu}_2 = [0, 0.25]^T$, and $\boldsymbol{\mu}_3 = [-0.25, 0]^T$. Note that the distinct parameters of GMM-2 and GMM-3 aim at modeling sparse and dense AP deployments, respectively. The pathloss model defined in (3.2) is used with the Cost 231 Hata propagation model described in [61, eq. (4.36), eq. (4.37)] and shadow fading is ignored as it is averaged out in (3.25) during the placement process.

For the simulations concerning load balancing, the setup consists of $M = 8$ APs and $K = 2000$ users over a 2×2 km^2 area. The user GMM has parameters $L = 3$, $\boldsymbol{\mu}_l = [-0.17, 0.17]^T$, $\boldsymbol{\mu}_2 = [0.17, 0.17]^T$, $\boldsymbol{\mu}_3 = [0.17, -0.17]^T$, $\sigma_1 = \sigma_2 = \sigma_3 = 100$, $p_1 = 0.6$, and $p_2 = p_3 = 0.2$. The remaining parameters remain the same as above. It is to be noted that although various real-life scenarios dictate different types of user distributions, the AP placement algorithms described in this paper can be applied to determine the AP locations for all distribution types.

3.7.2 Performance Measures

For the hybrid networks, we use the following quantitative measures to evaluate performances.

1. *Per-user achievable rate:* We use the SINR quantity ϕ_{k_m} in (3.5) to calculate the achievable per-user rate of user k_m as

$$R_{k_m} = \mathbb{E}\left\{\log_2\left(1 + \phi_{k_m}\right)\right\}. \tag{3.34}$$

Further, [61, Ch. 4] provides an alternative calculation

$$R_{k_m} = \frac{1}{\ln 2} e^{\mu_k} \mathrm{Ei}\left(\mu_k\right), \tag{3.35}$$

Table 3.1: List of Simulation Parameters

Parameter	Value
Number of T-APs M_f	8
Number of UAV-APs M_u	$\{2, 4, 6, 8\}$
Number of users K	2000
GMM-1 $\{L, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$	$\{3, [0.5, -0.5]^T, [0, 0.5]^T, [-0.5, 0]^T\}$
$p_1, p_2, p_3, \sigma_1, \sigma_2, \sigma_3\}$	$\{0.6, 0.6, 0.2, 100, 100, 100\}$
GMM-2 $\{L, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$	$\{3, [0.52, -0.52]^T, [0, 0.5]^T, [-0.5, 0]^T\}$
$p_1, p_2, p_3, \sigma_1, \sigma_2, \sigma_3\}$	$\{0.6, 0.6, 0.2, 200, 100, 100\}$
GMM-3 $\{L, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$	$\{3, [0.27, -0.27]^T, [0, 0.25]^T, [-0.25, 0]^T\}$
$p_1, p_2, p_3, \sigma_1, \sigma_2, \sigma_3\}$	$\{0.6, 0.6, 0.2, 200, 100, 100\}$
Pathloss $\{\gamma, c_0, c_1, r_0\}$	$\{2, 75.86, 7.59 \times 10^{-7}, 0.001$ km$\}$
Power ρ_r	200 mW
Bandwidth	20 MHz

where

$$\mu_k = \frac{1 + \rho_r \sum\limits_{\substack{m'=1 \\ m' \neq m}}^{M} \beta_{mk_{m'}}}{\rho_r \beta_{mk_m}}, \tag{3.36}$$

and $\mathrm{Ei}\,(x) = \int\limits_x^\infty \frac{e^{-t}}{t} \mathrm{d}t$ is the exponential integral.

2. *Sum rate:* The sum of all M per-user achievable rates, where each user is selected from a cell, is given by

$$S = \sum_{m=1}^{M} R_{k_m}. \tag{3.37}$$

The maximum number of iterations for the Lloyd and the proposed algorithms is set to 50. We use Monte Carlo simulations to generate achievable and sum rates, considering $10,000$ iterations, where we choose a set of users for transmission each time. For visualization, we use cumulative distribution function (CDF) plots, normalizing the rate values by its maximum quantity for relative comparison of the algorithms. Further, in order to quantify the performance, we utilize the 95%-likely metric that represents the best rate of the worst 5% of the users (users closer to cell borders). We denote this by $\mathcal{P}^{5\%}$, where $\mathcal{P}$ is R_{k_m} or S. The improvement ratio (IR) then quantifies the relative performance, expressed as percentage

$$\mathrm{IR} = \frac{\mathcal{P}^{5\%,\text{Placement B}} - \mathcal{P}^{5\%,\text{Placement A}}}{\mathcal{P}^{5\%,\text{Placement A}}} \times 100. \tag{3.38}$$

For load balancing, the following per-user measures are used.

1. *Access rate:* $R_{k_m}^{\mathrm{acc}}$, defined in (3.29), using ϕ_{k_m} from (3.5).

2. *Spectral access fraction:* This is a measure of the frequency with which user k_m communicates with its serving AP m. Formally, this is

$$U_{k_m} = \frac{1}{N_m}. \tag{3.39}$$

For relative comparison, the IR defined above is used here. We also note that the performance of CELA is not outlined here since it is an inferior algorithm to CELA-α.

3.7.3 Initialization of the Algorithms

It is well known that random initializations result in convergence of the Lloyd or the proposed Lloyd-type algorithms presented in this paper to different local minima. Many researchers try to circumvent this problem by running multiple instances of the entire algorithm and then averaging the results. To have a common starting point and for unbiased comparison of all algorithms, we develop, specifically for the GMM user distribution, a novel initialization scheme based on the bit allocation problem in VQ. The number of APs allocated to each group l of the GMM user distribution is

$$u_l = \overline{u} + \log_2 \frac{h_l}{H} + \log_2 \frac{g_l}{G}, \tag{3.40}$$

where

$$\overline{u} = \frac{M}{L}, \qquad h_l = 4\sqrt{|\mathbf{\Sigma}_l|}, \qquad g_l = K_l, \qquad H = \left(\prod_{l=1}^{L} h_l\right)^{\frac{1}{L}}, \qquad G = \left(\prod_{l=1}^{L} g_l\right)^{\frac{1}{L}}. \tag{3.41}$$

Here, L is the total number of groups, K_l is the number of users in group l, and $\mathbf{\Sigma}_l$ is the (sample) covariance of the users in group l. After allocating the M APs to the L groups, they are initialized randomly within each group. The proof of (3.40) in presented in Appendix 3.B.

It is to be noted that although the above initialization scheme is designed keeping the GMM in mind, it can be applied to any general distribution as the GMM presents a convenient way to express other distributions. As such, non-GMM distributions can be fit to GMMs either by employing the expectation-maximization (EM) algorithm [94] or through numerical methods which uses goodness-of-fit metrics such as the negative log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

3.7.4 Numerical Results

Experiment 1. To simulate the Inter-AP Lloyd algorithm, we first generate the initial AP locations with the novel initialization procedure discussed above. For $M = 16$, the APs are allocated among the $L = 3$ groups as $u_1 = 6$ and $u_2 = u_3 = 5$. It can be ascertained from (3.40) that the mixture component weights are primarily responsible for this allocation since the group variances $\mathbf{\Sigma}_l$ are all equal. To showcase the Inter-AP Lloyd algorithm, we use two distinct trade-off factors $\kappa_1 = 0.2 \times 10^8$ and $\kappa_2 = 1 \times 10^8$. Fig. 3.2 shows a user realization as well as the final AP locations obtained using the Lloyd and Inter-AP Lloyd algorithms. Note that we only show the final placement for $\kappa_2 = 1 \times 10^8$ for simplicity. The arrangement of the AP locations of the Inter-AP Lloyd algorithm are spread apart more than of the Lloyd algorithm and increases with higher κ. This is primarily due to the influence of the inter-AP distances term in the distortion function

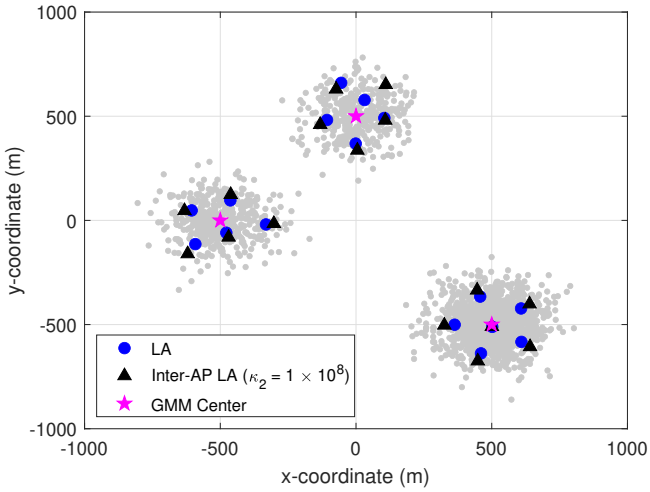

Figure 3.2: AP locations after convergence of the Lloyd and Inter-AP Lloyd algorithms for $\kappa_2 = 1 \times 10^8$ with $M = 16$ under GMM-1.

of the former, resulting in the APs being pushed away so that ICI is reduced, especially for those users that are near the cell boundaries. Now, in Fig. 3.3 we plot the CDF curves of the per-user achievable rate for the two values of trade-off factor $\kappa_1 = 0.2 \times 10^8$ and $\kappa_2 = 1 \times 10^8$. For comparison, we focus on the rate value of the worst 5% of the users, the 95%-likely rate, which mainly corresponds to those users near the cell boundaries, affected mostly by ICI. It is clear from the figure that the rates of these users are improved using the proposed Inter-AP Lloyd algorithm. On the other hand, users that are closer to the GMM centers, experience a reduction in desirable signal strength as the APs are pushed away, thus suffering a slight loss in capacity. This effect could be intuitively understood from examining desirable-signal/interference-signal trade-off induced by κ from (3.22), with the degree of trade-off depending on κ. Using the rate quantities measured by the value at which the 95%-likely rate line intersects the CDF curves, we find that κ_1 results in a 16.07% and κ_2 in a 42.75% improvement, establishing that an increase in the value of κ results in an increase in the 95%-likely rate. In Fig. 3.4, we examine the minimum rates guaranteed by the network when using both algorithms. Minimum rate is defined as the smallest among all the rates achieved by the served users during a transmission time interval. It is clearly shown that the minimum rate of the network that employs the Inter-AP Lloyd algorithm is always superior to that when the Lloyd algorithm is used. Similar to the previous figure, the performance gap increases as κ is assigned higher values. Through multiple experiments, we have observed that a low κ value $< 0.1 \times 10^8$ results in a performance similar to the Lloyd algorithm, while a value $> 2 \times 10^8$ results in convergence issues. Hence, the choice of κ is an important criterion in using the Inter-AP Lloyd algorithm. We can think of κ as a network design parameter, which allows for controlling the emphasis placed on the signal and the interference powers.

Experiment 2. Now, for the proposed HAPPA, we evaluate the effects of AP placement on the throughput performance of the hybrid network as we add additional UAV-APs under the GMM-2 configuration. For all cases, we set step-size $\delta = 0.5$ and trade-off factor $\kappa = 1 \times 10^8$. As a benchmark result, we calculate the performance of the system which has $M_f = 8$ T-APs and no UAV-APs, the fixed network, which

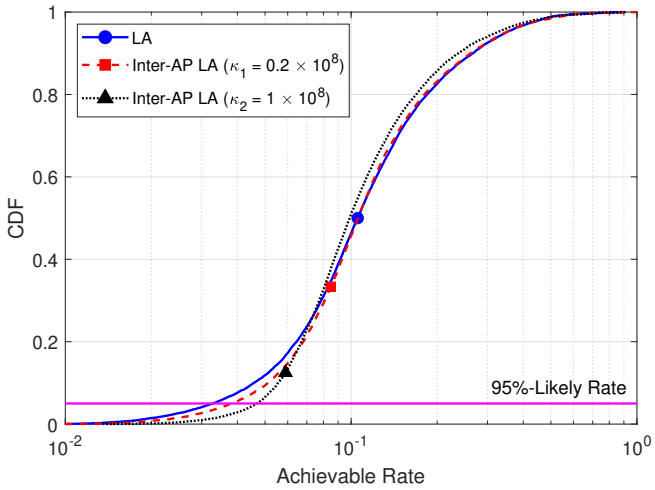

Figure 3.3: CDF plots of normalized per-user achievable rate for the Lloyd and Inter-AP Lloyd algorithms for $\kappa_1 = 0.2 \times 10^8$ and $\kappa_2 = 1 \times 10^8$ with $M = 16$ under GMM-1.

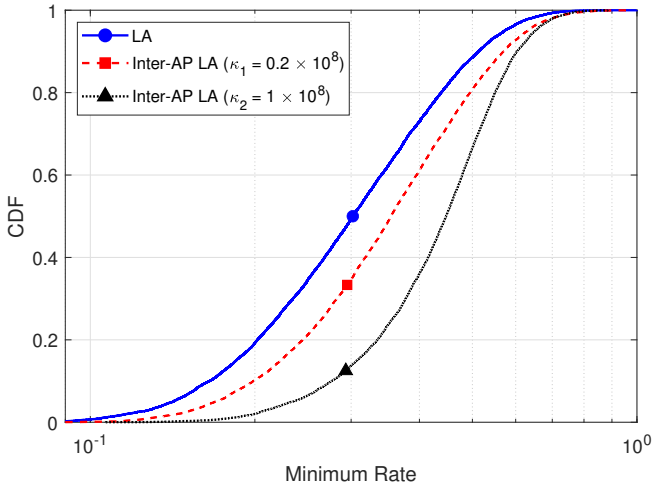

Figure 3.4: CDF plots of normalized minimum rate for the Lloyd and Inter-AP Lloyd algorithms for $\kappa_1 = 0.2 \times 10^8$ and $\kappa_2 = 1 \times 10^8$ with $M = 16$ under GMM-1.

is labelled as $\langle M_f = 8, M_u = 0 \rangle$. Then, we add 2, 4, 6, and 8 UAV-APs to the fixed network, resulting in up to $M = 16$ APs, the hybrid networks, labelled as $\langle M_f = 8, M_u = 2 \rangle$, $\langle M_f = 8, M_u = 4 \rangle$, $\langle M_f = 8, M_u = 6 \rangle$, and $\langle M_f = 8, M_u = 8 \rangle$, respectively. Finally, we also evaluate the performances of the system in the above networks when all M APs are UAV-APs, the fully flexible networks, labelled as $\langle M_f = 0, M_u = 10 \rangle$, $\langle M_f = 0, M_u = 12 \rangle$, $\langle M_f = 0, M_u = 14 \rangle$, and $\langle M_f = 0, M_u = 16 \rangle$, respectively. Note that the added UAV-APs are initialized using the proposed initialization algorithm. We start by showing in Fig. 3.5 the AP locations of fixed network $\langle M_f = 8, M_u = 0 \rangle$, the final UAV-AP locations obtained from HAPPA in the hybrid network $\langle M_f = 8, M_u = 8 \rangle$, and the final AP locations in the corresponding fully flexible network $\langle M_f = 0, M_u = 16 \rangle$ (where the Inter-AP Lloyd algorithm is run). Knowing that the T-APs were optimally placed for GMM-1, the UAV-APs position themselves according to the current GMM-2 configuration, in order to retrieve the capacity lost due to the

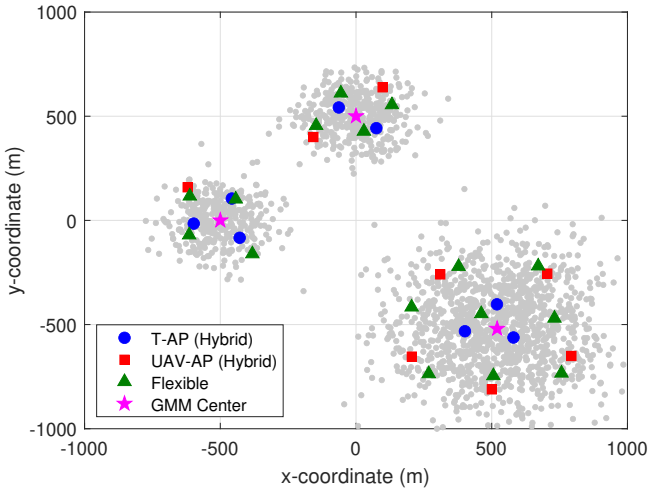

Figure 3.5: AP locations obtained after HAPPA with $M_f = 8$ T-APs and $M_u = 8$ UAV-APs (hybrid network) and Inter-AP Lloyd algorithm with $M = 16$ APs (fully flexible network) under GMM-2.

new user density. Clearly, in the fully flexible network $\langle M_f = 0, M_u = 16\rangle$, all APs are positioned optimally for GMM-2. For performance comparison, we calculate the sum rate in each network and the results are plotted in Fig. 3.6. The sum rate curve for the fixed network $\langle M_f = 8, M_u = 0\rangle$ clearly performs the worst. As we add UAV-APs to the system, the rate curves move to the right showing an increase in system sum rate. Thus, we demonstrate that, in this system, the HAPPA operating on a hybrid network efficiently takes care of the ICI even when the *degree of hybridity* $M_u/M \leq 50\%$. This implies that even when there are more T-APs than UAV-APs, i.e., less flexibility to place APs, significant improvements in system capacity are obtained. The curve for the fully flexible network $\langle M_f = 0, M_u = 16\rangle$ is also shown and only slightly outperforms the hybrid network $\langle M_f = 8, M_u = 8\rangle$. In terms of the 95%-likely rate, we measure the percentage improvement of each hybrid network relative to the fixed network and the values are tabulated in column 3 of Table 3.2. Furthermore, we compare the performance of each hybrid network to its corresponding fully flexible network for the various degrees of hybridity (e.g., comparison of $\langle M_f = 8, M_u = 2\rangle$ with $\langle M_f = 0, M_u = 10\rangle$), as tabulated in column 4 of Table 3.2. The performance gap between the hybrid and fully flexible networks is only 6.65% when the degree of hybridity is 20% and culminates in a minimal 2.02% gap when the hybridity is 50%. Note that in order to determine the minimum and sufficient number of UAV-APs required for the system to achieve a threshold sum rate η_{th}, we perform

$$U_{\min} = \arg\min_{M_u} S^{5\%}(M_f, M_u) \text{ s.t. } S^{5\%}(M_f, M_u) \geq \eta_{\text{th}}, \tag{3.42}$$

where $S^{5\%}(M_f, M_u)$ is the 95%-likely sum rate when M_u UAV-APs are added to M_f T-APs. The minimum number of UAV-APs can be obtained by sequential or bisectional search.

Experiment 3. In this experiment, we conduct a similar sum rate performance analysis of HAPPA as in experiment 2, however, under the compact GMM-3 user configuration. Fig. 3.7 shows the AP locations of the fixed network $\langle M_f = 8, M_u = 0\rangle$ and the resulting UAV-AP locations for the hybrid network $\langle M_f = 8, M_u = 8\rangle$ and the fully flexible network $\langle M_f = 0, M_u = 16\rangle$. For the hybrid network, the UAV-APs position

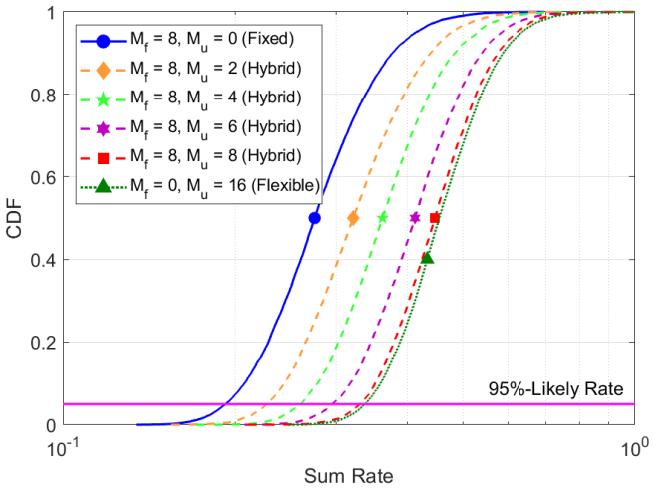

Figure 3.6: CDF plots of normalized sum rate for fixed $\langle M_f = 8, M_u = 0\rangle$, hybrid $\langle M_f = 8, M_u = 2, 4, 6, 8\rangle$, and fully flexible $\langle M_f = 0, M_u = 16\rangle$ networks under GMM-2.

Table 3.2: Hybrid Network Configuration, Degree of Hybridity, and 95%-Likely Sum Rate Improvement Ratios under GMM-2

Hybrid Network	$\frac{M_u}{M}$ [%]	$\frac{\text{Hybrid}}{\text{Fixed}}$ [%]	$\frac{\text{Flexible}}{\text{Hybrid}}$ [%]
$\langle M_f = 8, M_u = 2\rangle$	20	18.10	6.65
$\langle M_f = 8, M_u = 4\rangle$	33.33	35.61	4.84
$\langle M_f = 8, M_u = 6\rangle$	42.86	53.67	4.13
$\langle M_f = 8, M_u = 8\rangle$	50	71.92	2.02

themselves around the existing T-APs, which are suboptimally placed, to increase system capacity and the fully flexible network has the UAV-APs positioned optimally for GMM-3. In this scenario, we compare the sum rates achieved by the fixed network, the fully flexible network, and the hybrid networks with varying degrees of hybridity $\langle M_f = 8, M_u = 2\rangle$, $\langle M_f = 8, M_u = 4\rangle$, $\langle M_f = 8, M_u = 6\rangle$, and $\langle M_f = 8, M_u = 8\rangle$. For the sake of visual clarity, in Fig. 3.8, we focus on the comparison between the sum rate curves of the 50% hybrid network $\langle M_f = 8, M_u = 8\rangle$ and of the fully flexible network $\langle M_f = 0, M_u = 16\rangle$, since in the same figure, we also plot the corresponding sum rate curves for the 50% hybrid and fully flexible networks we obtained for the GMM-2, in experiment 2. The comparison of the 95%-likely rate values for all the considered degrees of hybridity are tabulated in Table 3.3. As expected, the fixed network performs the worst and performance substantially improves as the degree of hybridity increases. Notice that the fully flexible network only performs slightly better than the hybrid network even for a small degree of hybridity of 20%, which results in a small gap of nearly 8%. The gap shrinks to 5.65% for a 50% hybrid network. In Fig. 3.8, it is worth mentioning that in comparison to the rates yielded by GMM-2, the ones of GMM-3 are lower due to the increased ICI arising from the cells being relatively closer to one another. Quantitatively, relative to the corresponding curves yielded by the GMM-2, there is a 11.57% and 9.14% reduction in the 95%-likely

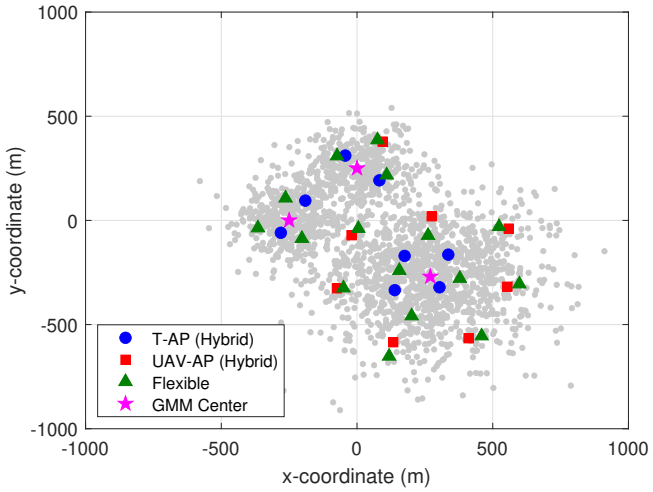

Figure 3.7: AP locations obtained after HAPPA with $M_f = 8$ T-APs and $M_u = 8$ UAV-APs (hybrid network) and Inter-AP Lloyd algorithm with $M = 16$ APs (fully flexible network) under GMM-3.

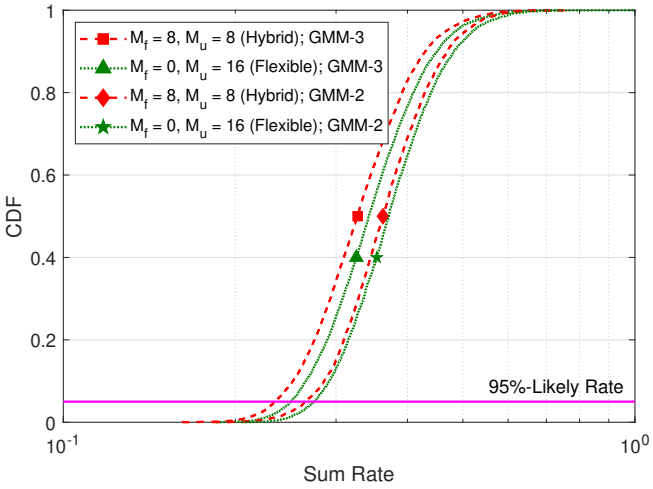

Figure 3.8: CDF plots of normalized sum rate for hybrid $\langle M_f = 8, M_u = 8 \rangle$ and fully flexible $\langle M_f = 0, M_u = 16 \rangle$ networks under GMM-3 and GMM-2.

values for the 50% hybrid and for the fully flexible networks, respectively, due to the GMM-3. Note that the effects of increased ICI are not only observed when comparing distinct GMMs but also within one GMM. This is observed in column 4 of Table 3.3, where the percentage improvements of a fully flexible network is determined over its hybrid counterpart for GMM-3. In summary, the proposed HAPPA is demonstrated to significantly improve sum rates in both cases when the GMM groups are far apart (lower overall ICI) and very close together (higher overall ICI). Thus, we can reasonably conclude that for user configurations in between the two extremes shown above, HAPPA will still be able to place APs optimally in order to recover the capacity lost as a result of changes in the user configuration, in the presence of ICI.

Experiment 4. In this experiment, we investigate the effects that the number of users in GMM-3 has on the performance of HAPPA. In this regard, we plot in Fig. 3.9, the sum rate curves achieved by

Table 3.3: Hybrid Network Configuration, Degree of Hybridity, and 95%-Likely Sum Rate Improvement Ratios under GMM-3

Hybrid Network	$\frac{M_u}{M}$ [%]	$\frac{\text{Hybrid}}{\text{Fixed}}$ [%]	$\frac{\text{Flexible}}{\text{Hybrid}}$ [%]
$\langle M_f = 8, M_u = 2 \rangle$	20	23.42	7.99
$\langle M_f = 8, M_u = 4 \rangle$	33.33	44.28	6.57
$\langle M_f = 8, M_u = 6 \rangle$	42.86	70.44	6.20
$\langle M_f = 8, M_u = 8 \rangle$	50	93.63	5.65

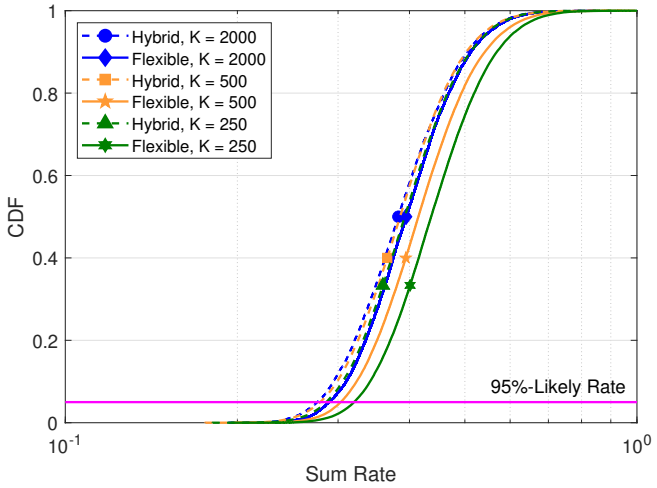

Figure 3.9: CDF plots of normalized sum rate for hybrid $\langle M_f = 8, M_u = 8 \rangle$ and fully flexible $\langle M_f = 0, M_u = 16 \rangle$ networks for $K = 2000$, 500, and 250 under GMM-3.

a $\langle M_f = 8, M_u = 8 \rangle$ hybrid network and the corresponding $\langle M_f = 0, M_u = 16 \rangle$ fully flexible network when $K = 2000$, $K = 500$, and $K = 250$. Two observations can be made from this figure. First, the sum rate performances of both the hybrid and fully flexible networks increase as the number of users is reduced. Second, the performance gap between hybrid and the fully flexible networks increases as the number of users is decreased. This follows from the fact that as the total number of users K reduces, the proportion of users that are far from the serving AP in each cell also reduces, thereby shifting the overall rate curve to the right. This also increases the importance of flexibility, therefore favoring the deployment that has more UAV-APs. Quantitatively, we calculate the performance gaps in terms of the 95%-likely rate. Among the hybrid networks, we observe a 1.20% and 3.74% gap for the curves when $K = 500$ and $K = 250$, respectively, over the curve when $K = 2000$. The values for the fully flexible networks are 4.82% and 10.58%, respectively. Further, the performance increase of the fully flexible networks over their hybrid counterparts starts at 5.65% when $K = 2000$, increases to 8.65% when $K = 500$, and to a higher 11.81% when $K = 250$.

Experiment 5. In this experiment, we compare the proposed HAPPA with the conventional exhaustive search. In particular, we perform the comparison of HAPPA with an exhaustive search algorithm to find the UAV-AP positions for the 50% hybrid network $\langle M_f = 8, M_u = 8 \rangle$ with GMM-2. The exhaustive

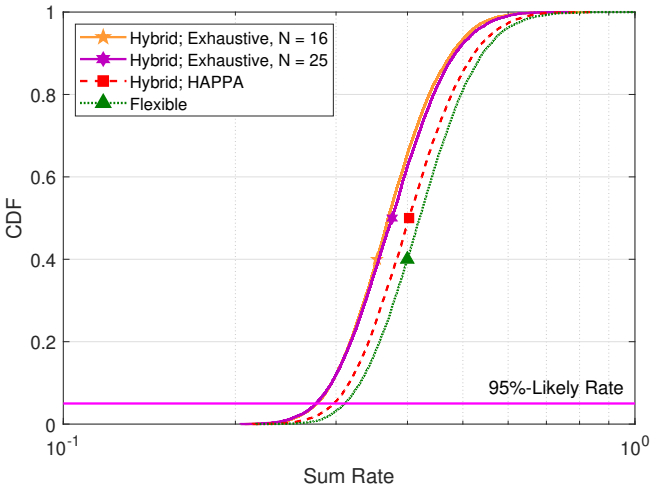

Figure 3.10: CDF plots of normalized sum rate for hybrid network $\langle M_f = 8, M_u = 8\rangle$ from exhaustive search, hybrid network $\langle M_f = 8, M_u = 8\rangle$ from HAPPA, and fully flexible network $\langle M_f = 0, M_u = 16\rangle$ under GMM-2.

search process is described as follows. We first split the geographical area under consideration into N grid points. Given the 8 T-APs, we have to determine the positions of the 8 UAV-APs at the grid points that generates the best 95%-likely sum rate performance. Since it is not beneficial for two APs to be colocated at the same grid point, the total number of combinations to distinctly place the 8 UAV-APs among N grid points is the binomial coefficient ${}^{N}C_8$. Note that N is chosen such that both HAPPA and the exhaustive search have similar implementation complexities for a fair performance comparison. The overall complexity order of HAPPA is $\mathcal{O}(KMI_SI_L)$ where $I_S = 5$ and $I_L = 15$ are average values obtained from conducting a large number of trials. As such, the complexity order for $K = 2000$ users and total $M = 16$ APs is 2.4×10^6. On the other hand, the exhaustive search algorithm has an overall complexity order of $KM \times {}^{N}C_8$, since for each combination, it is necessary to associate each of the K users to the M APs. Thus, $N = 16$ APs yields a complexity order of 4.12×10^6 for the exhaustive search, which is similar to that of HAPPA. With similar complexities ensuring a fair performance comparison, we then conduct the exhaustive search and generate the sum rate curve for the best combination, which is shown in Fig. 3.10 below. Clearly, it is observed that the performance of the exhaustive search with $N = 16$ grid points is inferior (the 95%-likely value is 4.22% lower) to HAPPA. To further investigate, we increased the number of grid points to $N = 25$, resulting in a complexity order of $KM \times {}^{25}C_8 = 3.46 \times 10^{10}$, which is clearly several orders of magnitude higher than that of HAPPA. The sum rate performance corresponding to $N = 25$ shown in Fig. 3.10 is also significantly lower (the 95%-likely value is 3.88% lower) than that of HAPPA and is only slightly better than that of $N = 16$. This shows that in spite of similar or even with significantly lower complexity, our proposed Lloyd-type algorithm, i.e., HAPPA, results in a placement of UAV-APs that yields a better sum rate performance in comparison to the exhaustive search method.

Experiment 6. In this experiment, we compare the proposed initialization scheme with the popular k-means++ initialization [95]. For this comparison, we consider a user distribution GMM-4, which has parameters $L = 2$, $\boldsymbol{\mu}_1 = [-0.5, 0]^T$, $\boldsymbol{\mu}_2 = [0.5, 0]^T$, $\sigma_1 = \sigma_2 = 100$, and $p_1 = p_2 = 0.5$, and is shown in Fig.

3.11. In order to show how our proposed initialization method described in Section 3.7.3 is advantageous over the k-means++ initialization scheme, we present the following three points of comparison.

Firstly, we examine the distribution of the AP locations between the GMM groups such that the number of users being served by each AP is considered. Since GMM-4 has equal mixture component probability (and hence, an equal number of users) and variances for both groups, the proposed scheme allocates an equal number of APs to both groups, i.e., if $M = 4$, then each group is allocated 2 APs. This balanced allocation, which is henceforth represented as $(2, 2)$, is true for any realization of the user distribution. In contrast, the k-means++ scheme does not always result in a balanced $(2, 2)$ allocation. Through multiple realizations, it is observed to result in unbalanced $(3, 1)$ and $(1, 3)$ allocations as well. Examples of the allocations are illustrated in Fig. 3.11, where we show the final AP locations obtained using the Inter-AP Lloyd algorithm, resulting in a balanced $(2, 2)$ allocation for the proposed scheme and an unbalanced $(3, 1)$ allocation for the k-means++ scheme. In a balanced allocation, the division of the $K = 2000$ users is nearly even among the $M = 4$ APs, resulting in each AP serving about 500 users. This is in contrast to an unbalanced $(1, 3)$ or $(3, 1)$ allocation where one AP alone would serve nearly 1000 users while the other 3 APs would serve about 333 users each. In Table 3.4, we present the results of 1000 realizations of initializations and the percentage of balanced and unbalanced allocations are noted for $M = 4$, 6, and 8. Note that for $M = 6$, the possible unbalanced allocations would be $(4, 2)$, $(2, 4)$, $(1, 5)$, and $(5, 1)$, while for $M = 8$, they would be $(5, 3)$, $(3, 5)$, $(6, 2)$, $(2, 6)$, $(7, 1)$, and $(1, 7)$. Table 3.4 shows us that while the proposed scheme always results in a balanced allocation, the k-means++ scheme results in unbalanced allocations 30.1% of the time when $M = 4$ and increases with M up to 44.5% when $M = 8$.

Secondly, to observe the effects of balanced and unbalanced AP allocations on the rate performance, we plot one curve for the proposed scheme and one curve each for the balanced and unbalanced allocations arising from the k-means++ scheme. We use the same user distribution as before with $M = 4$ APs and $K = 250$ users, and in Fig. 3.12, we show the CDF of the minimum rates achieved in each case. It is clear that regardless of the scheme, the balanced allocation results in the same performance. The unbalanced allocation yielded by the k-means++ scheme, however, performs considerably worse. This performance loss is due to the fact that one AP in a GMM group has to serve all the users within the group. As such, this AP has a higher average user-AP distance than the APs in the balanced allocation case, and contributes primarily to the lower achievable rates. Henceforth, we can conclude that the proposed initialization scheme is preferable to the k-means++ scheme since the latter has an incidence of unbalanced AP allocations, which in turn results in reduced achievable rates.

Finally, we demonstrate that the benefits of the proposed scheme is not only limited to rate improvements, but also reduced net distance travelled by the APs over time as the user realization changes and the APs need to update their positions. We again use GMM-4 with $K = 250$ and $M = 4$ and consider two user realizations, at time instances T_1 and T_2. Both initialization schemes are used at T_1 and the Inter-AP Lloyd algorithm is applied to place the APs. Then, using the final AP positions at T_1 as initial positions for T_2, the placement algorithm is again applied. Accordingly, we conduct 1000 trials to compare the distances traversed by the APs from T_1 to T_2 in both schemes. It is found that the average distance travelled by the

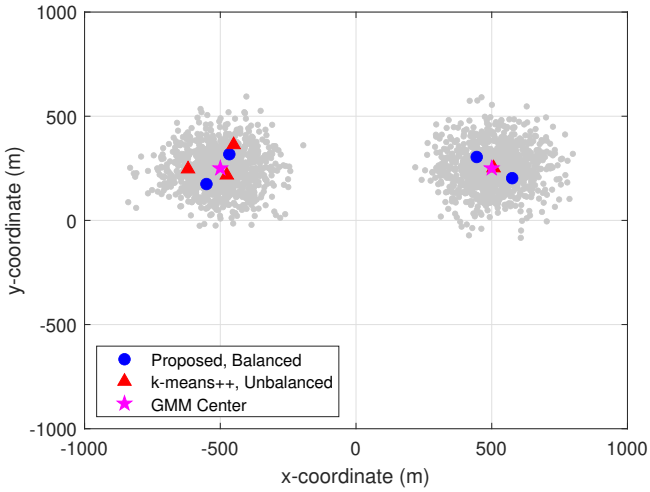

Figure 3.11: Representative example showing the final AP locations for the proposed and k-means++ initialization schemes for $M = 4$ under GMM-3. The proposed scheme always results in a balanced allocation while the k-means++ scheme may result in a balanced (not shown) or an unbalanced allocation, whose likelihood is given in Table 3.4.

Table 3.4: Percentage of Unbalanced Allocations Under the Proposed and k-means++ Initialization Schemes

M	Proposed [%]	k-means++ [%]
4	0	30.1
6	0	39.6
8	0	44.5

APs following the proposed scheme is 168.16 m, which is smaller compared to the averaged travelled distance of 175.08 m, yielded by the k-means++ scheme. The sole contributors to the higher traversed distances in the k-means++ scheme are the unbalanced allocations. Note that with smaller traversed distances, practical implementation issues such as the potential for collisions, UAV battery drainage, cell handovers, as well as signalling between NC and UAV-APs are reduced. Moreover, the benefits of the proposed scheme are also observed with the traditional Lloyd algorithm and HAPPA.

Experiment 7. In this last experiment, we obtain the AP placements, access rates, and the spectral access fractions of the LB-aware OWLA and CELA-α, and compare them to those of the Lloyd algorithm. In CELA-α, three trade-off values $\alpha = 0.9$, 1, and 1.75 are used. We choose the threshold as the distance of the AP to its nearest AP (cell-specific distance threshold). The AP locations obtained after the algorithms converge are shown in Fig. 3.13. OWLA results in APs that are placed away from those of the Lloyd algorithm, due to the LB-promoting weighted distortion measure. From CELA-α, we know that a higher value of α results in more user re-assignments. This is also evidenced by the fact that the AP locations are more different from those of the Lloyd algorithm as α increases. In order to quantitatively show the degree to which CELA-α perform user re-assignments, the occupancy of every cell for each of the considered α is provided in Table 3.5. We observe that while the occupancy level for equal occupancy would be $2000/8 = 250$, the occupancy levels vary significantly for the Lloyd algorithm. For smaller values of α, i.e., $\alpha = 0.9$, it is

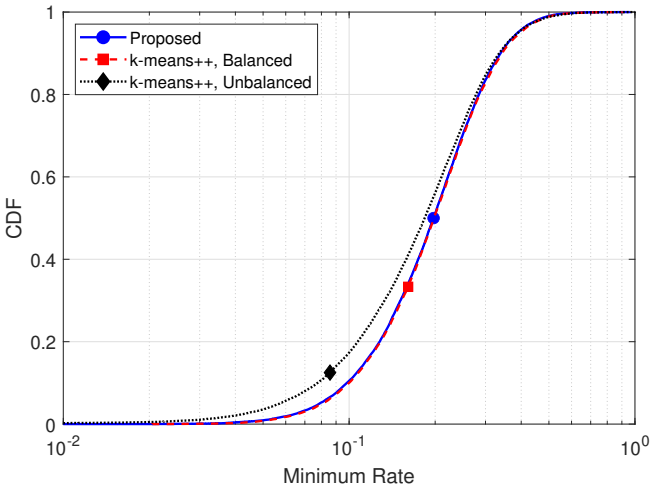

Figure 3.12: CDF plots of normalized minimum rate for the proposed and k-means++ balanced and unbalanced allocations with $M = 4$ and $K = 250$ under GMM-4.

observed that those cells with occupancy higher than the target in the Lloyd algorithm have their occupancy lowered. The opposite effect occurs for larger α values. As more users are re-assigned with the increase in α, more cells are able to attain the target value of 250 users. Particularly, for $\alpha = 1$, two cells and for $\alpha = 1.75$, an additional three cells attain this target occupancy. The occupancy observed in OWLA is also shown in the Table 3.5 and similarly to CELA-α, OWLA mitigates the issue of unbalanced loads in the cells.

Next, we show in Fig. 3.14 the cumulative distribution functions (CDFs) of the access rate. The rate curve corresponding to OWLA is observed to have a slightly inferior performance than that of the Lloyd algorithm. Also, as expected, when α increases to prioritize spectral access fairness among users within the cells, some degree of throughput loss is observed. Notice that the worst rate loss happens when $\alpha = 1.75$. Finally, Fig. 3.15 shows the CDFs of the spectral access fractions for OWLA and for the three values of α in CELA-α under consideration. It can be seen that this metric increases significantly with α and OWLA provides a performance slightly higher to that of $\alpha = 1$. Table 3.6 shows the percentage improvements for the access rate and spectral access fraction. Although rate reduction is observed to a degree, there is a significant improvement in spectrum access fraction. That is, the magnitude of access increase is much higher than that of the rate decrease. For instance, while the access rate suffers a reduction of 6.96%, the increase in spectral access fraction is nearly three-fold of that amount, at 20.83%, for $\alpha = 1.75$. Additionally, a key observation is that OWLA is able to achieve a higher spectral access fraction improvement at the cost of a lower access rate decrease, compared to CELA-α when $\alpha = 1$. As such, one should favor OWLA over CELA-α if a spectral access fraction of up to 12.5% above the one provided by Lloyd algorithm is required, for the considered user configuration. Above this mark, CELA-α is to be preferred. Finally, CELA-α has an inherent flexibility that enables the performance of the system to be governed by the trade-off factor α which can be based on specific system requirements, which OWLA cannot provide.

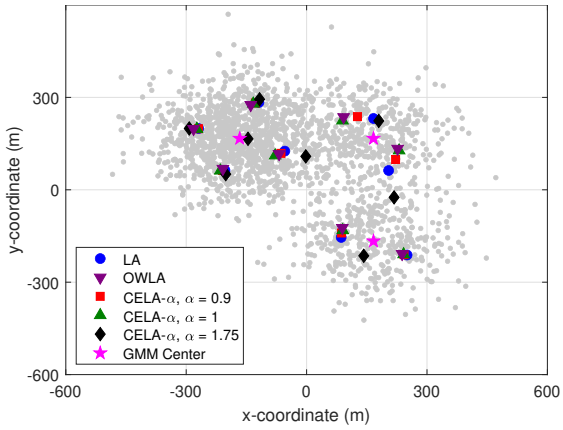

Figure 3.13: AP locations after convergence of the Lloyd algorithm, CELA-α with $\alpha = 0.9$, 1, 1.75, and OWLA for $M = 8$ and GMM-2.

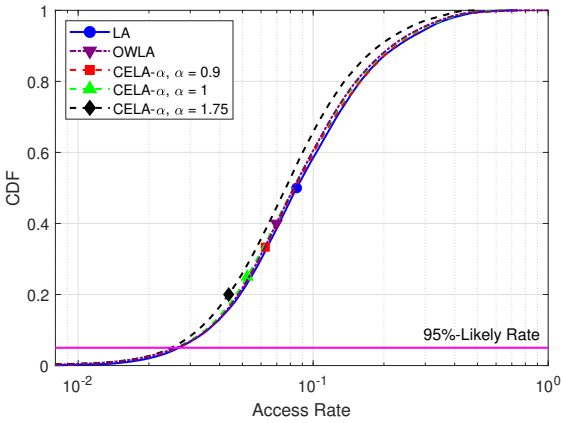

Figure 3.14: Per-user access rate CDF plots of the Lloyd algorithm, CELA-α with $\alpha = 0.9$, 1, 1.75, and OWLA for $M = 8$, and GMM-2.

Table 3.5: Cell Occupancy of LA, CELA-α, and OWLA

Algorithm	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_6$	$\mathcal{C}_7$	$\mathcal{C}_8$
LA	282	278	327	337	236	180	157	203
$\alpha = 0.9$	266	269	321	331	235	198	174	206
$\alpha = 1$	250	277	302	314	250	213	179	215
$\alpha = 1.75$	250	250	264	262	250	250	224	250
OWLA	281	288	302	286	221	219	189	214

3.8 Conclusion

In this chapter, we have addressed the AP placement problem in the small-cell uplink paradigm for hybrid network composed of terrestrial APs (T-APs) and AP-enabled UAVs (UAV-APs) in the presence of inter-cell interference (ICI). We accounted for ICI through signal-to-generated-interference-plus-noise ratio

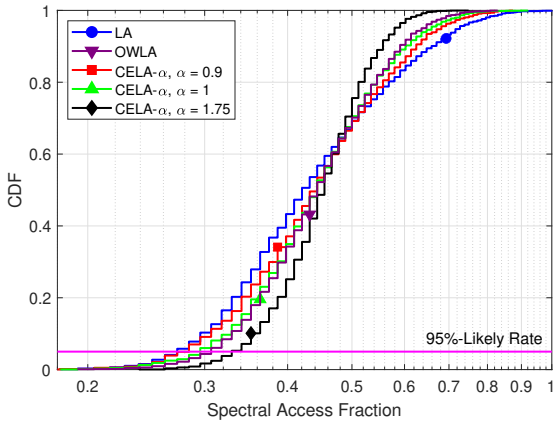

Figure 3.15: Spectral access fraction CDF plots of the Lloyd algorithm, CELA-α with $\alpha = 0.9$, 1, 1.75, and OWLA for $M = 8$, and GMM-2

Table 3.6: Percentage Improvements in 95%-Likely Achievable and Access Rates and Spectral Access Fraction for CELA-α and OWLA

Algorithm	Access Rate	Spectral Access Fraction
$\alpha = 0.9$	-1.46%	4.17%
$\alpha = 1$	-2.96%	8.33%
$\alpha = 1.75$	-6.96%	20.83%
OWLA	-2.28%	12.5%

(SGINR) instead of signal-to-interference-plus-noise ratio (SINR) and arrived at the Inter-AP Lloyd algorithm for fully flexible networks. It generated up to a 42.75% increase in 95%-likely achievable rate over the Lloyd algorithm. To account for the loss in capacity due to change in the user density, we devised a UAV-AP placement algorithm called the Hybrid AP Placement Algorithm (HAPPA). We showed that HAPPA, even with a small proportion of UAV-APs to T-APs, exhibited significant sum rate increase (up to 93.63%) over fixed networks and close to the ideal performance (as little as 2.02% difference) of fully flexible networks. Finally, we derived an initialization method for the Lloyd or any Lloyd-type algorithm, applicable to the Gaussian mixture model of user distribution. When compared to the popular k-means++ method, the proposed method always offered a relatively balanced (similar number of users served by each AP) allocation, in contrast to the unbalanced allocation that was possible in the k-means++ method at least 30% of the time, which resulted in a lower minimum rate over the balanced allocation.

We have also addressed the aspect of load balancing (LB) in throughput optimal small-cell access point placement. To account for LB in the placement process, we modified the Lloyd algorithm from vector quantization and presented two methods, namely the Occupancy Weighted Lloyd algorithm (OWLA) and Cell Equalized Lloyd algorithm-α (CELA-α), both of which yield increases in user access with minimal throughput loss. While OWLA utilizes a weighted distortion function, CELA-α adds an additional step in the Lloyd framework to achieve a degree of LB. Results show that both proposed algorithms achieve higher

spectral access (up to around 21%) while suffering a relatively minor reduction in throughput (up to around 7%), and CELA-α, through its trade-off factor α, allows for flexibility in deciding the degree of LB.

Chapter 3, in part, is a reprint with permission of the material as it appears in the papers: Govind Ravikumar Gopal, Bhaskar D. Rao, and Gabriel Porto Villardi, "Access point placement for hybrid UAV-terrestrial small-cell networks," in *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1826–1841, July 2021, and Govind Ravikumar Gopal, Bhaskar D. Rao, and Gabriel Porto Villardi, "Load balancing in small-cell access point placement," in *2023 IEEE 97th Vehicular Technology Conference (VTC-Spring)*, June 2023. The dissertation author was the primary investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-1617365 and Grant CCF-2124929, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

Appendices

3.A Proof of (3.19)

Consider the term

$$||\mathbf{p}-\mathbf{q}_{m'}||^2 = ||\underbrace{\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}}_{\mathbf{y}}+\underbrace{\mathbf{q}_{\mathcal{E}(\mathbf{p})}-\mathbf{q}_{m'}}_{\mathbf{x}}||^2. \tag{3.43}$$

From the conventional cell planning rules, it is clear that the distance between a user and its serving AP, denoted by $\mathbf{y}$, is smaller than the distance between the serving AP and any other interfering AP, denoted by $\mathbf{x}$. Hence, we can write

$$\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right| \leq \left|\left|\mathbf{q}_{\mathcal{E}(\mathbf{p})}-\mathbf{q}_{m'}\right|\right| \Rightarrow ||\mathbf{y}|| \leq ||\mathbf{x}||. \tag{3.44}$$

Each interfering AP m' can be classified into whether it is an immediate neighbor $\mathcal{IN}$ of AP $\mathcal{E}(\mathbf{p})$ or not, and correspondingly obtain the following relations

$$\begin{aligned} &||\mathbf{x}|| \geq ||\mathbf{y}||\,, \quad \forall m' \in \mathcal{IN}(\mathcal{E}(\mathbf{p})), \quad m' \neq \mathcal{E}(\mathbf{p}), \\ &||\mathbf{x}|| \gg ||\mathbf{y}||\,, \quad \forall m' \notin \mathcal{IN}(\mathcal{E}(\mathbf{p})), \quad m' \neq \mathcal{E}(\mathbf{p}). \end{aligned} \tag{3.45}$$

At this stage, we make the assumption that $||\mathbf{x}|| \gg ||\mathbf{y}||$ holds true $\forall m' \neq \mathcal{E}(\mathbf{p})$ and we simplify the term in (3.43) as

$$\begin{aligned} ||\mathbf{x}+\mathbf{y}||^2 &= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 + 2||\mathbf{x}||||\mathbf{y}||\cos\theta, \\ &= ||\mathbf{x}||^2\left(1+\frac{||\mathbf{y}||^2}{||\mathbf{x}||^2}+\frac{||\mathbf{y}||\cos\theta}{||\mathbf{x}||}\right), \\ &\approx ||\mathbf{x}||^2. \end{aligned} \tag{3.46}$$

This relation also holds true for $||\mathbf{p}-\mathbf{q}_{m'}||^\gamma$ with values of γ other than $\gamma = 2$. Hence, rearranging and taking the expectation gives us

$$\mathbb{E}_{\mathbf{p}}\left\{\frac{1}{||\mathbf{p}-\mathbf{q}_{m'}||^\gamma}\right\} \approx \frac{1}{\left|\left|\mathbf{q}_{m'}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^\gamma}. \tag{3.47}$$

3.B Proof of (3.40)

The bit allocation problem and the associated solution can be found in [41, Ch. 8] for the scalar (random variable) case. We extend this procedure to random vectors of dimension k and then derive the specific case for $k = 2$.

If we have N random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ that are to be quantized, then the overall distortion can be defined as

$$D = \sum_{i=1}^{N} g_i W_i(b_i), \tag{3.48}$$

where $W_i(b_i)$ is the distortion incurred when $\mathbf{X}_i$ is quantized using b_i bits and g_i's are some non-negative weights. From [41], $W_i(b_i)$ for k-dimensional vectors when the high-resolution approximation is applied is written as

$$W_i(b_i) = \frac{k}{k+2}\left(\frac{2\pi^{\frac{k}{2}}}{k\Gamma\left(\frac{k}{2}\right)}\right)^{-\frac{2}{k}} \times \left\{\int_{-\infty}^{\infty} [f_{\mathbf{X}_i}(\mathbf{x})]^{\frac{k}{k+2}}\, d\mathbf{x}\right\}^{\frac{k+2}{k}} 2^{-\frac{2b_i}{k}}. \tag{3.49}$$

where $\Gamma(\cdot)$ is the gamma function. We now use this expression to derive the bit allocation result for vectors of dimension k.

Lemma 1. *Consider N k-dimensional random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ of zero mean and covariance $\mathbf{\Sigma}_i$, and whose distributions $f_{\mathbf{X}_i}(\mathbf{x})$ are known for $i = 1, 2, \ldots, N$. Let $b_i = \log_2 q_i$ be the number of bits required in order to design a quantizer for $\mathbf{X}_i$ with q_i quantization levels. With B being the total number of bits available for the N quantizers (the bit budget), and the distortion defined as in* (3.49)*, the number of bits to be allocated to achieve optimal quantization, called the optimal bit allocation, is given by*

$$b_i = \overline{b} + \frac{k}{2}\log_2\frac{h_i}{H} + \frac{k}{2}\log_2\frac{g_i}{G}, \tag{3.50}$$

where:

$$\begin{aligned}
&\overline{b} = \frac{B}{N}, \qquad \epsilon(k) = \frac{k}{k+2}\left(\frac{2\pi^{\frac{k}{2}}}{k\Gamma\left(\frac{k}{2}\right)}\right)^{-\frac{2}{k}}, \qquad h_i = \epsilon(k)\left\{\int_{-\infty}^{\infty} [f_{\mathbf{X}_i}(\mathbf{x})]^{\frac{k}{k+2}}\, d\mathbf{x}\right\}^{\frac{k+2}{k}} = \epsilon(k)\,||f_{\mathbf{X}_i}(\mathbf{x})||_{\frac{k}{k+2}}^{\frac{k}{k+2}}, \\
&H = \left(\prod_{i=1}^{N} h_i\right)^{\frac{1}{N}}, \qquad G = \left(\prod_{i=1}^{N} g_i\right)^{\frac{1}{N}}.
\end{aligned} \tag{3.51}$$

Proof. Using the above notation in (3.49), we have

$$W_i(b_i) = h_i 2^{-\frac{2b_i}{k}}. \tag{3.52}$$

To minimize the overall distortion in (3.48) constrained by the bit budget, the objective function is

$$J = \sum_{i=1}^{N} g_i h_i 2^{-\frac{2b_i}{k}} + \lambda\left(\sum_{i=1}^{N} b_i - B\right), \tag{3.53}$$

which, on differentiating, we get

$$\begin{aligned}
&\frac{\partial J}{\partial b_i} = 0 \Rightarrow \lambda = \frac{2g_i h_i(\ln 2)}{k} 2^{-\frac{2b_i}{k}} \\
&\frac{\partial J}{\partial \lambda} = 0 \Rightarrow \sum_{i=1}^{N} b_i = B.
\end{aligned} \tag{3.54}$$

Solving this set of equations for b_i yields the required result. □

Noting that h_i is the norm of the pdf, we determine the closed-form expression when $\mathbf{X}_i$ is a multivariate Gaussian, as is the case in the GMM distribution.

Lemma 2. *For the multivariate k-dimensional Gaussian*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{3.55}$$

the η-norm defined by

$$||f_{\mathbf{X}}(\mathbf{x})||_\eta = \left\{\int_{-\infty}^{\infty} [f_{\mathbf{X}}(\mathbf{x})]^\eta \, d\mathbf{x}\right\}^{\frac{1}{\eta}}, \tag{3.56}$$

is given as

$$||f_{\mathbf{X}}(\mathbf{x})||_\eta = \left(\sqrt{(2\pi)^k |\mathbf{\Sigma}|}\right)^{\left(\frac{1}{\eta}-1\right)} \left(\frac{1}{\eta^{\frac{k}{2}}}\right)^{\frac{1}{\eta}}. \tag{3.57}$$

Proof.

$$\begin{aligned} ||f_{\mathbf{X}}(\mathbf{x})||_\eta &= \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \left\{\int_{-\infty}^{\infty} \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \eta\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x}\right\}^{\frac{1}{\eta}} \\ &= \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \left\{\sqrt{(2\pi)^k \left|\frac{1}{\eta}\mathbf{\Sigma}\right|}\right\}^{\frac{1}{\eta}} \\ &= \left(\sqrt{(2\pi)^k |\mathbf{\Sigma}|}\right)^{\left(\frac{1}{\eta}-1\right)} \left(\frac{1}{\eta^{\frac{k}{2}}}\right)^{\frac{1}{\eta}}. \end{aligned} \tag{3.58}$$

□

Theorem 1. *With M APs (the AP budget), the number of APs allocated to the L groups of a GMM with parameters p_l, $\boldsymbol{\mu}_l$, and $\mathbf{\Sigma}_l$, and K_l being the number of users in group l, $l = 1, 2, \ldots, L$, is*

$$u_l = \overline{u} + \log_2 \frac{h_l}{H} + \log_2 \frac{g_l}{G}, \tag{3.59}$$

where

$$\overline{u} = \frac{M}{L}, \qquad h_l = 4\sqrt{|\mathbf{\Sigma}_l|}, \qquad g_l = K_l, \qquad H = \left(\prod_{l=1}^{L} h_l\right)^{\frac{1}{L}}, \qquad G = \left(\prod_{l=1}^{L} g_l\right)^{\frac{1}{L}}. \tag{3.60}$$

Proof. Since the users are distributed in the $\mathbb{R}^2$ plane, we have $k = 2$. The AP budget is M, number of groups is L, and the distortion weights g_l correspond to the mixture component weights p_l, which in turn correspond to the number of users K_l. Since each GMM group l has $f_{\mathbf{X}_l}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_l, \mathbf{\Sigma}_l)$, from Lemma 2 we obtain

$$h_l = \frac{1}{2\pi} ||f_{\mathbf{X}_l}(\mathbf{x})||_{\frac{1}{2}} = 4\sqrt{|\mathbf{\Sigma}_l|}. \tag{3.61}$$

Making these parallels and simplifications in (3.50) of Lemma 1, we obtain the desired expression. □

Chapter 4

Cell-Free AP Placement

We now examine the problem of uplink cell-free access point (AP) placement in the context of optimal throughput. In this regard, we formulate two main placement problems, namely the sum rate and minimum rate maximization problems, and discuss the challenges associated with solving the underlying optimization problems with the help of some simple scenarios. As a practical solution to the AP placement problem, we suggest a vector quantization (VQ) approach. The suitability of the VQ approach to cell-free AP placement is investigated by examining three VQ-based solutions. First, the standard VQ approach, that is the Lloyd algorithm (using the squared error distortion function) is described. Second, the tree-structured VQ (TSVQ), which performs successive partitioning of the distribution space is applied. Third, a probability density function optimized VQ (PDFVQ) procedure is outlined, enabling efficient, low complexity, and scalable placement, and is aimed at a massive distributed multiple-input-multiple-output scenario. While the VQ-based solutions do not explicitly solve the cell-free AP placement problems, numerical experiments show that their sum and minimum rate performances are good enough, and offer a good starting point for gradient-based optimization methods. Among the VQ solutions, PDFVQ, with its distinct advantages, offers a good trade-off between sum and minimum rates.

4.1 Introduction

The problem of AP placement in cell-free systems is fairly novel and much prior research has not been conducted. As an example, [96] investigates the deployment of UAVs in a cell-free network to maximize minimum SINR using a gradient approach while considering pilot contamination and that not every AP communicates with all users. The authors of [97] consider the placement of APs in a distributed massive MIMO system as a combinatorial problem to minimize transmit powers while considering antenna radiation patterns and different channel models. While a cell-free system is not explicitly defined, the system model considered mimics such a system. A graph-based approach is found to yield significant power savings while ensuring placement with good coverage. In [98], 3-D placement of UAV-APs is considered to maximize the downlink sum rate and uses an alternating optimization method. A prior work by our group [60]

examines both sum rate and minimum rate maximization problems. This work solves the two problems using compressed sensing techniques by dividing the geographical area into regular grids. However, the approximations used do not solve the two placement problems optimally.

With the expected densification in future networks, it is necessary to design AP placement schemes that are not only practical, but can scale easily with the number of APs which is characteristic in a massive MIMO scenario. Moreover, as alluded to above in the examples of AP placement scenarios, such methods should be able to easily adapt to changing user environments. While the VQ approach (the Lloyd algorithm) have been utilized in prior works (e.g., [13]), none of the previous works have investigated the application of the Lloyd algorithm and other VQ approaches to solve the cell-free AP placement problem. In the VQ framework as applied to AP placement, the user positions are clustered and the cluster centers are the AP locations. The VQ technique considers a single user that communicates to its nearest AP with the objective function that utilizes a distortion function averaged over the random position of this user. It does not match the cell-free model where all users communicate to all APs. In spite of these limitations, VQ-based solutions have some features that make them worthy of consideration. By design, VQ offers not only a distributed solution (where APs are placed at different locations as opposed to a colocated solution) but can encourage cooperation (as is expected in the cell-free model) by placing APs closer to one another in areas of higher user density. Additionally, VQ solutions provide good initial points for gradient and learning-based methods to solve specific throughput problems. Hence, in this chapter, we explore and compare multiple VQ techniques, each with its own benefits, that can solve for AP locations in a cell-free network, with throughput as the performance measure.

Contributions

To the best of our knowledge, analysis of the cell-free AP placement problem in the context of throughput optimality along with the suitability of the VQ framework and the application of VQ-based methods to solve the same have not been addressed in literature. Hence, in this work, our contributions are as follows.

- We formulate the two main throughput optimal cell-free AP placement problems, namely the sum rate and minimum rate maximization problems. While these problems have been previously studied in small-cell works, a detailed discussion in the context of cell-free networks has not been presented. Starting from the simpler sum SNR problem, analysis of the sum rate problem is conducted and simple examples are shown to describe the possible solutions and to highlight the challenges associated with the general AP placement problem. The minimum rate problem is then discussed, also with some examples.

- Three VQ-based techniques to place APs are proposed and explored, namely standard VQ, which is the Lloyd algorithm, tree-structured VQ (TSVQ), and probability density function optimized VQ (PDFVQ). While the Lloyd algorithm provides a well-established method to place APs, there are the disadvantages of complexity and scalability. TSVQ, through successive partitioning of the user area,

places APs in such a way so as to foster cooperation. PDFVQ on the other hand, allows an efficient, less computationally intensive, and easily adaptable AP placement solution by using bit allocation, transform coding, and scalar quantization, and is especially suitable for a scaled network with a large number of APs.

- To complete the discussion on cell-free AP placement, we also investigate a limited AP cooperation scenario which utilizes the linear minimum mean square error (LMMSE) combiner. The performance of the placement solutions for this scenario is also provided.

4.2 System Model

In the cell-free system model, K single-antenna users are distributed over a geographical area with a probability density function $f_{\mathbf{P}}(\mathbf{p})$, with $\mathbf{p} \in \mathbb{R}^2$ as the random vector denoting the user position. M single-antenna APs serve these users, where $\mathbf{q} \in \mathbb{R}^2$ is the AP location. With $m = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, K$, a narrowband fading channel is considered with

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}, \tag{4.1}$$

where β_{mk} and $h_{mk} \sim \mathcal{CN}(0, 1)$ are the large- and small-scale fading coefficients, respectively, independent of each other and over coherent intervals. A general expression for β_{mk} is

$$\beta_{mk} = \frac{c z_{mk}}{||\mathbf{p} - \mathbf{q}_m||^{\gamma}}, \tag{4.2}$$

where c is a constant, z_{mk} is the shadow fading coefficient, and γ is the pathloss exponent. All APs cooperate with each other and are connected via error-free backhaul links to the network controller (NC). We consider the traditional cell-free uplink regime where all APs serve a smaller number of users in the same time-frequency resource. The uplink received signal at AP m is

$$y_m = \sum_{k=1}^{K} \sqrt{\rho_r} g_{mk} s_k + w_m, \tag{4.3}$$

where for user k, ρ_r is the transmit power, s_k is the data symbol with $\mathbb{E}\{|s_k|^2\} = 1$, and $w_m \sim \mathcal{CN}(0, 1)$ is the additive noise. The received signal vector at the NC from all M APs using (5.3) can be written as

$$\mathbf{y} = \sum_{k=1}^{K} \sqrt{\rho_r} \mathbf{g}_k s_k + \mathbf{w}, \tag{4.4}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_M]^T$, $\mathbf{g}_k = [g_{1k}, g_{2k}, \ldots, g_{Mk}]^T$, and $\mathbf{w} = [w_1, w_2, \ldots, w_M]^T$. When a combiner $\mathbf{v}_k$ is used to estimate data symbols of user k as $\hat{s}_k = \mathbf{v}_k^H \mathbf{y}$, the per-user achievable rate is $R_k = \mathbb{E}\{\log(1 + \phi_k^{\mathbf{v}_k})\}$, where the expectation is over all the small-scale and shadow fading coefficients, and the SINR [19] is

$$\phi_k^{\mathbf{v}_k} = \frac{\rho_r \mathbf{v}_k^H \mathbf{g}_k \mathbf{g}_k^H \mathbf{v}_k}{\mathbf{v}_k^H \mathbf{v}_k + \sum\limits_{\substack{k'=1 \\ k' \neq k}}^{K} \rho_r \mathbf{v}_k^H \mathbf{g}_{k'} \mathbf{g}_{k'}^H \mathbf{v}_k}. \tag{4.5}$$

One such combiner is the zero forcing (ZF) detector, and results in the processed signal at the NC as $\mathbf{r} = \left(\mathbf{G}^H\mathbf{G}\right)^{-1}\mathbf{G}^H\mathbf{y}$, where $\mathbf{G} = [g_{mk}]$ is a $M \times K$ matrix consisting of the channel coefficients [60]. The achievable per-user SNR in this case is

$$\psi_k^{\mathrm{ZF}} = \frac{\rho_r}{[(\mathbf{G}^H\mathbf{G})^{-1}]_{kk}}. \tag{4.6}$$

Using an asymptotic approximation for the SNR as outlined in [15,60], the per-user SNR can also be written as

$$\frac{1}{M}\psi_k^{\mathrm{ZF}} \xrightarrow[M\to\infty]{\text{a.s.}} \rho_r\overline{\beta}_k, \tag{4.7}$$

where

$$\overline{\beta}_k \triangleq \lim_{M\to\infty}\frac{1}{M}\sum_m \beta_{mk}. \tag{4.8}$$

4.3 Throughput Formulations for the Cell-Free AP Placement Problem

There are two main formulations for cell-free AP placement in terms of throughput optimality:

- *Sum rate maximization*, which involves the sum of the rates of all users, as follows

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\max}\sum_{k=1}^{K}\log(1+\phi_k^{\mathbf{v}_k}). \tag{4.9}$$

  Note that this problem is identical to the average rate maximization problem by assuming a sample mean and taking the average over the user distribution $f_{\mathbf{P}}(\mathbf{p})$.

- *Minimum rate maximization*, where the minimum of the rates among all of the users is maximized

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\max}\ \min_k\ \log(1+\phi_k^{\mathbf{v}_k}). \tag{4.10}$$

  The notion of fairness, which is important in a cell-free system since all users are served by all APs, is enforced by the minimum rate problem in (4.10) as opposed to the sum rate problem in (4.9). In practice, the 95%-likely rate, which represents the best rate among the worst 5% of the users, is used as a measure to evaluate the network minimum rate performance. Hence, the max-min rate is adjusted to the 95%-likely rate, which is more robust. This is subsequently addressed in Section 4.4.4.

It should be noted that finding solutions that address both of the above metrics, although challenging, is ideal. It is desirable to achieve an optimal trade-off between sum rate and minimum rate. In the ensuing sections, we discuss the above formulations. Our analysis of the sum rate maximization problem is preceded by the simpler sum SNR maximization problem.

4.3.1 Sum SNR Maximization

When throughput is measured by utilizing SNR alone, we can obtain a simpler sum throughput problem to (4.9) by using the ZF SNR ψ_k^{ZF} from (5.5) and by replacing the summation with an expectation (the factor of $1/K$ has been neglected). The average is taken over the user position and the set $\mathcal{A} = \{z_{mk}, \forall m\}$ consisting of all shadow fading coefficients between each user and AP. The optimization problem is written as

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\max}\ \mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\psi_k^{\mathrm{ZF}}\right\}. \tag{4.11}$$

We can simplify the above objective function using the approximation in (5.6) and β_{mk} from (5.2) in the following manner

$$\begin{aligned}
\mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\psi_k^{\mathrm{ZF}}\right\} &= \mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\rho_r \sum_{m=1}^{M} \beta_{mk}\right\}, \\
&= \mathbb{E}_{\mathbf{p}}\left\{\rho_r \sum_{m=1}^{M} \mathbb{E}_{\mathcal{A}}\left\{\beta_{mk}\right\}\right\}, \\
&\overset{(a)}{=} \mathbb{E}_{\mathbf{p}}\left\{\sum_{m=1}^{M} \frac{c'}{||\mathbf{q}_m - \mathbf{p}||^{\gamma}}\right\}, \\
&\overset{(b)}{=} c' \sum_{m=1}^{M} \mathbb{E}_{\mathbf{p}}\left\{\frac{1}{||\mathbf{q}_m - \mathbf{p}||^{\gamma}}\right\},
\end{aligned} \tag{4.12}$$

where $c' = c\rho_r \mathbb{E}_{\mathcal{A}}\{z_{mk}\}$, $\mathbf{p}$ is the position of user k in (a), and (b) uses the fact that the expectation is a linear operator. Note that c' can be ignored as it does not effect the optimization problem.

The following observations can be made regarding the solution of the above objective function in (4.12).

- A colocated solution is evident since there is no dependence between the terms in the summation associated with each AP m and hence, the optimization for each AP can be performed separately. The suggested colocated solution may not be a unique solution and multiple global and/or local maxima may exist for the optimization problem. The complete characterization of the solution, however, depends on two factors, namely the pathloss exponent γ used as well as the shape of the user distribution $f_{\mathbf{P}}(\mathbf{p})$ over which the expectation is taken, e.g., uni-modal versus multi-modal density functions.

- In (4.12), although the norm $||\mathbf{q}_m - \mathbf{p}||$, i.e., the Euclidean distance is strictly convex [99] in $\mathbf{q}_m$, its inverse is neither concave nor convex and is also undefined at $\mathbf{q}_m = \mathbf{p}$ (this can, however, be avoided by adding a small positive quantity to the denominator, which could also account for the height of the AP). The sum of the expectation of the inverse over the APs thus also is neither convex or concave. Additionally, when a uni-modal distribution is assumed, we can expect a colocated solution alone with a unique maximum. However, when a multi-modal distribution is considered, it is expected that multiple global maxima exist and distributed solutions may be obtained. This is explored in Section 4.3.3.

It is worth noting that the sum SNR problem has been previously addressed in part in [60]. In this work, while dividing the user area into regular grid points, the sum rate maximization problem is

upperbounded to a sum SNR problem and using a compressed sensing framework, approximated to a linear program (called the max-sum algorithm). However, as is expected from a sum SNR problem (as discussed above), most APs are concentrated around high user density regions (a near-colocated solution). Although a high sum rate can be achieved with this solution, users far away from the APs are severely affected in terms of throughput, resulting in poor minimum rate performance. Thus, this solution is not suitable for cell-free AP placement when fairness is considered.

4.3.2 Sum Rate Maximization

Returning to the sum rate maximization problem, we rewrite (4.9) by using the simplifications assumed before, as follows

$$\underset{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_M}{\arg\max}\ \mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\log\left(1+\psi_k^{\mathrm{ZF}}\right)\right\}, \tag{4.13}$$

and the objective function, utilizing the approximation in (5.6), can be rewritten as

$$\mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\log\left(1+\psi_k^{\mathrm{ZF}}\right)\right\} = \mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\log\left(1+\rho_r\sum_{m=1}^{M}\beta_{mk}\right)\right\}. \tag{4.14}$$

Similar to the SNR problem outlined before, this objective function is neither concave nor convex. However, unlike the former, the term associated with each AP m in (4.14) cannot be decoupled from the terms associated with the rest of the APs. Hence, for both uni-modal and multi-modal distributions, we can expect only distributed solutions that maximize the sum rate.

In summary, for both the sum SNR and sum rate problems there may be multiple local optima suggesting that the optimization problem is complex and challenging.

4.3.3 Examples for Sum SNR and Sum Rate Maximizations

Given the abovementioned complexity in solving the sum SNR and sum rate problems, we now attempt to understand the problems and their solutions better. For this purpose, we explore simple examples where the aforementioned two problems are solved, and where the multiple local optima are studied to develop intuition and insight.

User distribution considered

For tractability, we consider a simple 1-D scenario where users are distributed along a line, and the placement of four APs for both the sum SNR and sum rate maximization problems. For this purpose, we assume a bi-modal distribution since it is the simplest among multi-modal distributions that can exhibit the multiple maxima as discussed in Section 4.3.1. With the user position denoted by p, the PDF of the bi-modal Gaussian considered here is

$$f_P(p) = \pi_1\mathcal{N}(p|\mu_1,\sigma_1^2) + \pi_2\mathcal{N}(p|\mu_2,\sigma_2^2), \tag{4.15}$$

where for each Gaussian i, $i = 1, 2$, π_i is the probability such that $\pi_1 + \pi_2 = 1$, μ_i is the mean, and σ_i is the standard deviation. To generate different solution structures for the two optimization problems, we consider two distinct configurations of the user distribution:

Conf. 1: $\pi_1 = \pi_2 = 0.5$, $\mu_1 = -3$, $\mu_2 = 3$, and $\sigma_1 = \sigma_2 = 1$.

Conf. 2: $\pi_1 = 0.35$, $\pi_2 = 0.65$, $\mu_1 = -3$, $\mu_2 = 4$, and $\sigma_1 = \sigma_2 = 1$.

Configuration 1 is symmetric about the origin while configuration 2 is asymmetric.

Definitions of SNR and rate

The four APs have locations q_1, q_2, q_3, and q_4, and the ZF SNR[1] for user k can be calculated as

$$\begin{aligned} \psi_k^{\mathrm{ZF}} &= \beta_{1k} + \beta_{2k} + \beta_{3k} + \beta_{4k}, \\ &= \frac{1}{(p_k - q_1)^2} + \frac{1}{(p_k - q_2)^2} + \frac{1}{(p_k - q_3)^2} + \frac{1}{(p_k - q_4)^2}, \end{aligned} \tag{4.16}$$

where, for simplicity, the transmit power is set to unity and the definition of the large-scale fading coefficient β_{mk} (from (5.2)) assumes that shadow fading is absent, the pathloss exponent is two, and the constant is set to one. These assumptions do not change the conclusions that are obtained in this section. The sum SNR and sum rate quantities are then defined as $\sum_{k=1}^{K} \psi_k^{\mathrm{ZF}}$ and $\sum_{k=1}^{K} \log(1 + \psi_k^{\mathrm{ZF}})$, respectively. Note that for implementation purposes, a small quantity ϵ is added to the denominator of β_{ik}, $i = 1, 2, 3, 4$, to prevent ψ_k^{ZF} from approaching infinity.

AP location solutions considered

To evaluate and understand the sum SNR and sum rate performances of the system, we study various AP placement scenarios. First, a colocated solution is considered where all four APs are situated at the same location. This location is found by sweeping the AP position denoted by q across the span of the user locations. Second, multiple semi-distributed solutions are selected. Instead of all APs at one location, they can be allocated to each of the Gaussians and placed at their respective means. In this scenario, we consider three situations, namely when two APs each are placed at μ_1 and μ_2, three APs are at μ_1 and one AP is at μ_2, and one AP is at μ_1 and three APs are at μ_2. These three situations are termed 'Distributed (2+2)', 'Distributed (3+1)', and 'Distributed (1+3)', respectively. Third, a fully distributed scenario involves starting from a distributed (2+2) solution and moving the two APs within each Gaussian away from each other until the maximum is achieved. Note that this fully distributed solution represents only a local maximum. Finally, we have the solution obtained by applying the standard Lloyd algorithm to the user distribution. In this iterative solution, the users are clustered using the squared Euclidean distance and the cluster centers (centroids) are determined to be the AP locations. More explicit details are provided in Section 4.4.1.

[1]While the asymptotic approximation for the SNR expression used does strictly not hold here, it is good enough to illustrate the complexities of the maximization problems considered.

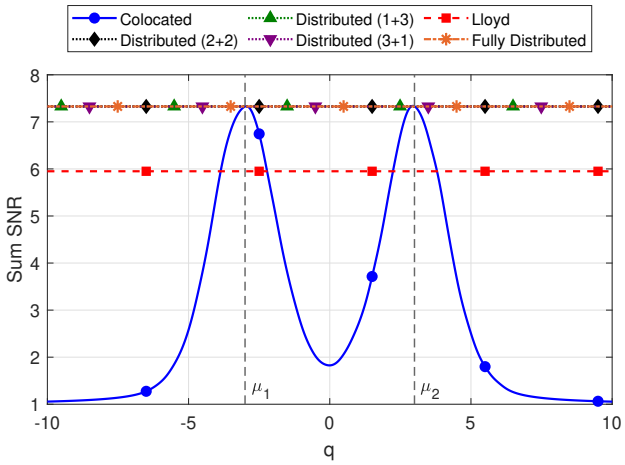

Figure 4.1: Sum SNR for different AP location scenarios under user configuration 1.

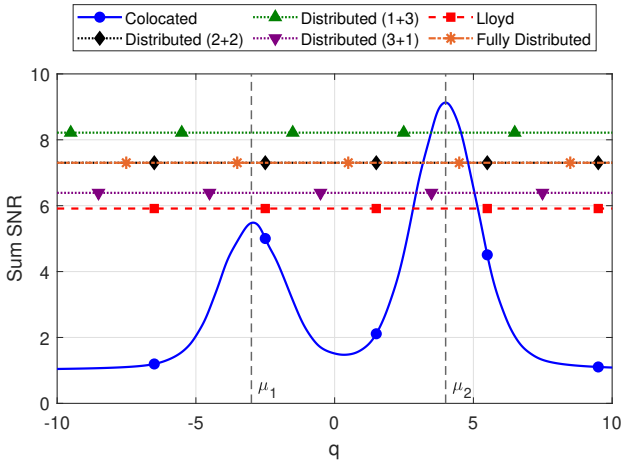

Figure 4.2: Sum SNR for different AP location scenarios under user configuration 2.

Results for sum SNR

The results of sum SNR under user configurations 1 and 2 are shown in Fig. 4.1 and Fig. 4.2, respectively. Note that for the colocated solution, the sum SNR obtained as a function of location q is plotted. For the other solutions, a line is drawn corresponding to the sum SNRs obtained.

For *configuration 1*, it is observed that the peak for the colocated system occurs at the two means μ_1 and μ_2. The three distributed scenarios as well as the fully distributed scenario offer the same peak sum SNR value as the colocated case. It is noted that the AP locations in the fully distributed case are the same as in distributed (2+2). Further, the Lloyd solution (which is also a fully distributed scenario) offers a lower sum SNR value. The sum SNR maximization for configuration 1 thus has multiple local maxima, including both colocated and semi-distributed solutions.

For *configuration 2*, the Gaussian with mean μ_2 has a higher probability, with the result that a colocated solution with all four APs at mean μ_2 yields the highest sum SNR. Among the distributed solutions, it is observed that a higher allocation of APs at μ_2 favors a higher sum SNR, however, with the colocated

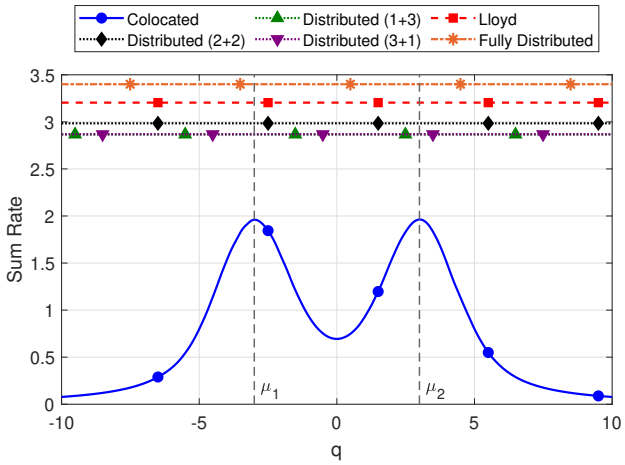

Figure 4.3: Sum rate for different AP location scenarios under user configuration 1.

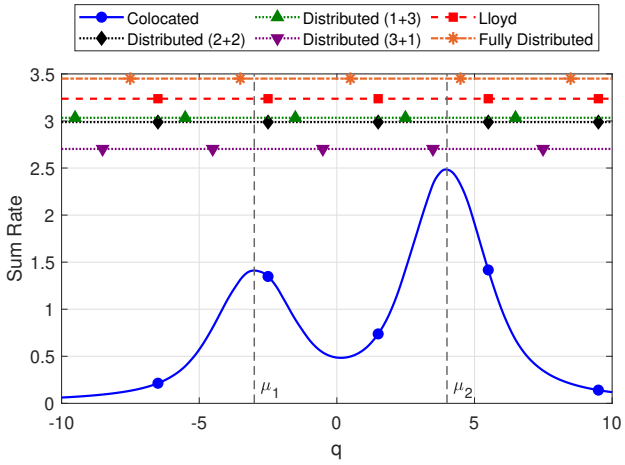

Figure 4.4: Sum rate for different AP location scenarios under user configuration 2.

solution offering the highest. Again, both the fully distributed and distributed (2+2) solutions have the same AP locations and sum SNR values, with the Lloyd solution performing the worst. In summary, the above results for sum SNR show that the AP locations that maximize the same may be colocated or distributed depending on the user distribution, as discussed in Section 4.3.1.

Results for sum rate

The results of the sum rate metric for user configurations 1 and 2 are shown in Fig. 4.3 and Fig. 4.4, respectively, and differ from the sum SNR plots above. In configuration 1, it is clear that a colocated solution performs poorly and that distributed and fully distributed solutions offer a higher sum rate. Since the distribution is symmetric, both distributed (1+3) and distributed (3+1) solutions have the same sum rate, which is lower than that of distributed (2+2). The Lloyd solution performs better than the distributed solutions as no APs are colocated, and the fully distributed solution performs the best. Note that the fully distributed solution does not represent the optimum solution due to the method with which

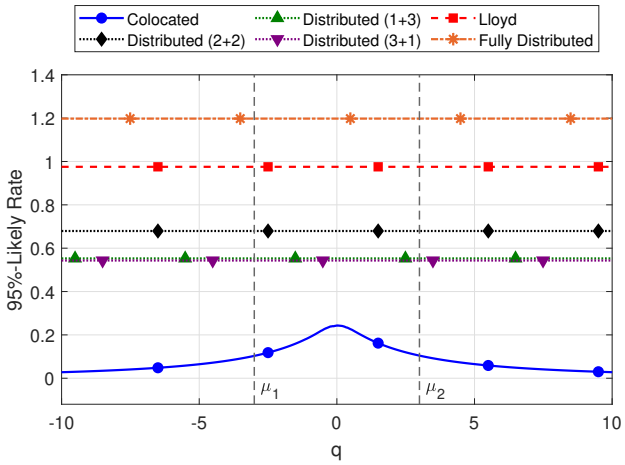

Figure 4.5: 95%-likely rate for different AP location scenarios under user configuration 1.

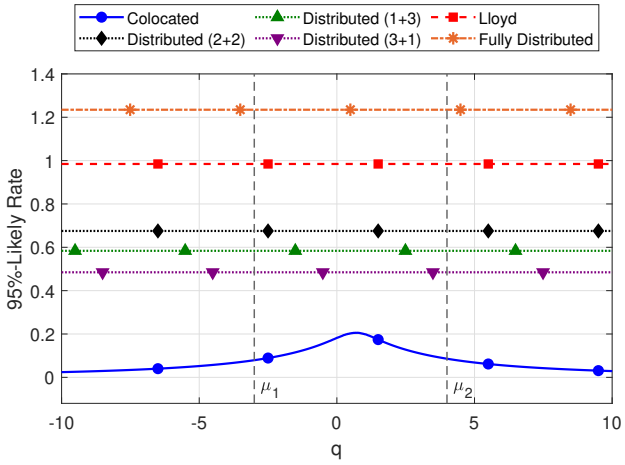

Figure 4.6: 95%-likely rate for different AP location scenarios under user configuration 2.

the locations are generated. In configuration 2, the colocated solution still performs the worst. However, since the Gaussian at mean μ_2 carries a higher proportion of users, the distributed (1+3) solution performs better than both distributed (2+2) and distributed (3+1). Finally, as in configuration 1, both the Lloyd and the fully distributed solution perform better than the distributed solutions.

Results for minimum rate

We now plot the 95%-likely rates corresponding to the different AP locations for configurations 1 and 2, in Fig. 4.5 and Fig. 4.6, respectively. Note that the fully distributed solution in these plots maximizes the 95%-likely rates as opposed to the sum SNR or sum rate. Even though all distributed solutions generate the same peak sum SNR for configuration 1 as shown in Fig. 4.1, the users will not achieve the same minimum rate performance in all cases. For this user configuration, distributed (1+3) and distributed (3+1) have lower 95%-likely rates than distributed (2+2). Additionally, the Lloyd solution has a higher minimum rate and the fully distributed solution generates the highest 95%-likely rate. Clearly, the colocated solution

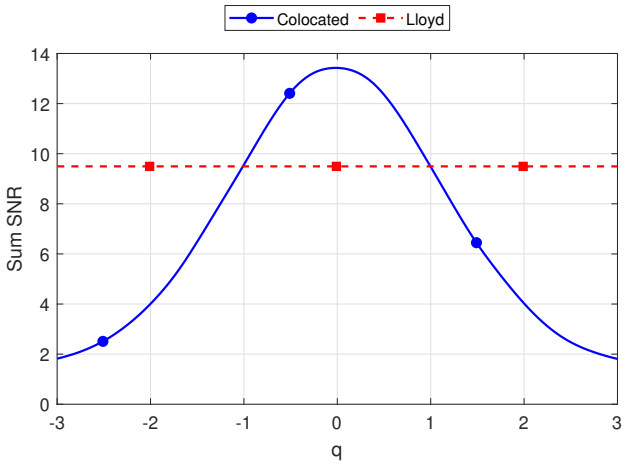

Figure 4.7: Sum SNR for colocated and Lloyd solutions under uni-modal distribution.

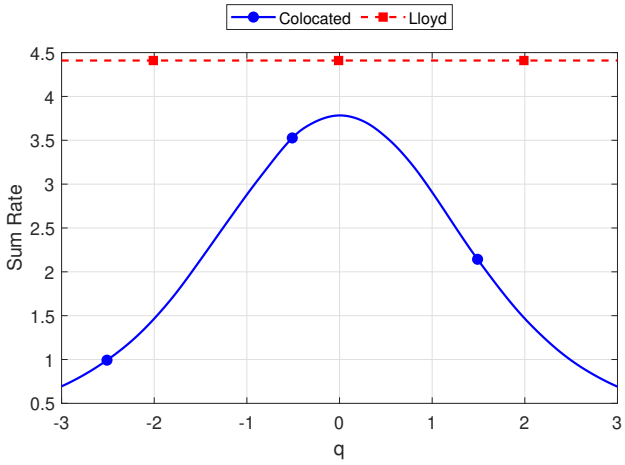

Figure 4.8: Sum rate for colocated and Lloyd solutions under uni-modal distribution.

exhibits the worst performance. The same observations are noted for configuration 2, with the difference being that distributed (1+3) generates a higher 95%-likely rate than distributed (3+1) due to the higher user proportion around mean μ_2.

Results for a uni-modal user distribution

As a final note and to complete the discussion on the influence of the distribution shape on the solutions, we also plot the sum SNR and sum rate performances for a uni-modal Gaussian distribution (with zero mean and unit variance), in Fig. 4.7 and Fig. 4.8, respectively. Like the bi-modal examples shown prior, the colocated solution is preferred in the sum SNR case over the distributed Lloyd solution. The opposite is true for the sum rate case. It is also to be noted that unlike the bi-modal scenario, there is only one colocated solution that provides the peak sum SNR.

In conclusion, all observations made in the context of the simple examples support the complex nature of the AP placement for sum SNR and sum rate maximization problems (Section 4.3.1 and Section

4.3.2).

4.3.4 Minimum Rate Maximization

For the sum SNR maximization case, we showed both colocated and semi-distributed cases were favored. It was also postulated that the minimum rate performance of these solutions would be inferior to that provided by a fully distributed solution (like the Lloyd solution). Even in the sum rate maximization case, we have not addressed the minimum rates of the users explicitly and we can expect that there are users that lie far away from the AP locations, resulting in significantly low throughput to these users, which is not ideal from a fairness point of view. The minimum rate maximization problem is thus of interest and has been defined above in (4.10). In this chapter, we do not discuss this problem as it has been addressed and solved using a grid-based approach in [60,61] and is called the max-min algorithm. Such a solution is useful since the problem can be converted into a convex problem, which can be solved easily. It is to be noted that the grid structure assumed leads to approximate solutions and a finer grid is necessary to obtain the optimal locations. In Section 5.7, we compare the performance of the max-min algorithm with the Lloyd algorithm.

4.4 Vector Quantization Approaches

In this section, as practical solutions to cell-free AP placement, we investigate how VQ techniques can be applied to sum rate and minimum rate maximizations. While acknowledging that all users are served by all APs in the cell-free system under consideration, we assume that for the purposes of AP placement, VQ is implemented by assuming that each user is associated with its geographically nearest AP by adopting the squared Euclidean distance (error) as the distortion measure of interest. For completeness, we first start by presenting the standard VQ approach and then motivate why the VQ framework is useful for the placement problem, followed by the other VQ techniques.

4.4.1 Standard VQ

The standard and simplest VQ technique is the Lloyd algorithm [41] that utilizes the squared Euclidean distance between APs and users as the distortion measure. This squared error (SE) distortion between a user at $\mathbf{p}$ and an AP at $\mathbf{q}_m$ is denoted as follows

$$d_{\mathrm{SE}}(\mathbf{p}, \mathbf{q}_m) = ||\mathbf{p} - \mathbf{q}_m||^2 . \tag{4.17}$$

Details of how the Lloyd algorithm can be applied to AP placement have been elucidated in Chapter 2. We provide the algorithm (Algorithm 12) below, where the K user positions denoted by $\mathbf{p}_k$ are generated as realizations from the user distribution $f_{\mathbf{P}}(\mathbf{p})$.

Why Use the VQ Approach? Although we focus on the standard VQ technique as the illustration here, the advantages listed below follow for the subsequent VQ techniques described.

- *Distributedness:* The VQ approaches, through their formulations, are designed to provide a distributed solution. The standard Lloyd algorithm using squared error distortion minimizes the distance of a user

Algorithm 12 Lloyd Algorithm

1: Initialize random AP locations $\mathbf{q}_1^{(0)}, \mathbf{q}_2^{(0)}, \ldots, \mathbf{q}_M^{(0)}$.

2: Use the nearest neighbor condition (NNC) to determine the cells $\mathcal{C}_1^{(i+1)}, \mathcal{C}_2^{(i+1)}, \ldots, \mathcal{C}_M^{(i+1)}$ such that

$$\mathcal{C}_m^{(i+1)} = \left\{ \mathbf{p}_k : d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq d_{\mathrm{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m \right\}.$$

3: Use the centroid condition (CC) to determine the AP locations $\mathbf{q}_1^{(i+1)}, \mathbf{q}_2^{(i+1)}, \ldots, \mathbf{q}_M^{(i+1)}$ such that

$$\mathbf{q}_m^{(i+1)} = \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \mathbf{p}_k.$$

4: Repeat from step 2 until convergence (MSE falls below a threshold).

to its closest AP, averaged over the entire distribution, obtains a distributed solution by ensuring that at least one AP is close to each user. It is important to note that as a consequence, the Lloyd algorithm solution also addresses the minimum rate of the system in an effective manner, while the minimum rate metric itself is not explicitly contained in its objective function. Additionally, the Lloyd algorithm alone exhibits a space-filling advantage [**?**] which refers to the generation of the Voronoi cells $(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M)$ and therefore the AP locations that span the space of the user distribution efficiently.

- *Cooperation:* While allowing distributedness, the VQ formulations which do not explicitly account for cooperation, also place APs close to one another. This clustering occurs especially in areas of high user density, thus encouraging cooperation that is expected in a cell-free system and addresses the system sum rate. In the specific the case of the Lloyd algorithm, the objective function indirectly fosters cooperation since the AP density is proportional to a power of the user density under a high resolution approximation [41], as follows

$$g_{\mathbf{P}}(\mathbf{p}) = \frac{f_{\mathbf{P}}^{\frac{1}{2}}(\mathbf{p})}{\int f_{\mathbf{P}}^{\frac{1}{2}}(\mathbf{p}')\mathrm{d}\mathbf{p}'}. \tag{4.18}$$

- *Initialization:* Although it is possible to determine solutions for both problems through gradient methods, such methods usually generate local optima resulting in lower-than-expected performances. The VQ approaches, due to the distributed nature of their solutions, are able to provide suitable starting points for such gradient-based methods.

4.4.2 Tree-Structured VQ

Tree-structured VQ (TSVQ) [41] is an alternate VQ approach where the codebook search time is reduced compared to standard VQ. In this technique, the input training set is partitioned into a hierarchy of Voronoi regions, which allows a tree to be generated for encoding. Thus, TSVQ differs from the standard VQ discussed above in that the final required codebook is generated by the successive splitting of intermediate codebooks, starting with a single centroid. The Lloyd algorithm (Algorithm 12) is applied to each stage of the hierarchy and its application is confined to the partitioned training set of the previous stage.

Why use TSVQ? Designed primarily for relatively fast codebook search properties, TSVQ has the following benefits.

- *Initialization:* An advantage of TSVQ is that initialization of the AP locations is not required, since it starts with a single centroid. Compared to the standard VQ, this avoids the random initializations (considered sub-par) and the calculations needed for advanced initialization methods such as k-means++ [95].
- *Cooperation:* Due to the successive splitting of the intermediate codepoints, we can expect pairs of APs to be closer to each other compared to standard VQ. This enables cooperation among APs, increasing the system sum rate.
- *Flexibility:* When the number of APs is changed for the same user distribution, codepoints can either be merged (when the number of APs is reduced) or split (when increased). The choice of codepoints for merging or splitting will be on the basis of the rate performance associated with the codepoints.

It should be noted that in general, TSVQ does not find the closest AP to each user and there is a small decrease expected in the performance of TSVQ (in terms of mean squared error) when compared to standard VQ.

TSVQ Algorithm. In our implementation of TSVQ for cell-free AP placement, we limit ourselves to balanced binary trees, i.e., at each stage, every intermediate codepoint is split into two codepoints, to favor the lowest complexity. The TSVQ algorithm is outlined in Algorithm 13. The stages are indexed by j (root node is stage 0) and the set of codepoints is represented by $\mathcal{P}_j$ for stage j. The codepoints are indexed by i, and the number of codepoints increases at each stage. The codepoints generated after the algorithm converges are the required AP positions. $\mathcal{R}$ indicates the set of all users while $\mathcal{R}_i$ denotes the set of users associated with codepoint i. The splitting of codepoints is performed by generating two perturbations of the original codepoint.

Algorithm 13 Tree-Structured Vector Quantization Algorithm

1: Initialize $\mathcal{P}_0$ with the codepoint of all users in $\mathcal{R}$.
2: Split the codepoint(s) in $\mathcal{P}_j$ into two.
3: Apply Lloyd algorithm (Algorithm 12) to each split pair for the set of users $\mathcal{R}_i$ at stage j.
4: Partition $\mathcal{R}_i$ at stage j into two sets corresponding to new codepoints.
5: Update $\mathcal{P}_j$ with new codepoints.
6: Repeat from step 2 for next stage $j+1$ until M codepoints are generated.

4.4.3 PDF Optimized VQ

So far, we have considered a full-scale version of VQ (the Lloyd algorithm) and TSVQ, which involves hierarchical codebook generation. Both flavors of VQ, however, come with disadvantages such as complexity, scalability, and learnability which will be elaborated shortly. In this section, we outline the

probability density function optimized VQ (PDFVQ) procedure, first developed in [100], for use in cell-free AP placement that can address these shortcomings. In PDFVQ, an efficient quantizer is generated by first estimating the distribution PDF (of the users) using the expectation-maximization (EM) algorithm [94] and assuming a Gaussian mixture model (GMM). The GMM considered is of the form

$$f_{\mathbf{P}}(\mathbf{p}) = \sum_{l=1}^{L} p_l \mathcal{N}\left(\mathbf{p}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right), \tag{4.19}$$

where L is the number of mixture components (clusters), and p_l, $\boldsymbol{\mu}_l$, and $\boldsymbol{\Sigma}_l$ are the probability, mean, and covariance matrix, respectively, of mixture component l. Then, given the total bit budget (which corresponds to the total number of APs), by leveraging both bit allocation and transform coding, closed-form expressions are defined to allocate bits to each cluster and along every dimension (x- and y-coordinates of the 2-D user density) so that scalar quantizers can be used to generate the required codebook (the AP locations).

Why use PDFVQ? Although designed for high-dimensional source inputs, this procedure addresses the shortcomings of the two prior VQ approaches in the following manner.

- *Complexity:* In the two prior VQ approaches, the computational complexity and memory required are high. In TSVQ, the Lloyd algorithm is applied a number of times depending on the codebook splits performed to obtain the required number of APs. The overall complexity of the Lloyd algorithm is $\mathcal{O}(2KMI)$, where the factor '2' is due to the 2-D user density and I denotes the number of iterations required for convergence. It is also observed that the number of executed iterations I increases with the number of APs M. In addition to the simple expressions for AP allocations for each cluster and dimension of the 2-D user density, through the use of scalar quantizers for each dimension, PDFVQ enjoys lower complexity as the quantizers work on both a lower number of users and APs, consequently performing a lower number of iterations. Further, the scalar quantizers can also be implemented by means of look-up tables [101] that determine the codebook. Thus, there exists a simple closed-form mapping from user density to the codepoints (AP locations). The memory requirements in PDFVQ are also small as a result of transform coding.

- *Learnability:* It is expected that the user density varies over time, in which case the AP positions must be adapted to account for this change. Both prior VQ schemes are not amenable to a learning environment as they cannot adapt quickly to changes in the user distribution as well as number of APs, and their computationally complex procedures must be repeated to optimize to the new environment. In PDFVQ, when the source distribution changes, both the density estimation and codebook generation steps must be implemented, however, only the parameters of the density have to be learned (the EM algorithm is easily updated using existing values). If the number of APs only is changed, then the closed-form expressions and scalar quantizers alone are needed, resulting in computational savings and faster adaptability.

- *Scalability:* The above discussions on complexity and learnability also point to the scalability aspects of the Lloyd, TSVQ, and PDFVQ algorithms. For the Lloyd algorithm alone, as the number of APs increases, the whole procedure must be repeated with increased complexity. In the case of TSVQ,

additional successive splitting steps must be performed, with the number of times that the Lloyd algorithm is implemented doubling at every step. In PDFVQ, the closed-form expressions that perform allocations do not change with the increase in APs. These allocations and scalar quantizers can be easily and quickly implemented, enabling an 'on-the-fly' placement. Thus, PDFVQ facilitates scalability, both with respect to dimensions (of the user distribution) as well as the number of APs.

Due to the abovementioned advantages, PDFVQ is thus suitable for AP placement both in an environment where the user distribution changes over time as well as the large-scale massive MIMO scenario with a very high number of APs. It is also worth noting that the PDFVQ procedure is independent of the chosen distortion measure for the scalar quantizers, but for simplicity, we will consider that the Lloyd algorithm is used.

PDFVQ Algorithm. The PDFVQ procedure from [100] introduced in the beginning of this section applied to AP placement (involving only two dimensions) is outlined in Algorithm 14 below. Note that in this paper, we model the user distributions as GMMs (Section 5.7), which avoids the need for parameter estimation. The total number of bits available b_{tot} is $\log_2 M$. It is important to note that in both step 1 and step 2, the resulting bit allocations b_l and $b_{l,j}$, respectively, are *not* expected to be integers. The number of levels computed in step 2 corresponding to the bit allocation $b_{l,j}$, that is, $V_{l,j} = 2^{b_{l,j}}$, is ultimately rounded off to the nearest integer, and represents the number of APs. As 'bit' generally refers to an integer quantity, to avoid subsequent confusion, we omit the word 'bit' when discussing allocations. The codebooks generated in step 3 for each cluster are the required AP locations.

Algorithm 14 PDF Optimized Vector Quantization Algorithm

1: Determine the allocation b_l to cluster l given the total budget b_{tot} using

$$2^{b_l} = 2^{b_{\text{tot}}} \frac{\sqrt{p_l c_l}}{\sum_{j=1}^{L} \sqrt{p_j c_j}}, \tag{4.20}$$

where $c_l = \sqrt{\lambda_{l,1}\lambda_{l,2}}$, $\boldsymbol{\lambda}_l = \text{diag}(\lambda_{l,1}, \lambda_{l,2})$, and $\boldsymbol{\Sigma}_l = \mathbf{Q}_l \boldsymbol{\lambda}_l \mathbf{Q}_l^T$ is the eigen value decomposition.

2: With each cluster l, compute the allocation along each dimension $b_{l,j}$, $j = 1, 2$, using

$$b_{l,j} = \frac{b_l}{2} + \frac{1}{2} \log \left[\frac{\lambda_{l,j}}{c_l} \right]. \tag{4.21}$$

Compute and round off the corresponding level $V_{l,j} = 2^{b_{l,j}}$.

3: Generate the codebook $\mathcal{R}_l$ for each cluster using

$$\mathcal{R}_l = \{\mathbf{q} | \mathbf{q} = \mathbf{Q}_l \mathbf{y} + \boldsymbol{\mu}_l, \mathbf{y} \in \mathcal{T}_l\}, \tag{4.22}$$

where $\mathcal{T}_l$ is the set of vectors given by the Cartesian product $\mathcal{T}_l = \mathcal{S}_{l,1} \times \mathcal{S}_{l,2}$, with $\mathcal{S}_{l,j}$, $j = 1, 2$ being the optimal $V_{l,j}$-level scalar quantizer of a univariate Gaussian with variance $\lambda_{l,j}$.

4.4.4 Gradient Approaches

As alluded to above, the VQ approaches do not explicitly solve the sum rate and minimum rate maximization problems, but provide a practical solution. However, if gradient ascent were to be applied to solve the problems so that either the sum rate or minimum rate can be improved further, the proper initialization of the AP locations is critical to avoid sub-par local optima. Hence, the VQ-based methods could provide reasonably good starting points to apply gradient ascent for both problems. Accordingly, we present both of the gradient calculations below, and are called the *max-sum* and *max-min gradient*, respectively.

Max-sum gradient

To maximize sum rate, the gradient update expression with j as the iteration index is

$$\mathbf{q}_m^{(j+1)} = \mathbf{q}_m^{(j)} + \delta \frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \sum_{k=1}^{K} \log \left(1 + \rho_r \sum_{m=1}^{M} \beta_{mk} \right) \right\}, \ \forall m, \tag{4.23}$$

where in the objective function from (4.14), we have neglected the shadow fading z_{mk} and replaced the expectation with the sum over the users. The gradient in (5.18) is calculated as

$$\frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \sum_{k=1}^{K} \log \left(1 + \rho_r \sum_{m=1}^{M} \beta_{mk} \right) \right\} = \frac{\gamma \rho_r}{2} \sum_{k=1}^{K} \frac{1}{1 + \psi_k^{\mathrm{ZF}}} \frac{(\mathbf{p}_k - \mathbf{q}_m^{(j)})}{||\mathbf{p}_k - \mathbf{q}_m^{(j)}||^{\gamma+2}}. \tag{4.24}$$

Max-min gradient

For minimum rate, taking the gradient of the rate of the worst user is fragile since the absolute value of the minimum rate can vary significantly across the iterations of gradient ascent causing convergence issues. Additionally, we are also interested in evaluating the performance of the cell-free system in terms of the 95%-likely rate, which quantifies the best rate of the worst 5% of the users. Accordingly, we consider the sum of rates corresponding to the worst 5% of the users, represented by $\mathcal{K}_{5\%}$, as

$$\mathbf{q}_m^{(j+1)} = \mathbf{q}_m^{(j)} + \delta \frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \sum_{k \in \mathcal{K}_{5\%}} \log \left(1 + \rho_r \sum_{m=1}^{M} \beta_{mk} \right) \right\}, \ \forall m, \tag{4.25}$$

which uses the same simplifications as in (5.18), and the gradient is

$$\frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \log \left(1 + \rho_r \sum_{m=1}^{M} \beta_{mk} \right) \right\} = \frac{\gamma \rho_r}{2} \sum_{k \in \mathcal{K}_{5\%}} \frac{1}{1 + \psi_k^{\mathrm{ZF}}} \frac{(\mathbf{p}_k - \mathbf{q}_m^{(j)})}{||\mathbf{p}_k - \mathbf{q}_m^{(j)}||^{\gamma+2}}. \tag{4.26}$$

4.5 Implementation and Evaluation of Limited Cooperation

Recall that in the system model, we had a combiner $\mathbf{v}_k$ chosen to estimate the data symbol of user k as

$$\hat{s}_k = \mathbf{v}_k^H \mathbf{y} = \sqrt{\rho_r} \mathbf{v}_k^H \mathbf{g}_k s_k + \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \mathbf{v}_k^H \mathbf{g}_{k'} s_{k'} + \mathbf{v}_k^H \mathbf{w}, \tag{4.27}$$

and the signal-to-noise-plus-interference ratio (SINR) of user k derived from (4.27) to be

$$\phi_k^{\mathbf{v}_k} = \frac{\rho_r \mathbf{v}_k^H \mathbf{g}_k \mathbf{g}_k^H \mathbf{v}_k}{\mathbf{v}_k^H \mathbf{v}_k + \sum\limits_{\substack{k'=1 \\ k' \neq k}}^{K} \rho_r \mathbf{v}_k^H \mathbf{g}_{k'} \mathbf{g}_{k'}^H \mathbf{v}_k}. \tag{4.28}$$

We now consider a limited AP cooperation scenario where the APs are grouped into N AP clusters (referred to simply "clusters" henceforth), where no AP is a member of two clusters. Each cluster has M_n APs where $n = 1, 2, \ldots, N$, such that $\sum\limits_{n=1}^{N} M_n = M$. In each cluster, K_n users communicate to the M_n APs such that $\sum\limits_{n=1}^{N} K_n = K$. A user is associated to the cluster to which its *geographically closest* AP belongs. Note that we will use k_n to index the user in cluster n such that $k_n = 1, 2, \ldots, K_n$. Further, each cluster has a local NC to which all APs within the cluster are connected to. This NC receives the received signals of its *constituent* APs in order to design appropriate combiners for data symbol estimation. In cluster n, the received signal at the M_n APs (expressed as a $M_n \times 1$ vector) is

$$\mathbf{y}_n = \sqrt{\rho_r}\mathbf{g}_{k_n} s_{k_n} + \underbrace{\sum_{\substack{k'_n=1 \\ k'_n \neq k_n}}^{K_n} \sqrt{\rho_r}\mathbf{g}_{k'_n} s_{k'_n}}_{\text{Intra-cluster interference}} + \underbrace{\sum_{\substack{n'=1 \\ n' \neq n}}^{N} \sum_{k_{n'}=1}^{K_{n'}} \sqrt{\rho_r}\mathbf{g}_{k_{n'}} s_{k_{n'}}}_{\text{Inter-cluster interference}} + \mathbf{w}_n, \tag{4.29}$$

where $\mathbf{g}_{k_n}$ is a $M_n \times 1$ vector containing the channel coefficients from the user k_n to the M_n APs and s_{k_n} represents the data symbol of user k_n. $\mathbf{w}_n$ is a $M_n \times 1$ additive noise vector. For simplicity, we can write the above equation as

$$\mathbf{y}_n = \sqrt{\rho_r}\mathbf{g}_{k_n} s_{k_n} + \mathbf{z}_{k_n}, \tag{4.30}$$

where $\mathbf{z}_{k_n}$ denotes the sum of intra-cluster interference, inter-cluster interference, and the noise, with respect to user k_n.

The linear minimum mean square error (LMMSE) combiner is assumed here, with the combiner $\mathbf{v}_{k_n}$ corresponding to user k_n defined as

$$\begin{aligned} \mathbf{v}_{k_n}^{\text{LMMSE}} &= \mathbf{C}_{\mathbf{y}_n}^{-1} \mathbf{C}_{s_{k_n}\mathbf{y}_n}^{H}, \\ &= \rho_r \mathbf{C}_{\mathbf{y}_n}^{-1} \mathbf{g}_{k_n}, \end{aligned} \tag{4.31}$$

where $\mathbf{C}_{\mathbf{y}_n}$ denotes the auto-covariance matrix of received signal $\mathbf{y}_n$ and $\mathbf{C}_{s_{k_n}\mathbf{y}_n}$ is the cross-covariance matrix of the data symbol s_{k_n} with the received signal $\mathbf{y}_n$. The per-user instantaneous SINR ϕ_{k_n} of user k_n can be calculated as follows. When the LMMSE combiner is applied to the received signal in (4.30), we obtain the following equation

$$(\mathbf{v}_{k_n}^{\text{LMMSE}})^H \mathbf{y}_n = \sqrt{\rho_r}(\mathbf{v}_{k_n}^{\text{LMMSE}})^H \mathbf{g}_{k_n} s_{k_n} + (\mathbf{v}_{k_n}^{\text{LMMSE}})^H \mathbf{z}_{k_n}. \tag{4.32}$$

From this equation, the signal power is

$$\begin{aligned} \rho_s &= \mathbb{E}_{s_{k_n}} \left\{ \left| \sqrt{\rho_r}(\mathbf{v}_{k_n}^{\text{LMMSE}})^H \mathbf{g}_{k_n} s_{k_n} \right|^2 \right\}, \\ &= \rho_r \mathbb{E}_{s_{k_n}} \left\{ \left((\mathbf{v}_{k_n}^{\text{LMMSE}})^H \mathbf{g}_{k_n} s_{k_n} \right)^H (\mathbf{v}_{k_n}^{\text{LMMSE}})^H \mathbf{g}_{k_n} s_{k_n} \right\}, \\ &= \rho_r \mathbf{g}_{k_n}^H \mathbf{v}_{k_n}^{\text{LMMSE}} (\mathbf{v}_{k_n}^{\text{LMMSE}})^H \mathbf{g}_{k_n}, \end{aligned} \tag{4.33}$$

and the interference-plus-noise power can be calculated as

$$\rho_{i+n} = \mathbf{g}_{k_n}^H \mathbf{C}_{\mathbf{z}_{k_n}} \mathbf{g}_{k_n}. \tag{4.34}$$

Thus, the SINR is

$$\begin{aligned} \phi_{k_n} &= \frac{\rho_s}{\rho_{i+n}}, \\ &= \frac{\rho_r^3 \left(\mathbf{g}_{k_n}^H \mathbf{C}_{\mathbf{y}_n}^{-1} \mathbf{g}_{k_n}\right)^2}{\mathbf{g}_{k_n}^H \mathbf{C}_{\mathbf{z}_{k_n}} \mathbf{g}_{k_n}}, \end{aligned} \tag{4.35}$$

where the second expression arises from applying (4.31) to (4.33).

It is simple to extend the above LMMSE combiner to the canonical cell-free network, i.e., with a single cluster. By writing (**??**) in terms of user k, we have

$$\mathbf{y} = \sqrt{\rho_r} \mathbf{g}_k s_k + \mathbf{z}_k, \tag{4.36}$$

where $\mathbf{z}_k = \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \mathbf{g}_{k'} s_{k'} + \mathbf{w}$ is the sum of the interfering signals from the other users and the noise. The LMMSE combiner in this case is

$$\begin{aligned} \mathbf{v}_k^{\text{LMMSE}} &= \mathbf{C}_{\mathbf{y}}^{-1} \mathbf{C}_{s_k \mathbf{y}}^H, \\ &= \rho_r \mathbf{C}_{\mathbf{y}}^{-1} \mathbf{g}_k, \end{aligned} \tag{4.37}$$

and the per-user SINR is

$$\phi_k = \frac{\rho_r^3 \left(\mathbf{g}_k^H \mathbf{C}_{\mathbf{y}}^{-1} \mathbf{g}_k\right)^2}{\mathbf{g}_k^H \mathbf{C}_{\mathbf{z}_k} \mathbf{g}_k}. \tag{4.38}$$

In this regard, we make the following two remarks:

- The LMMSE combiner assumed here is gene-aided since the vector of channel coefficients corresponding to user k is required to be known.
- In the canonical cell-free network, the LMMSE combiner requires that the NC collects the received signals $\mathbf{y}$ by all APs along with the vector of the channel coefficients $\mathbf{g}_k$ corresponding to user k. On the contrary, in the case of limited cooperation, only the APs in a cluster send their received signals and channel coefficients of user k to the local NC to determine the combiner, resulting in backhaul savings.

4.6 Simulation Methodology and Results

4.6.1 Simulation Parameters

In a geographical area of dimensions 2 km $\times$ 2 km, we consider $M = 32$ APs and $K = 4$ users, since $M \gg K$ for a cell-free system. For the purposes of placement, however, we use 2000 users that are distributed according to a GMM of the form in (5.27), with covariance $\mathbf{\Sigma}_l = \sigma_l^2 \mathbf{I}$, where σ_l is the standard deviation of mixture component l and $\mathbf{I}$ is the identity matrix. The GMM used in our simulations has parameters $L = 3$, $\boldsymbol{\mu}_1 = [0.5, -0.5]^T$, $\boldsymbol{\mu}_2 = [0, 0.5]^T$, $\boldsymbol{\mu}_3 = [-0.5, 0]^T$, $\sigma_1 = \sigma_2 = \sigma_3 = 100$, $p_1 = 0.6$, and $p_2 = p_3 = 0.2$. The

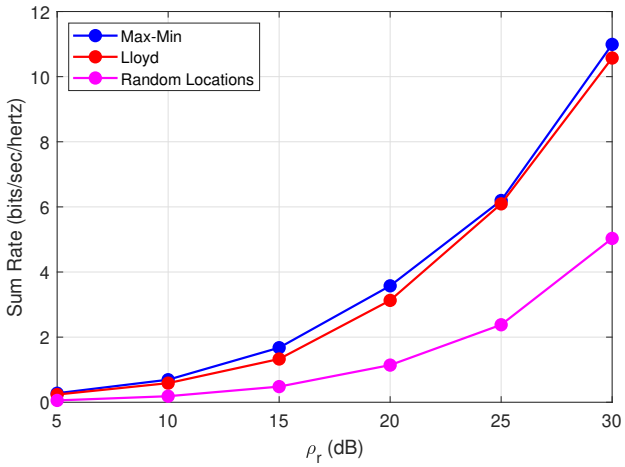

Figure 4.9: Sum rates versus transmit power ρ_r for the max-min and Lloyd algorithms with $M = 32$

pathloss model from [61, (4.34)] is used with shadow fading ignored for simplicity and the transmit power ρ_r of the users is increased from 5 to 30 dB. The max-sum and max-min gradient methods utilize step sizes of $\delta = 10^3$ and $\delta = 3 \times 10^4$, respectively.

4.6.2 Performance Measures

The per-user achievable rate is used, and is defined for user k as

$$R_k = \mathbb{E}\left\{\log_2\left(1 + \psi_k^{\mathrm{ZF}}\right)\right\}, \tag{4.39}$$

with ψ_k^{ZF} from (5.5) and the rate values are generated using Monte Carlo iterations. Further, algorithms are run multiple times and the solution that yields the best result is chosen. The maximum number of iterations for the Lloyd algorithm is set to 50. For comparison among the algorithms, both the sum rate and the 95%-likely rate measures are used. The relative performance between algorithms (say, algorithm 2 over algorithm 1) can be calculated by using the following measure expressed as percentage

$$\text{Improvement Ratio} = \frac{\mathcal{P}^{\text{Algorithm 2}} - \mathcal{P}^{\text{Algorithm 1}}}{\mathcal{P}^{\text{Algorithm 1}}} \times 100, \tag{4.40}$$

where $\mathcal{P}$ is either the sum rate or 95%-likely rate.

4.6.3 Numerical Results

Experiment 1. Here, we quantify the relative rate performances of the max-min [60] (with 2500 grid points) and Lloyd algorithms, and in Fig. 4.9, we plot the sum rates. It is observed here that both the max-min and Lloyd algorithms provide comparable performances. Fig. 4.10, on the other hand, plots the 95%-likely rate, which is the 5^{th} percentile of the CDF of the rate values, as a function of the transmit power, where the Lloyd algorithm shows only a slight improvement over the max-min algorithm.

We study briefly this minimum rate performance of the two algorithms. Although Fig. 4.10 shows that the Lloyd algorithm fares better in terms of the 95%-likely rate than the max-min algorithm, this

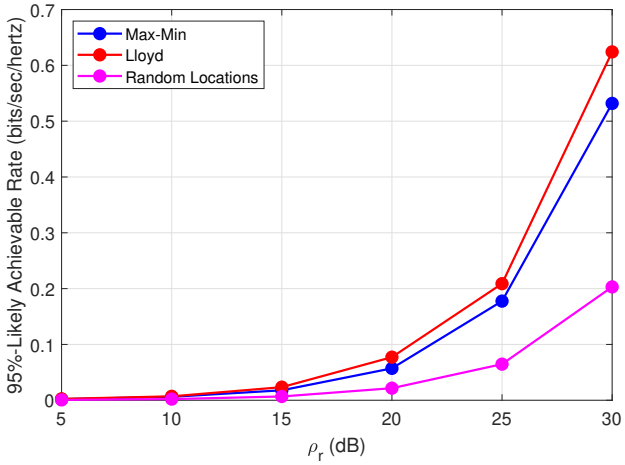

Figure 4.10: 95%-likely rates versus transmit power ρ_r for the max-min and Lloyd algorithms with $M = 32$

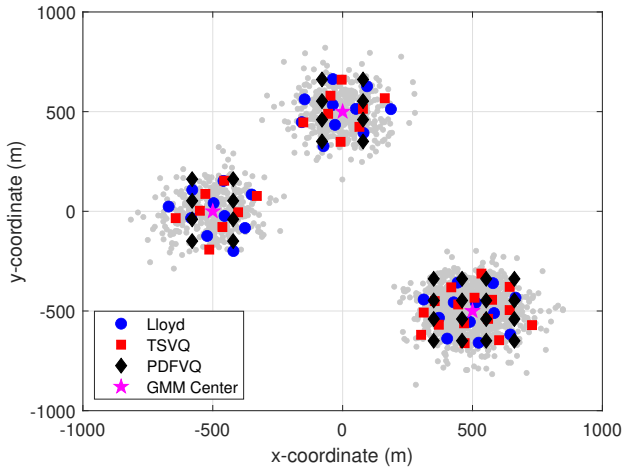

Figure 4.11: Final AP locations of the Lloyd algorithm, TSVQ, and PDFVQ at $\rho_r = 30$ dB.

Table 4.1: Rate Improvement of TSVQ and PDFVQ Relative to the Lloyd Algorithm at $\rho_r = 30$ dB

Algorithm	Sum Rate	95%-Likely Rate
TSVQ	3.82%	−7.98%
PDFVQ	2.78%	−0.82%

occurrence is due to the selection of the GMM configuration as well as the grid resolution (number of grid points). Accordingly, the max-min algorithm has superior performance to the Lloyd algorithm, when the grid resolution is high and the GMM is heavy-tailed, i.e., where there are some users further away from where the majority are (figure omitted due to space constraints). This occurs because the Lloyd algorithm places APs while minimizing both the Euclidean distances as well as the probability of cells. In contrast, the max-min algorithm always places APs such that the minimum rate of each user is satisfied. However, the Lloyd algorithm still provides a reasonable solution keeping in mind the implementation complexities related to the increased grid resolution in the max-min algorithm.

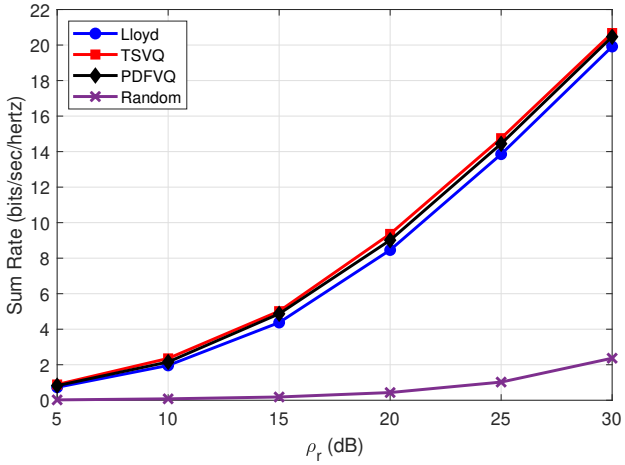

Figure 4.12: Sum rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random AP locations

.

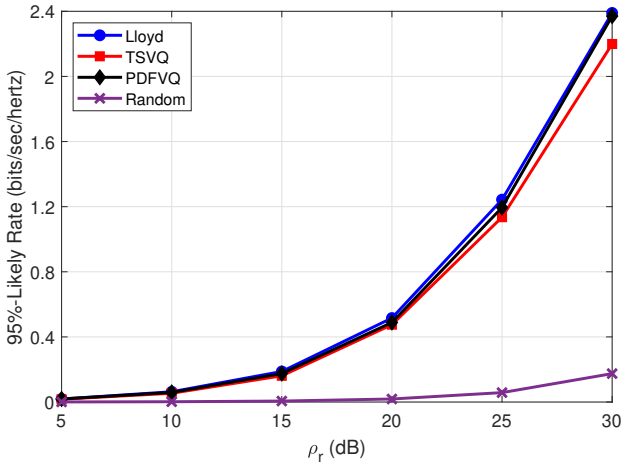

Figure 4.13: 95%-likely rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random AP locations

.

Experiment 2. We compare the performances of TSVQ and PDFVQ to the standard Lloyd algorithm. Their final AP locations and the distribution of the 2000 users (as gray circles) are shown in Fig. 4.11, where the Lloyd algorithm results in APs that are more distributed than TSVQ. This is expected since the TSVQ algorithm successively splits Voronoi regions and after completion does not necessarily associate each user to its closest AP. For PDFVQ, the allocation procedure within each cluster results in 3.85, 2.93, and 2.93 APs along each dimension, respectively, since the GMM clusters are symmetric along both x- and y-coordinates. Although rounding off these values would result in 4, 3, and 3 APs along each coordinate, the total number of APs would then be 34. Thus, to limit to $M = 32$ APs, we select 4, 2, and 2 APs along the x-coordinate and 4, 4, and 4 APs along the y-coordinate. This combination is selected as it provides the best result through repeated trials. The AP locations are observed to be more regular and grid-like due to the scalar quantizers used in the transform coding design, when compared to the standard Lloyd algorithm or TSVQ. The sum and 95%-likely rates of the three VQ-based algorithms are plotted in Fig. 4.12 and

Fig. 4.13, respectively. Additionally, for comparison, the performances when the APs are randomly placed are plotted (the AP locations themselves are omitted in Fig. 4.11 to avoid cluttering the figure), showing an expected low performance. The improvements in sum and 95%-likely rates, expressed as percentages, of TSVQ and PDFVQ over the Lloyd algorithm at transmit power $\rho_r = 30$ dB are noted in Table 4.1. It is seen that while the sum rate of TSVQ is greater than that of the Lloyd algorithm (by nearly 4%), the 95%-likely rate is worse (by nearly 8%). For PDFVQ, the sum rate offered is close to that of TSVQ while the 95%-likely rate approaches but does not equal that of the Lloyd algorithm (with the difference being less than 1%), due to the fact that the AP positions are far away from a small group of users in two mixture components. While for the user distribution considered PDFVQ offers a solution providing higher sum rate and similar 95%-likely rate to the Lloyd algorithm, it can be expected that this performance will reduce as the variance of the mixture components is increased.

Experiment 3. In Experiment 1, we compared the VQ approaches for a GMM consisting of Gaussian mixture components with spherical (proportionate to the identity matrix) covariance matrices. In order to show an example of PDFVQ applied to a full (non-diagonal) covariance matrix, we now consider a user distribution where the second component of the GMM considered is modified to have a covariance matrix of $\boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma^2 & 2\sigma^2/3 \\ 2\sigma^2/3 & 2\sigma^2 \end{bmatrix}$ with $\sigma = 100$ and the mean of the third component $\boldsymbol{\mu}_3 = [-0.5, -0.5]^T$. When PDFVQ is applied to this distribution, 4, 2, 3 APs along the x-coordinate and 4, 5, and 3 APs along the y-coordinate are allocated, totaling 35 APs. The best allocation for the desired 32 APs is found to be 4, 2, and 4, and 4, 4, and 2 APs along the x- and y-coordinates, respectively. The positions of such APs along with those of the Lloyd algorithm and TSVQ are shown in Fig. 4.14. The sum rate and 95%-likely rates corresponding to these locations are provided in Fig. 4.15 and Fig. 4.16, respectively, and the rate improvements are tabulated in Table 4.2. Again, the performances of the random AP locations are significantly worse than the VQ-based methods. Similar to the GMM with spherical Gaussians in Experiment 1, we observe that PDFVQ is able to match the sum rate performance of TSVQ. However, unlike above, the 95%-likely rate, like TSVQ, is lower than the Lloyd algorithm by nearly 13% since the space-filling advantage of the Lloyd algorithm is lost. Based on the above two experiments, it could be concluded that PDFVQ is a reasonable alternative to TSVQ that provides a similar or superior performance.

Although it is not the case here, we also note that the clusters of the GMM may be placed quite close to one another or may even overlap marginally. When PDFVQ is applied, the APs associated with the overlapping clusters may be closely placed and is inefficient. In such a scenario, starting from the cluster pair with most overlap (the degree of which can be computed by a measure such as overlap rate [102]), the AP allocation of the cluster with the higher 95%-likely rate could be decreased and re-allocated to another cluster with a poor performance. This process can be repeated for all overlapping clusters to improve the 95%-likely rate of the network.

Experiment 4. In this experiment, we use the gradient approaches outlined in Section 4.4.4 to improve the sum rate and 95%-likely rate performances shown in Experiment 2 for the GMM with a full covariance matrix. First, we show that using the max-sum gradient ascent, we can increase the sum rates of each of the Lloyd, TSVQ, and PDFVQ performances, as shown in Fig. 4.17. Note that in order to view all

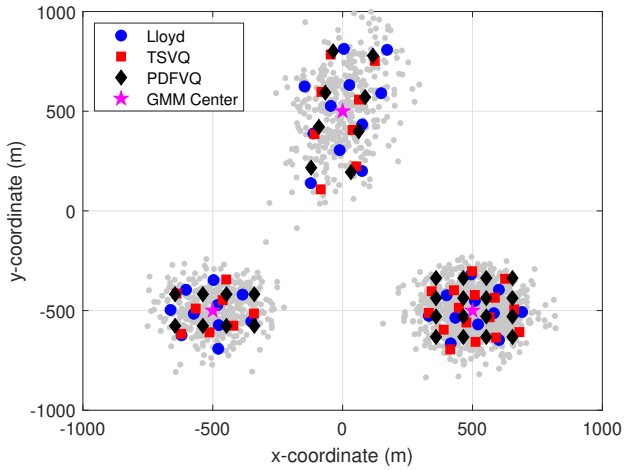

Figure 4.14: Final AP locations of the Lloyd algorithm, TSVQ, and PDFVQ with a full covariance matrix at $\rho_r = 30$ dB.

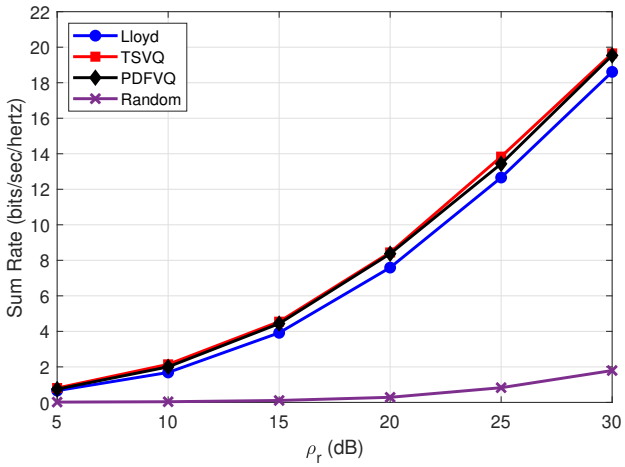

Figure 4.15: Sum rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random AP locations with a full covariance matrix.

Table 4.2: Rate Improvement of TSVQ and PDFVQ Relative to the Lloyd Algorithm With a Full Covariance Matrix at $\rho_r = 30$ dB

Algorithm	Sum Rate	95%-Likely Rate
TSVQ	5.59%	−13.67%
PDFVQ	4.98%	−12.85%

the curves better, we have included an additional figure that focuses on the power levels $\rho_r = 20$, 25, and 30 dB. For comparison with an existing method, we also plot the performance when the max-sum gradient ascent algorithm is applied to randomly initialized AP positions. This initialization considers that the APs are allocated to each GMM cluster according to PDFVQ and then positioned using the i.i.d. Gaussian distribution of the cluster. Table 4.3 lists the percentage rate improvements of the max-sum gradient applied

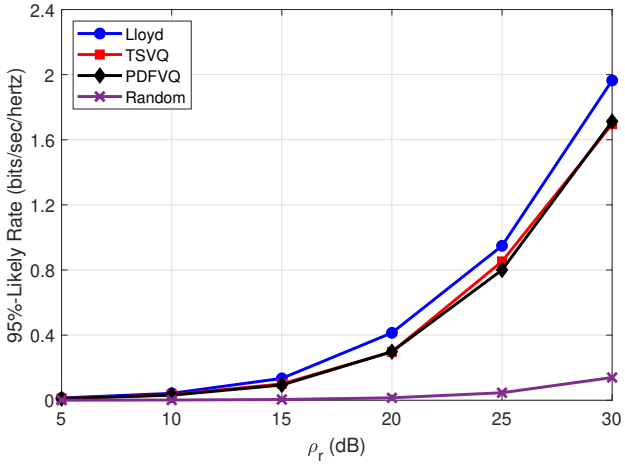

Figure 4.16: 95%-likely rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random AP locations with a full covariance matrix.

to each VQ approach, over the Lloyd algorithm alone. The gradient ascent applied to both the TSVQ and PDFVQ AP locations yields the highest sum rate (nearly 10% over the Lloyd algorithm) while the rate obtained when the ascent operation is applied with the Lloyd algorithm as the starting point is nearly the same as the sum rates of PDFVQ and TSVQ. This occurrence is due to the fact that the gradient ascent iterations with the Lloyd AP solutions as the initial points converge to a local optimum which is different from that obtained when the ascent is applied to PDFVQ or TSVQ. The 95%-likely rate performances shown in Fig. 4.18, when the max-sum gradient is applied to PDFVQ and TSVQ, do not change significantly (an increase is observed at $\rho_r = 30$ dB) and decrease when the gradient is applied to the Lloyd AP positions. Thus, in terms of sum rate, PDFVQ provides the best solution out of all VQ approaches and a further increase in sum rate (about 4% over PDFVQ) without negatively affecting the minimum rate performance (about 5% increase) is achieved by using the max-sum gradient ascent. For the randomly initialized case, it is observed that for a similar sum rate performance as the other techniques, the 95%-likely rate is remarkably inferior, thus showing the superiority of the VQ-initialized gradient approaches. Next, the max-min gradient ascent is applied and the 95%-likely rate is observed to increase as shown in Fig. 4.19. The best rate is obtained when the ascent algorithm is applied to the Lloyd solution as opposed to when it is applied to either TSVQ or PDFVQ where it is able to match the performance of the Lloyd algorithm. Table 4.4 informs us that while a 14% improvement in the 95%-likely rate is achieved by applying the gradient to the Lloyd solution, the difference of the gradient applied to TSVQ or PDFVQ from the Lloyd algorithm is only up to 4%. The sum rates corresponding to the 95%-likely rates are plotted in Fig. 4.20 (which also includes an additional zoomed in figure), where as a result of the 95%-likely rate improvement, the sum rate when the ascent is applied to the Lloyd algorithm is the least value. In contrast, despite the increase in 95%-likely rate, the sum rate performances when ascent is applied to TSVQ and PDFVQ are nearly the same as TSVQ and PDFVQ itself. Hence, with its simpler design, PDFVQ along with max-min gradient offers a good tradeoff between a 95%-likely rate similar to and a sum rate higher (by over 5%) than the Lloyd algorithm. It should, however, be mentioned that if 95%-likely rate (or minimum rate) is the sole performance measure

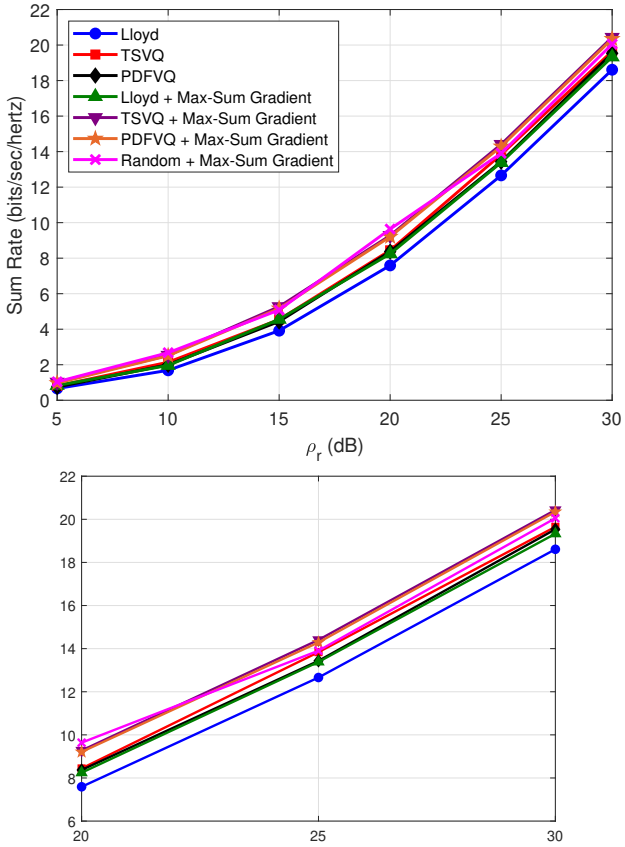

Figure 4.17: Sum rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random locations with a full covariance matrix along with max-sum gradient. The figure on the bottom zooms in on the sum rates for $\rho_r = 20$, 25, and 30 dB.

Table 4.3: Rate Improvement of the VQ Approaches with the Max-Sum Gradient Relative to the Lloyd Algorithm at $\rho_r = 30$ dB

Algorithm	Sum Rate	95%-Likely Rate
Lloyd + Max-Sum Gradient	3.88%	−5.60%
TSVQ + Max-Sum Gradient	9.80%	−7.52%
PDFVQ + Max-Sum Gradient	9.34%	−8.66%

of interest, the Lloyd algorithm alone is a straightforward choice. The gradient approach requires the choice of an appropriate step size for convergence while the Lloyd algorithm is known to converge [103]. Finally, as we observed for the max-sum case, while a similar sum rate is observed when the max-min gradient is applied to randomly initialized AP locations, the 95%-likely rate is significantly inferior to the other situations.

Experiment 5. In this experiment, we quantify the effect of a time-varying user density on the network performance and show the need for an easily adaptable AP placement algorithm. We consider a simple situation where the users are *initially* distributed as a single-cluster GMM ($L = 1$ in (5.27)) with

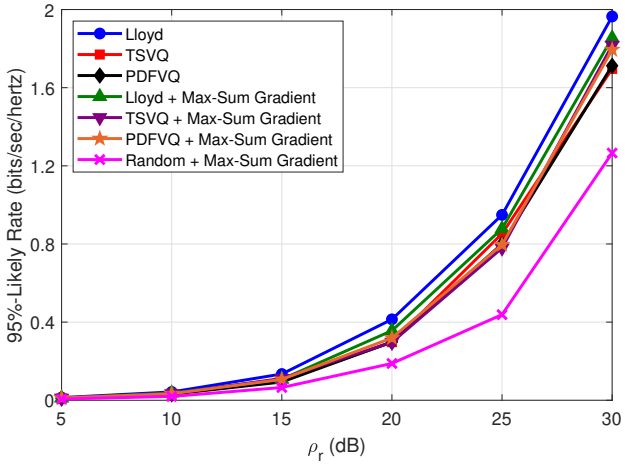

Figure 4.18: 95%-likely rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random locations with a full covariance matrix along with max-sum gradient.

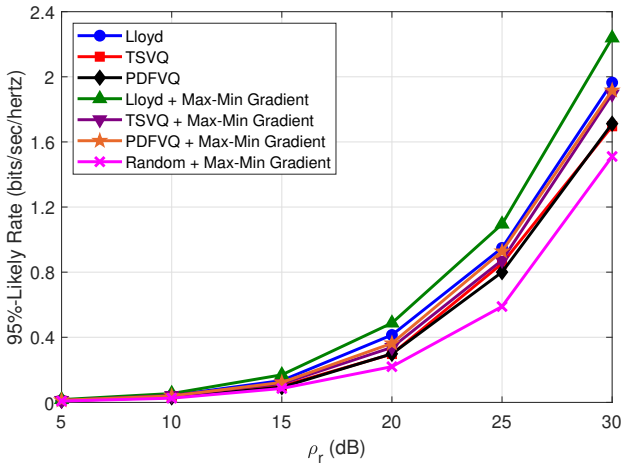

Figure 4.19: 95%-likely rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random locations with a full covariance matrix along with max-min gradient.

Table 4.4: Rate Improvement of the VQ Approaches with the Max-Min Gradient Relative to the Lloyd Algorithm at $\rho_r = 30$ dB

Algorithm	95%-Likely Rate	Sum Rate
Lloyd + Max-Min Gradient	13.93%	−3.73%
TSVQ + Max-Min Gradient	−3.60%	5.73%
PDFVQ + Max-Min Gradient	−2.48%	5.30%

mean $\boldsymbol{\mu} = [0, 0]^T$ and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2/3 \\ \sigma^2/3 & \sigma^2/2 \end{bmatrix}$ where $\sigma = 200$, and call it density A. Assuming a total of $M = 18$ APs, we calculate the PDFVQ AP locations for this user density. Over time, we consider that the user distribution changes to density B, where the users are more spread out along another direction to that of density A, with covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2/2 & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 \end{bmatrix}$ and $\sigma = 300$. These two user densities and the PDFVQ

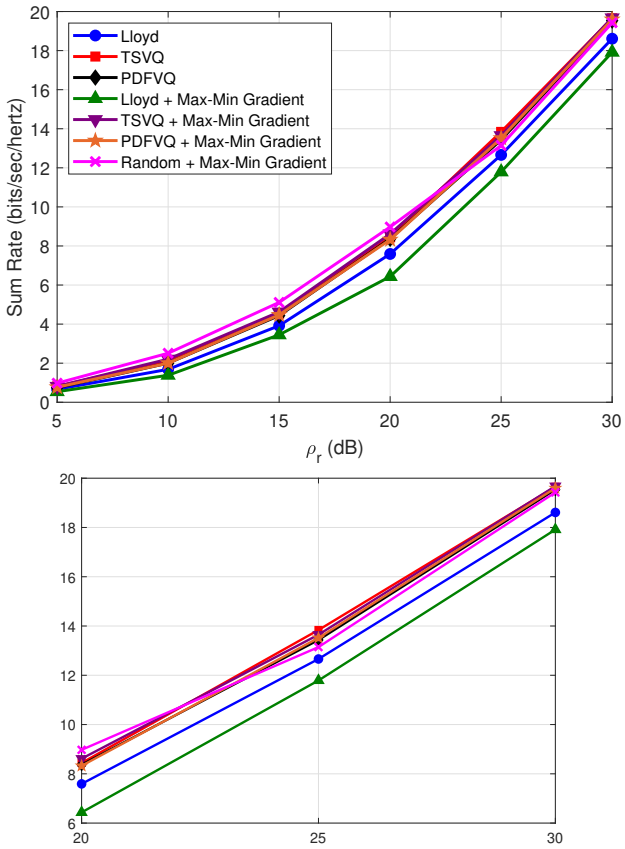

Figure 4.20: Sum rate as a function of ρ_r for the Lloyd algorithm, TSVQ, PDFVQ, and random locations with a full covariance matrix along with max-min gradient. The figure on the bottom zooms in on the sum rates for $\rho_r = 20$, 25, and 30 dB.

AP locations determined for density A are shown in Fig. 4.21. We evaluate the sum rate and 95%-likely rate performances when the AP locations are matched to density A and consider both user densities A and B, so that the loss due to mismatch of the AP locations to the user density B is also shown. For completeness, we also compute the rate values when PDFVQ AP locations are determined and matched to density B. As expected, in both the sum rate plotted in Fig. 4.22 and the 95%-likely rate plotted in Fig. 4.23, there is a significant loss in performance for the PDFVQ AP locations which are matched to density A, when the users re-position to density B. The relative losses in the sum rate and 95%-likely rate at $\rho_r = 30$ dB are 32.67% and 89.54%, respectively. To prevent this performance decrease, there is a need to re-calculate the AP locations for the new user density. Clearly, when the AP locations are matched via PDFVQ to the user density B, the rate performances in Fig. 4.22 and Fig. 4.23 are improved over the diminished performance of the mismatched case. With easy adaptability and the other aforementioned advantages, PDFVQ offers the best method for cell-free AP placement among the VQ techniques discussed in this paper. In future work, we will address how the changing user densities can be learned for use in PDFVQ.

Experiment 6. Here, the relative performance between PDFVQ and the Lloyd algorithm when the number of APs is increased as in a massive MIMO system is shown. Three values of number of APs M

(a) Density A. (b) Density B.

Figure 4.21: The two user densities A and B considered for Experiment 4 along the PDFVQ AP locations matched to density A.

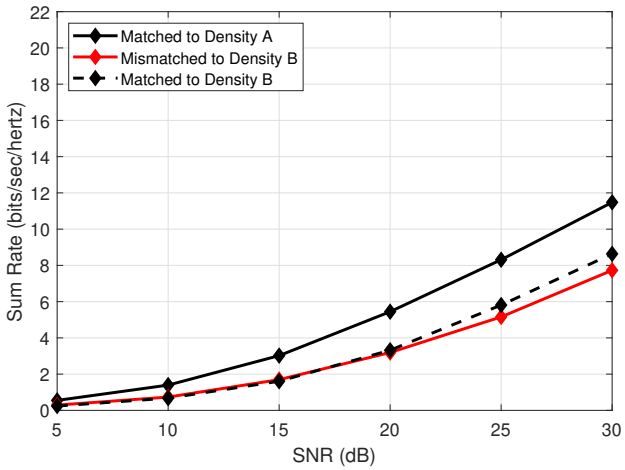

Figure 4.22: Sum rate as a function of ρ_r for Experiment 4.

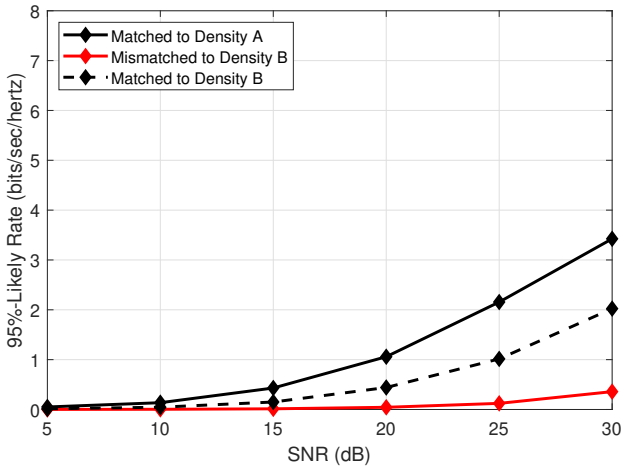

Figure 4.23: 95%-likely rate as a function of ρ_r for Experiment 4.

used are 64, 256, and 1024, corresponding to which number of users K are 8, 32, and 128, respectively, since $M \gg K$ for a cell-free system. We assume a total of 5000 users, distributed as a GMM of the form in (5.27) with parameters $L = 4$, $\boldsymbol{\mu}_1 = [0.5, -0.5]^T$, $\boldsymbol{\mu}_2 = [0.5, 0.5]^T$, $\boldsymbol{\mu}_3 = [-0.5, 0.5]^T$, $\boldsymbol{\mu}_4 = [-0.5, -0.5]^T$, $p_1 = p_2 = 0.35$, and $p_3 = p_4 = 0.15$. The covariance matrices of the clusters are defined as $\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$,

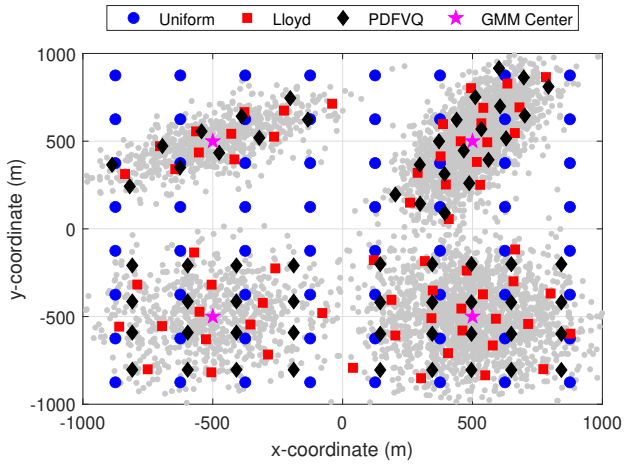

Figure 4.24: User distribution and AP locations of the Lloyd and PDFVQ algorithms when $M = 64$ and $\rho_r = 30$ dB.

Table 4.5: Allocations along the x- and y-coordinates, and per cluster generated by the PDFVQ algorithm and the number of APs per cluster generated by the Lloyd algorithm when $M = 64$, 256, and 1024

Cluster Index	$M = 64$				$M = 256$				$M = 1024$			
	PDFVQ			Lloyd	PDFVQ			Lloyd	PDFVQ			Lloyd
	x	y	Total	Total	x	y	Total	Total	x	y	Total	Total
1	5	4	20	22	10	10	100	87	20	19	380	359
2	3	6	18	17	5	12	60	63	10	25	250	271
3	2	5	10	11	4	10	40	51	8	21	168	166
4	4	4	16	14	7	8	56	55	15	15	225	228

$\mathbf{\Sigma}_2 = \begin{bmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & 4\sigma^2 \end{bmatrix}$, $\mathbf{\Sigma}_3 = \begin{bmatrix} 4\sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix}$, and $\mathbf{\Sigma}_4 = \mathbf{\Sigma}_1$, with $\sigma = 200$. Note that while measuring the performance, the sum rate is normalized by the number of users so that fair comparison may be made across the different number of users corresponding to the three AP levels considered. Additionally, we compare the two VQ approaches against a uniform placement where the APs are located at regular grid points and distributed throughout the geographical area. The user distribution, the uniformly placed APs, and the placements achieved by the Lloyd and PDFVQ algorithms for $M = 64$ APs and at $\rho_r = 30$ dB are shown in Fig. 4.24 (the AP locations for other AP levels are omitted for lack of space). The allocations along each coordinate are given in Table 4.5, which also notes the per cluster number of APs for both VQ based algorithms.

The sum rate performances of the VQ based algorithms and when the APs are uniformly placed as a function of the user transmit power ρ_r for $M = 64$, 256, and 1024, are shown in Fig. 4.25, where for each M value, the Lloyd algorithm shows superior performance to PDFVQ. Clearly, the uniformly distributed APs perform the worst since their placement is not congruent with the density of the users. Also, as expected, when M is increased, the sum rate performance for all three AP locations is improved. The PDFVQ APs (as observed in Fig. 4.24), are more spread out across the clusters than the Lloyd algorithm. While both VQ

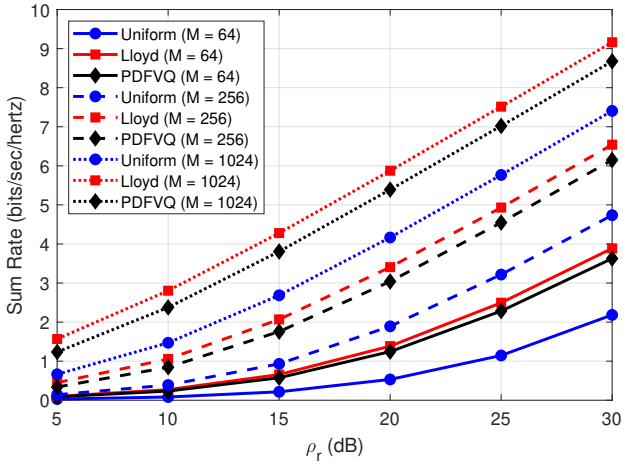

Figure 4.25: Sum rate as a function of ρ_r for the Lloyd and PDFVQ algorithms when $M = 64$, 256, and 1024.

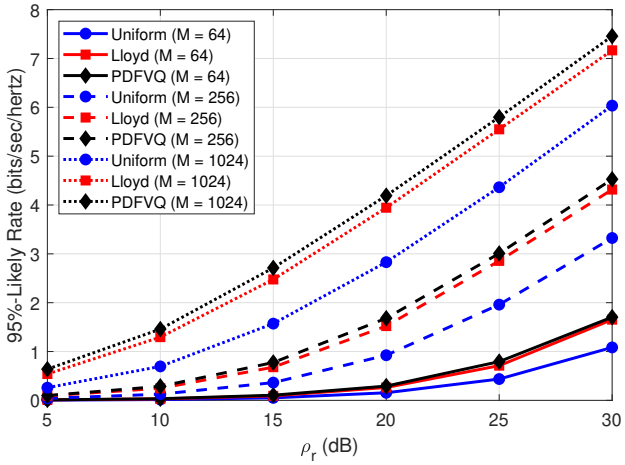

Figure 4.26: 95%-likely rate as a function of ρ_r for the Lloyd and PDFVQ algorithms when $M = 64$, 256, and 1024.

approaches do place more APs in areas of higher user density (which is at and around the cluster centers), the Lloyd algorithm is able to take advantage of both dimensions simultaneously, while the PDFVQ algorithm considers them separately. As a result, the Lloyd algorithm places more APs nearer the cluster centers and achieves a higher sum rate. The spread of the PDFVQ solutions give it an advantage as it is able to serve the users at the edge of the clusters, more so than the Lloyd algorithm. This translates to a higher 95%-likely rate performance for the PDFVQ algorithm. As the number of APs M is raised, 95%-likely rate as well as the difference in 95%-likely rate between the PDFVQ and Lloyd algorithm increases. The percentage changes of the sum rate and 95%-likely rate performances of the PDFVQ algorithm relative to the Lloyd algorithm at $\rho_r = 30$ dB are shown in Table 4.6. For the GMM considered, we observe that a nearly 5% improvement in the 95%-likely rate is achieved at the cost of a 6% decrease in the sum rate, for $M = 256$. Thus, in summary, while PDFVQ performs inferior to the Lloyd algorithm (in terms of sum or average rate) and is not optimal for both sum and 95%-likely rate, the low complexity and easy scalability justify the sub-optimality.

Table 4.6: Sum rate and 95%-likely rate improvements of the PDFVQ algorithm over the Lloyd algorithm at $\rho_r = 30$ dB when $M = 64$, 256, and 1024 expressed as percentage

Rate Measure	$M = 64$	$M = 256$	$M = 1024$
Sum Rate	-6.68%	-5.98%	-5.32%
95%-Likely Rate	3.07%	4.91%	4.06%

Table 4.7: Rate Improvement of the Three-Cluster Scenario Relative to the Single Cluster Scenario for the Lloyd and PDFVQ AP Locations at $\rho_r = 30$ dB and under GMM-1

Algorithm	Sum Rate	95%-Likely Rate
Lloyd	-6.48%	-13.15%
PDFVQ	-7.12%	-17.61%

Experiment 7. In this experiment, we show the results of the limited cooperation set-up. While the number of APs, number of users, and the GMM user distribution follow the prior experiments (except Experiment 6), we consider an additional GMM user density. To distinguish between them, we refer to the former as GMM-1 and the latter as GMM-2. GMM-2 has the same parameters as GMM-1, but with the means $\boldsymbol{\mu}_1 = [0.325, -0.325]^T$, $\boldsymbol{\mu}_2 = [0, 0.325]^T$, and $\boldsymbol{\mu}_3 = [-0.325, -0.325]^T$. Note that the parameters of GMM-1 and GMM-2 are chosen to model sparse and dense AP deployments, respectively.

For both the Lloyd and PDFVQ AP locations designed for GMM-1, we evaluate and compare the cases where the clustering is considered and when a single cluster (representing the canonical cell-free system) is implemented. Fig. 4.27 shows the instance where the PDFVQ AP locations are clustered according to the three mixture components. The sum rate and 95%-likely rate curves for the abovementioned cases are shown in Fig. 4.28 and Fig. 4.29, respectively. The performance for random AP locations without clustering is also included in the figures, clearly inferior to that of the others. In terms of sum rate, the Lloyd and PDFVQ solutions with clustering show similar performances (PDFVQ shows only a 2.38% lower performance than the Lloyd AP locations at $\rho_r = 30$ dB). The 95%-likely rate performances show that the Lloyd algorithm performs better than the PDFVQ solution (which shows a 13.61% lower performance than the Lloyd solution at $\rho_r = 30$ dB). More importantly, in both the sum rate and 95%-likely rate plots, the single cluster scenario is found to possess significantly higher rates for the Lloyd and PDFVQ solutions when compared to the three-cluster scenario. In other words, limited cooperation brings about a performance loss with respect to the canonical cell-free scenario. The differences in performance, expressed as percentage changes of the limited cooperation scenario over the canonical cell-free scenario, are tabulated in Table 4.7.

We now consider the denser GMM-2 for the user distribution, as shown in Fig. 4.30, also displaying the PDFVQ AP locations and clusters. As in the case of GMM-1, the rate performances of the Lloyd and PDFVQ solutions with three clusters and a single cluster consideration are plotted in Fig. 4.31 and Fig. 4.32. The relative performances of the curves remain the same as observed in GMM-1, however, the performance loss when the three-cluster scenario is used over the single cluster scenario is larger here. This is evident in

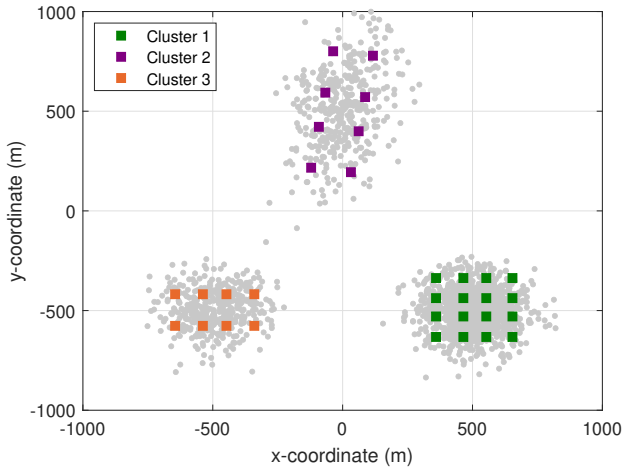

Figure 4.27: User configuration and the AP locations of the PDFVQ algorithm showing AP clustering under GMM-1.

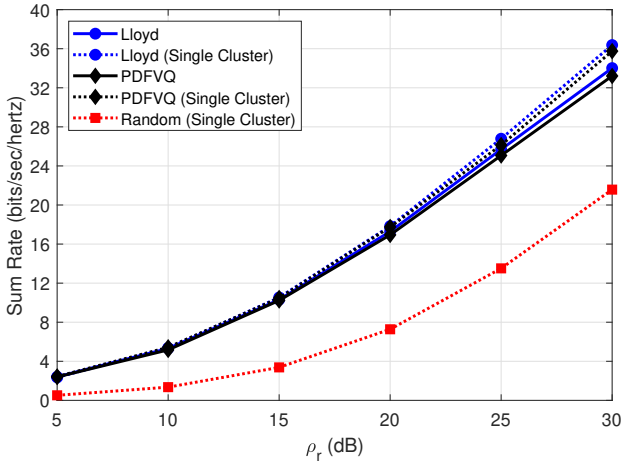

Figure 4.28: Sum rate as a function of ρ_r for the Lloyd algorithm and PDFVQ with clustering and when a single cluster is considered under GMM-1.

Table 4.8: Rate Improvement of the Three-Cluster Scenario Relative to the Single Cluster Scenario for the Lloyd and PDFVQ AP Locations at $\rho_r = 30$ dB and under GMM-2

Algorithm	Sum Rate	95%-Likely Rate
Lloyd	-8.75%	-17.88%
PDFVQ	-9.99%	-28.62%

Table 4.8 which shows these losses as a percentage. The increase in losses can be attributed to the fact that the inter-cluster interferences are larger in the case of GMM-2 than GMM-1 (since the clusters are closer together).

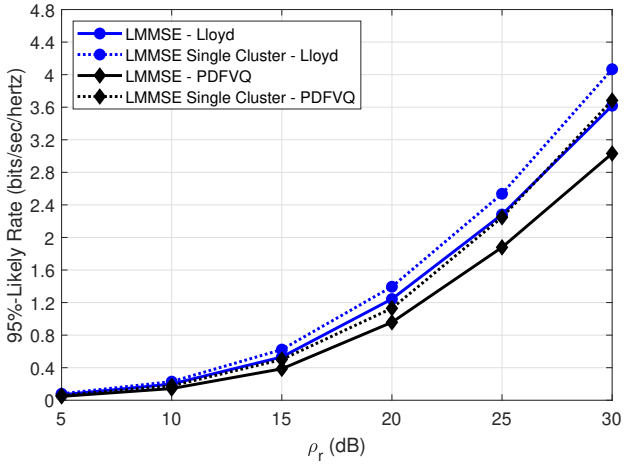

Figure 4.29: 95%-likely rate as a function of ρ_r for the Lloyd algorithm and PDFVQ with clustering and when a single cluster is considered under GMM-1.

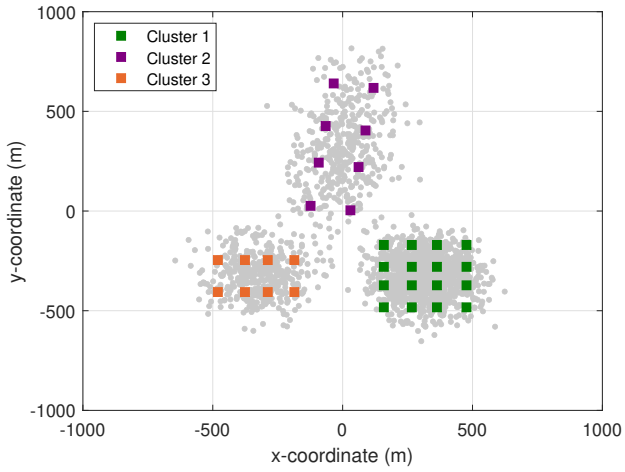

Figure 4.30: User configuration and the AP locations of the PDFVQ algorithm showing AP clustering under GMM-2.

4.7 Conclusion

In this paper, we have addressed access point (AP) placement in cell-free massive multiple-input-multiple-output (MIMO) systems under a throughput criteria. We investigated the two main optimization problems in this regard, namely the sum rate and minimum rate maximization problems. To understand their solution frameworks, simple examples were constructed and analyzed exposing the difficulty in solving the problems. Therefore, as a practical approach, the use of vector quantization (VQ)-based methods, namely the popular Lloyd algorithm, tree-structured VQ (TSVQ), and probability density function optimized VQ (PDFVQ), to cell-free AP placement, was investigated. Among the three algorithms presented, although the tree-structured VQ (TSVQ) provides better sum rate (as it fosters cooperation among APs by placing them closer) compared to the Lloyd algorithm, it suffers from high complexity, poor scalability, and the inability to easily adapt to new environments. PDFVQ, which overcomes the aforementioned shortcomings, allowed a more efficient generation of the codebook and generated a sum rate similar to and 95%-likely rate higher than

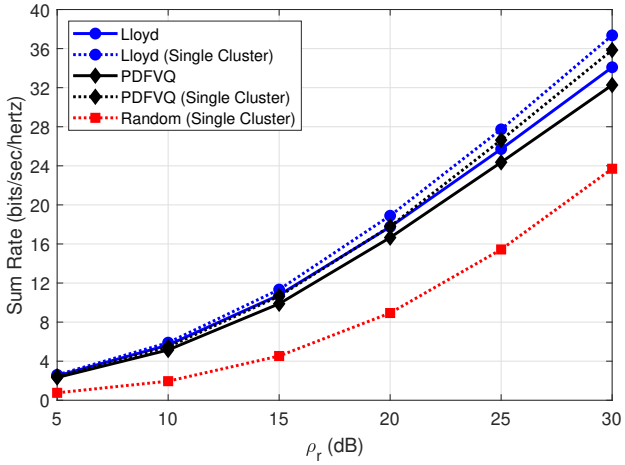

Figure 4.31: Sum rate as a function of ρ_r for the Lloyd algorithm and PDFVQ with clustering and when a single cluster is considered under GMM-2.

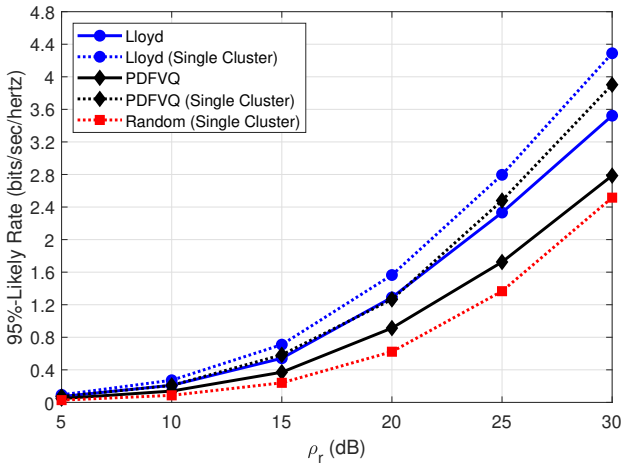

Figure 4.32: 95%-likely rate as a function of ρ_r for the Lloyd algorithm and PDFVQ with clustering and when a single cluster is considered under GMM-2.

TSVQ and close to the Lloyd algorithm. Additionally, for gradient-based maximization methods, PDFVQ is found to provide good initial points. It was observed numerically that, over the Lloyd algorithm, an increase of 9% in sum rate and a difference of just 2.5% in the 95%-likely rate was achieved by applying max-sum and max-min gradient ascent algorithms, respectively, with the PDFVQ AP locations as starting points. Thus, PDFVQ offers a convenient, less computationally intensive, and easily scalable AP placement technique for cell-free networks.

Chapter 4, in part, is a reprint with permission of the material as it appears in the paper: Govind Ravikumar Gopal and Bhaskar D. Rao, "Vector quantization methods for access point placement in cell-free massive MIMO systems," which is under review in *IEEE Transactions on Wireless Communications*. The dissertation author was the primary investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-2124929 and Grant CCF-2225617, in part

by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

Chapter 5

Practical Cell-Free AP Deployment

In this chapter, we address uplink cell-free distributed massive multiple-input-multiple-output (MIMO) access point (AP) deployment from the perspective of maximizing throughput. Different from the previous chapter, we develop a multi-step AP deployment procedure that incorporates several important and practical features, namely user position determination, limited AP cooperation, and fine-tuning of the AP positions. Starting from an initial deployment, an AP clustering step is proposed as the first step so that the user positions can be determined. To reduce the backhaul requirements in cell-free systems, we suggest agglomeration clustering to support limited cooperation among the APs followed by multilateration to calculate the user locations. Then, an AP placement algorithm is employed that utilizes the computed user positions. Finally, in order to fine-tune individual AP locations for both sum rate and minimum rate improvement (the two optimization problems of interest), we present the local max-sum and local majorization-minimization methods for sum rate improvement, and the local max-min method for minimum rate improvement. The entire multi-step procedure addresses major challenges associated with cell-free AP deployment. Over the reasonably good sum and minimum rate performances of the main AP placement step, the fine-tuning procedure is numerically shown to improve the appropriate rate values.

5.1 Introduction

Massive multiple-input-multiple-output (MIMO) systems have become increasingly prominent in the recent decade as they empower the current 5G and the upcoming Beyond 5G networks to meet their network requirements, both in terms of throughput and user experience. A large number of antennas allow MIMO systems to exhibit higher spectral and energy efficiencies, along with reduced interference through beamforming [7, 8]. Distributed antenna systems (DAS), particularly distributed MIMO, are advantageous over colocated systems as they enable higher average rates [10, 16–18]. Such distributed systems can be generally classified as non-cooperative and cooperative systems, with traditional small-cell systems constituting the former and the more lately popular cell-free systems comprising the latter.

Cell-free systems are networks where a number of users are all served by a large number of antennas

or access points (APs) [12, 104]. The "cell-free" paradigm originates from the concept of network MIMO [105], and by not splitting users into cells, higher spectral efficiency and interference mitigation are achieved in such systems compared to their small-cell counterparts. Cell-free systems have not yet been deployed, largely due to the scalability issues related to the backhaul necessary for information exchange between APs and the network controller (NC) or the central processing unit [22, 106]. Therefore, small-cell systems are continued to be used in current network deployments. However, cell-free systems show a promising outlook for near future network deployments [107]. Consequently, numerous problems and challenges in cell-free networks have been investigated. Previous works have included studies on power optimization and energy efficiency [20, 26–28], rate maximization [29, 30], clustering (user- and cell-centric) [31–35], limited fronthaul [36], pilot assignment [37], reconfigurable intelligent surfaces [38], and federated learning [39].

In this chapter, our focus is on AP deployment in cell-free systems, of which AP placement forms an important part. Optimization of the AP locations in terms of throughput is the key consideration. AP placement is useful in situations such as conference venues and stadiums where the user demands change over time. It also finds use in emergency scenarios, especially in natural (or man-made) disaster situations where terrestrial infrastructure has been knocked down and AP-enabled unmanned aerial vehicles (UAVs) must establish a temporary network [108]. In practice, the distribution of users (user density) can provide a good measure of user demand and is thus used in designing placement problems. It should be noted that different realizations of the user positions can generate the same user density. Past research focusing on AP placement has primarily been in the small-cell paradigm (e.g., [80] and references therein). In the cell-free regime, AP placement is relatively new and there has been limited prior work conducted (for instance, [60, 96–98]). For descriptions of previous placement works in small-cell and cell-free systems, refer to [80] and [109], respectively, or the previous chapters. In addition to AP placement, two major existing challenges are the determination of the user positions or of their density and the viability of limited cooperation among the APs. For example, the vector quantization (VQ) based methods described in [109] (and in the previous chapter) and the gradient-based method in [96] both require the knowledge of the user positions to implement the algorithms. Further, the other VQ methods described for small-cell placement in [80] (an in chapters 2 and 3) also require the locations of the users. Hence, it is essential to calculate the positions of the users. Additionally, the reduction of backhaul forms an important component of scalable cell-free AP deployment. Thus, limited cooperation among the APs serves as a method to reduce the information exchange among the APs.

The position or density determination of users has been covered in past works under the class of problems concerning tracking users [110, 111] and crowd density estimation, which uses wireless signals [112–115] or images [116, 117] alone, or even jointly [118, 119]. For example, the work of [119] looks at vision-aided transmitter identification, where wireless data (receive power) is assisted by images from a camera, using deep learning. The most popular avenue for position determination has been multilateration in the field of robotics, e.g. [120, 121] and references therein, where a set of three (trilateration) or more measured distances to known landmarks or reference points are used to localize the robot. This technique as been also implemented in the wireless domain which uses measured RSS to locate mobile users [122–124].

While [122] utilizes RSS measurements and a method based on extreme value theory, Taylor approximation, and Bayesian filtering to determine indoor node positions, [123] proposes a simple geometric approach to trilateration while also concurrently estimating the pathloss exponent. Their subsequent work [124] extends the geometric interpretation to the 3D case. Further, to improve the localization results, neural networks (NNs) are used to provide estimates of the pathloss exponent and standard deviations of distances between the reference points and the mobile device. Similarly, limited cooperation has been previously studied in various formats, including user-centric and cell-centric networks [31–34], dynamic clustering [35, 125, 126], and in the context of resource allocation for interference mitigation between APs [127]. For instance, [33] deals with data transmission and power control for scalable cell-free systems by considering cell-centricity for cooperation between APs and user-centricity for serving users. In [125], the authors also address scalability, however, through dynamic cooperative clustering for initial access, pilot assignment, downlink precoding, and uplink combining. Virtual cells utilizing hierarchical clustering are formed in the work of [127] with a view to create two types of limited cooperation models aiming to increase system sum rate.

The cell-free AP deployment problem has been gaining popularity only in the recent years. Among the existing placement techniques (discussed in [60, 96–98]), the user positions are required to be known, however, their determination is not covered by the aforementioned works. Moreover, we point out that all APs must cooperate fully for AP placement, i.e., the NC receives the positions of all users and determines AP locations. While total cooperation is necessary for placement, partial or limited cooperation, also not investigated in the context of AP placement, can allow for user position determination and reduce the information exchange between APs. For example, we propose that by first assuming that the users (whose positions are initially unknown) are associated with their closest APs, the APs can be grouped into non-overlapping clusters allowing for limited cooperation. Following this step, the within-cluster APs share information only among each other to determine the user positions associated with the cluster. This can be performed at designated cluster heads, thereby reducing backhaul requirements. Once the user positions are determined, the APs can then send the information to the NC for AP placement. Further, with this user position knowledge, we can allow the APs to individually update or fine-tune their positions according to network rate requirements. Thus, in this work, our focus is not on the placement of APs alone, but also on the abovementioned challenges associated with it. Consequently, we propose a multi-step cell-free AP deployment scheme that starts with unknown user positions (the AP positions are always assumed to be available), forms cooperative AP clusters to determine the user positions, performs AP placement, and finally fine-tunes the AP positions to achieve a higher rate performance. At each stage of this scheme, a varying degree of cooperation (or in other words, distributedness) is assumed. It should be noted that distributed algorithms, apart from requiring less information exchange between APs, helps in system scalability. *We note that in this chapter, we will use the word 'deployment' to refer to the multi-step scheme while the word 'placement' shall refer to the AP location determination step within the multi-step deployment scheme.*

Contributions

To the best of our knowledge, a solution for the cell-free AP deployment problem in the context of throughput optimality while addressing user position determination, limited cooperation, and fine-tuning of the AP positions, has not been provided in literature. While making the assumption that there exists an initial or existing deployment, in this work, our contributions are as follows.

- As the **first** step to the deployment scheme, an AP clustering method based on agglomeration clustering is presented that groups APs into non-overlapping clusters. The APs within each cluster share the the large-scale fading coefficients of the users with each other (a limited cooperation scenario) in the **second** step of the scheme so that through multilateration, the positions of the users associated with the cluster are determined.
- To place APs in the **third** step, two main throughput optimizations, namely the sum rate and minimum rate maximization problems, are considered. Here, out of the many placement solutions, we consider the Lloyd algorithm or the standard VQ method.
- In the **fourth** and final step, a fine-tuning procedure is proposed. An iterative process in which a subset of APs is first selected and whose positions are then updated is designed. AP position update procedures to improve the system sum rate, which are the local max-sum and local majorization-minimization (MM) methods, are conducted by the APs individually. To improve the minimum rate of the system, the local max-min method is also presented.

The remainder of this chapter is organized as follows. Section 5.2 outlines the cell-free model used throughout the paper. A brief description of the end-to-end procedure for deployment is provided in Section 5.3. Section 5.4 then describes the AP clustering, followed by user position determination utilizing multilateration in Section 5.5. In Section 5.6, the fine-tuning of the AP positions to increment system throughput for two placement problems are elucidated. Numerical simulation results are shown and described in Section 5.7. Finally, we provide concluding remarks in Section 5.8. Throughout this chapter, we use bold symbols to denote vectors, $\mathbb{E}\{\cdot\}$ is the expectation operator, $||\cdot||$ represents the ℓ_2-norm of a vector, and all logarithms are to the base 2.

5.2 System Model

The system model considered in this chapter is based on the previous chapter. We describe it here briefly for completeness. In an area, K single-antenna users are distributed with the probability density function $f_{\mathbf{P}}(\mathbf{p})$, where random vector $\mathbf{p} \in \mathbb{R}^2$ denotes the user position. The users are served by M single-antenna APs, where $\mathbf{q} \in \mathbb{R}^2$ indicates the AP location. The narrowband fading channel coefficient between the m^{th} AP and k^{th} user with $m = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, K$ is

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}, \tag{5.1}$$

where β_{mk} and $h_{mk} \sim \mathcal{CN}(0,1)$ are the large- and small-scale fading coefficients, respectively, independent of each other and over coherent intervals. A general expression for the large-scale fading coefficient (LSFC) β_{mk} is

$$\beta_{mk} = \frac{cz_{mk}}{||\mathbf{p} - \mathbf{q}_m||^{\gamma}}, \tag{5.2}$$

where c is a constant, z_{mk} is the shadow fading coefficient, and γ is the pathloss exponent. We consider the canonical cell-free system in the uplink regime where all APs serve a smaller number of users in the same time-frequency resource. Pursuant to this model, the received signal at AP m is

$$y_m = \sum_{k=1}^{K} \sqrt{\rho_r} g_{mk} s_k + w_m, \tag{5.3}$$

where for user k, ρ_r is the transmit power, s_k is the data symbol with $\mathbb{E}\{|s_k|^2\} = 1$, and $w_m \sim \mathcal{CN}(0,1)$ is the additive noise. To estimate the data symbols, the NC that connects all APs together via error-free backhaul accumulates the received signals written as follows

$$\mathbf{y} = \sum_{k=1}^{K} \sqrt{\rho_r} \mathbf{g}_k s_k + \mathbf{w}, \tag{5.4}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_M]^T$, $\mathbf{g}_k = [g_{1k}, g_{2k}, \ldots, g_{Mk}]^T$, and $\mathbf{w} = [w_1, w_2, \ldots, w_M]^T$.

When a zero forcing (ZF) combiner is used, the resulting signal at the NC is $\mathbf{r} = \left(\mathbf{G}^H \mathbf{G}\right)^{-1} \mathbf{G}^H \mathbf{y}$, where $\mathbf{G} = [g_{mk}]$ is a $M \times K$ matrix consisting of the channel coefficients. The achievable per-user SNR for user k in this case is

$$\psi_k^{\mathrm{ZF}} = \frac{\rho_r}{[(\mathbf{G}^H \mathbf{G})^{-1}]_{kk}}. \tag{5.5}$$

An asymptotic approximation for the per-user SNR [15, 60] is

$$\frac{1}{M} \psi_k^{\mathrm{ZF}} \xrightarrow[M \to \infty]{\text{a.s.}} \rho_r \overline{\beta}_k, \tag{5.6}$$

where

$$\overline{\beta}_k \triangleq \lim_{M \to \infty} \frac{1}{M} \sum_{m} \beta_{mk}. \tag{5.7}$$

The achievable rate then is $R_k = \mathbb{E}\{\log(1 + \psi_k^{\mathrm{ZF}})\}$, where the expectation is over all the small-scale and shadow fading coefficients.

5.3 End-to-End AP Deployment

In this section, we discuss the entire proposed AP deployment procedure and the associated challenges. Table 5.1 provides a brief summary of the procedure. To develop the process, cooperation forms an important part. Hence, also noted in the table are the levels of cooperation among the APs expected in each step as well as the relevant algorithms. We provide here a brief summary and implementation of steps involved, and their details are provided in the ensuing sections.

Table 5.1: Steps in the procedure for end-to-end cell-free AP deployment. LSFC stands for the large-scale fading coefficient.

No.	Step	Details	Cooperation	Algorithm(s)
0	Start	Users positions (unknown) APs positions (known)	-	-
1	AP Clustering	• Performed through *agglomeration clustering*. • Minimum cluster size must be three APs (required for user position determination in step 2). • User-AP association is based on the lowest Euclidean distance. Each user associates with one AP for the purposes of the placement process alone.	Full cooperation	Algorithm 15
2	User position determination	• Each AP in a cluster receives LSFC from each user in the cluster. • Within-cluster APs share LSFC coefficients with each other. • Knowing/estimating the pathloss exponent and pathloss constant corresponding to a LSFC model, user positions are computed through *multilateration*.	Limited cooperation	(5.15) and (5.16)
3	AP placement	• Based on knowledge of user positions (step 2), an AP placement solution can directly be employed. • APs ultimately know the positions of their associated users.	Full cooperation	-
4	Fine-tuning AP positions	• Per-user SNR can be calculated by serving AP by knowing user positions, other AP positions, and pathloss model, and hence calculating the LSFCs of users with other non-serving APs. • Preliminary step is *AP subset selection* through the k-means++ procedure where APs sufficiently far away are selected so that updates in their position do not affect any other APs. • Subsequent step is the *AP position update* involving the selected APs and their associated users, and includes the following techniques: – Local max-sum gradient ascent (for sum rate maximization) – Local majorization-minimization (MM) (for sum rate maximization) – Local max-min gradient ascent (for minimum rate maximization)	Full cooperation	Algorithm 16 and Algorithm 17

- We start with an existing deployment (e.g., uniform or some prevailing positions) of APs whose positions are assumed to be known by the NC. This information is also shared to all APs. The NC however is not aware of the current user positions (or their distribution).
- *AP clustering:* The AP positions are clustered by the NC using agglomeration clustering (AC) keeping in mind that there must at least be three APs in each cluster so that the user positions can be found in the next step. The NC sends the cluster indices back to the APs while also assigning one AP within each cluster as the *cluster head* (CH). Note that for placement purposes, a user associates to the AP closest to it. This could also correspond to the AP with which it experiences the lowest pathloss. The APs share this information with the NC. When the clusters are formed, all associated users to each within-cluster APs also form part of the cluster.
- *User position determination:* The scenario of limited cooperation now comes into play where the LSFCs between each user and AP in each cluster are shared to the CH. This allows for multilateration so that the CH can determine the positions of the users in the cluster.
- *AP placement:* The CHs share the calculated within-cluster user positions to the NC so that the AP placement process can be performed. While many placement techniques exist, we choose the Lloyd algorithm among the VQ-based techniques motivated and described in [109]. At the end of this placement step, the NC informs each AP of their new positions as well as the associated users and their positions. Note that these user locations may be calculated again at a later stage through limited cooperation (as in the first and second steps).
- *Fine-tuning of the AP positions:* This fine-tuning procedure is also conducted at the NC where first, a subset of APs that are sufficiently far apart are chosen so that the update in the placement of one does not significantly affect the others. The SNR of each user is calculated by its associated AP since its position, the pathloss model, and the locations of the other APs are known. The AP update equations corresponding to the sum rate and minimum rate maximizations are then performed locally at each AP. Note that the NC could perform these updates for the APs in the selected subset, however, it is not necessary. The NC then chooses another subset and the AP position update is performed. The NC further repeats the entire process until the required performance is achieved or a chosen number of iterations are conducted. Further, in each iteration, the user-AP association is repeated and the NC communicates this association to the APs.

We now make some remarks about the above deployment procedure.

- *New deployment from the old deployment.* Once the deployment process is concluded and the new AP positions are determined, the NC compares the new and old AP positions in the form of pairwise distances between the AP positions in both cases. The NC then determines where each AP should move to its new location such that the cumulative distance moved by the APs is at a minimum.
- *Limited cooperation cell-free.* After deployment, we can further consider that the APs cooperate only partially as opposed to complete cooperation in the canonical cell-free setup. This limited cooperation

scenario can be achieved by the NC performing the AP clustering step (step 1) following the fine-tuning (step 4). To evaluate the performance of this setup, the CH by knowing the received signals and the channel coefficients from a user to all within-cluster APs, can design a suitable combiner for data symbol estimation and SINR calculation for the user.

- *Updating AP positions without the placement step.* After the NC knows the positions of the users (step 2), the fine-tuning procedure may directly be performed without the placement process. This is useful in situations where the user distribution is deemed by the NC not to have changed significantly. This determination is made by the NC by measuring a change (normally a loss) in the average system performance caused by the changed user configuration.

To demonstrate the end-to-end deployment, we discuss in the subsequent sections the implementation of clustering, user position determination, and the fine-tuning of the AP positions. Note that since AP placement has been discussed previously in a multitude of works, we do not include it here to avoid repetition. We also note that the two main placement optimization problems to be considered, namely the sum rate and minium rate maximization problems have also been defined in [109] and the previous chapter.

5.4 AP Clustering

To perform clustering of the APs, we utilize a popular technique called as *agglomeration clustering* (AC) [128]. AC is a type of hierarchical clustering method, particularly, a bottom-up approach where each data point is placed initially into its own cluster and then gradually merged to form larger clusters. The combination of clusters is based on their proximities to each other, which are evaluated using some cluster distance measures. A graphical representation of this process is called a dendrogram in which each cluster combination representing one step of AC is shown in order (until a single cluster is obtained for the data) and at the cluster distance level at which the combination takes place (called the height in the dendrogram).

Benefits of AC. Compared to the well-known k-means or Lloyd algorithm [41] clustering methods from VQ, AC benefits from the following advantages [129]:

- The Lloyd algorithm requires initialization and the final clustering result is then dependent on the initial points chosen for the cluster centers. AC does not require any selection of initial points since the data points start as individual clusters. Thus, there is a degree of consistency in the clustering result when AC is applied.

- If a slight modification to the number of clusters is required in the Lloyd algorithm, then a complete update of the clustering must be performed, i.e., the entire clustering algorithm must be run again. However, in the case of AC, since the each clustering step involves the combination of two existing clusters, the number of clusters may be changed by either undoing or performing more combinations (the dendrogram is quite useful in this scenario). This allows for great flexibility in deciding the number of clusters required.

The combination of clusters within the AC process is decided using the measured "distance" between the clusters. In particular, the distance between individual elements in the cluster is denoted by d, typically the Euclidean distance, while we refer to the inter-cluster distance as *linkage*, denoted by D. Given $\mathcal{A}$ and $\mathcal{B}$ as two clusters, with cluster elements defined as $a \in \mathcal{A}$ and $b \in \mathcal{B}$, we have the following linkages [130]:

- Single: $D_S(\mathcal{A}, \mathcal{B}) = \min_{a \in \mathcal{A}, b \in \mathcal{B}} d(a, b)$
- Complete: $D_C(\mathcal{A}, \mathcal{B}) = \max_{a \in \mathcal{A}, b \in \mathcal{B}} d(a, b)$
- Average: $D_A(\mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}||\mathcal{B}|} \sum_{a \in \mathcal{A}, b \in \mathcal{B}} d(a, b)$
- Centroid: $D_M(\mathcal{A}, \mathcal{B}) = d(\bar{\mathbf{r}}_{\mathcal{A}}, \bar{\mathbf{r}}_{\mathcal{B}})$

where $\bar{\mathbf{r}}_{\mathcal{A}}$ represents the centroid of all elements in cluster $\mathcal{A}$. In general, a proximity matrix can be generated which contains the linkages between every possible pair of clusters and is updated as AC progresses. The AC algorithm as applied to AP clustering is defined in Algorithm 15. Note that in our implementation, there should be at least three APs within a cluster indicating the number of reference points required for the multilateration step to be performed.

Algorithm 15 Agglomeration Clustering (AC)

1: Initialize all APs as individual clusters.
2: Generate the proximity matrix for the set of clusters based on the selected linkage scheme.
3: Join the two closest clusters.
4: Repeat from step 2 until minimum cluster size is three.

User-AP Association. For the placement process, we consider that the each user associates to its geographically closest AP, using the squared Euclidean distance. This is similar to the nearest neighbor condition in the Lloyd algorithm [109]. Formally, for AP m, the set of users, denoted by $\mathcal{C}_m$, that are associated with it are

$$\mathcal{C}_m = \{\mathbf{p}_k : ||\mathbf{p}_k - \mathbf{q}_m||^2 \leq ||\mathbf{p}_k - \mathbf{q}_l||^2, \forall l \neq m\}. \tag{5.8}$$

5.5 User Position Estimation

While user position determination has been conducted in the past, it is necessary in cell-free AP deployment and is assisted by the clustering in the previous step. Though in the system model (Section 5.2) we consider the canonical cell-free model where all APs cooperate, we assume this to be true only to estimate the data symbols and for the evaluation of the network performance. From the previous section, we have obtained clusters of APs (with a minimum cluster size of three); the users that are associated with those APs also form the cluster. Consequently, in this limited cooperation scenario, we consider that the APs are grouped into N AP clusters (referred to simply "clusters" henceforth), where no AP is a member of two clusters. Each cluster has M_n APs where $n = 1, 2, \ldots, N$, such that $\sum_{n=1}^{N} M_n = M$. In each cluster,

K_n users communicate to the M_n APs such that $\sum_{n=1}^{N} K_n = K$. Note that we will use m_n and k_n to index the AP and user, respectively, in cluster n such that $m_n = 1, 2, \ldots, M_n$ and $k_n = 1, 2, \ldots, K_n$.

Each cluster performs multilateration to determine the user positions within the cluster. In this process, each AP in the cluster shares the LSFCs of all constituent users with all other within-cluster APs. In other words, in cluster n, all M_n APs, through mutual information exchange, know the LSFCs of all K_n users in the cluster, i.e., $\beta_{m_n k_n}$, $\forall m_n, k_n$. In order to compute the position of user k_n, we consider the LSFC model in (5.2). By averaging out the shadow fading z_{mk} and letting $c' = c\mathbb{E}\{z_{m_n k_n}\}$, we have

$$\beta_{m_n k_n} = \frac{c'}{||\mathbf{p}_{k_n} - \mathbf{q}_{m_n}||^{\gamma}}, \ \forall m_n. \tag{5.9}$$

On rearranging, we obtain

$$\begin{aligned} ||\mathbf{p}_{k_n} - \mathbf{q}_{m_n}||^{\gamma} &= \frac{c'}{\beta_{m_n k_n}}, \\ \Rightarrow \quad ||\mathbf{p}_{k_n} - \mathbf{q}_{m_n}||^{2} &= \left(\frac{c'}{\beta_{m_n k_n}}\right)^{\frac{2}{\gamma}}. \end{aligned} \tag{5.10}$$

After expanding the ℓ_2-norm in the above expression and similar to [123], we perform the pairwise subtraction of each expression for $m_n = 2, 3, \ldots, M_n$ from that of the first ($m_n = 1$) AP, resulting in

$$2(\mathbf{q}_{m_n} - \mathbf{q}_1)^T \mathbf{p}_{k_n} = \left(\frac{c'}{\beta_{1,k_n}}\right)^{\frac{2}{\gamma}} - \left(\frac{c'}{\beta_{m_n k_n}}\right)^{\frac{2}{\gamma}} + \mathbf{q}_{m_n}^T \mathbf{q}_{m_n} - \mathbf{q}_1^T \mathbf{q}_1. \tag{5.11}$$

In a matrix formulation, (5.11) can be written as

$$\mathbf{A}_n \mathbf{p}_{k_n} = \mathbf{b}_{k_n}, \tag{5.12}$$

where

$$\mathbf{A}_n = 2 \begin{bmatrix} (\mathbf{q}_2 - \mathbf{q}_1)^T \\ (\mathbf{q}_3 - \mathbf{q}_1)^T \\ \vdots \\ (\mathbf{q}_{M_n} - \mathbf{q}_1)^T \end{bmatrix}, \tag{5.13}$$

and

$$\mathbf{b}_{k_n} = \begin{bmatrix} \left(\frac{c'}{\beta_{1,k_n}}\right)^{\frac{2}{\gamma}} - \left(\frac{c'}{\beta_{2,k_n}}\right)^{\frac{2}{\gamma}} + \mathbf{q}_2^T \mathbf{q}_2 - \mathbf{q}_1^T \mathbf{q}_1 \\ \left(\frac{c'}{\beta_{1,k_n}}\right)^{\frac{2}{\gamma}} - \left(\frac{c'}{\beta_{3,k_n}}\right)^{\frac{2}{\gamma}} + \mathbf{q}_3^T \mathbf{q}_3 - \mathbf{q}_1^T \mathbf{q}_1 \\ \vdots \\ \left(\frac{c'}{\beta_{1,k_n}}\right)^{\frac{2}{\gamma}} - \left(\frac{c'}{\beta_{M_n k_n}}\right)^{\frac{2}{\gamma}} + \mathbf{q}_{M_n}^T \mathbf{q}_{M_n} - \mathbf{q}_1^T \mathbf{q}_1 \end{bmatrix}. \tag{5.14}$$

If a cluster consists of only three APs, then from (5.12) we can write the user position estimate as $\hat{\mathbf{p}}_{k_n} = \mathbf{A}_n^{-1} \mathbf{b}_{k_n}$ [123]. However, when $M_n > 3$, (5.12) is an overdetermined system of equations and the user position estimate is obtained using ordinary least squares as follows

$$\hat{\mathbf{p}}_{k_n} = (\mathbf{A}_n^T \mathbf{A}_n)^{-1} \mathbf{A}_n^T \mathbf{b}_{k_n}. \tag{5.15}$$

Additionally, we note that determination of the user positions utilizing (5.15) requires the knowledge of the pathloss exponent γ and constant c. These parameters may be estimated using the technique similar to that

described in [123]. Additionally, as a consequence of estimation of the abovementioned pathloss parameters, the vector $\mathbf{b}_{k_n}$ in (5.12) is thus an estimate, denoted by $\hat{\mathbf{b}}_{k_n}$. Then, the estimate of the user position is calculated by using the weighted least squares [124], as follows

$$\hat{\mathbf{p}}_{k_n} = (\mathbf{A}_n^T \mathbf{W}_{k_n}^{-1} \mathbf{A}_n)^{-1} \mathbf{A}_n^T \mathbf{W}_{k_n}^{-1} \hat{\mathbf{b}}_{k_n}, \tag{5.16}$$

where $\mathbf{W}_{k_n}$ is the covariance matrix of $\hat{\mathbf{b}}_{k_n}$. Finally, the user positions determined in each cluster are shared by the APs to the NC for the AP placement step that follows.

5.6 Fine-Tuning AP Positions

In this section, we present a procedure to fine-tune the AP positions once the user positions have been determined at the APs and the placement of APs is performed. The fine-tuning of the AP locations is specifically dependent on which of the two main placement metrics, namely the sum and minimum rates, is to be improved.

Note that when the user positions were determined by clustering APs, only the APs within a cluster are aware of the LSFCs between them and the constituent users. Following the prior placement process, the SNRs of the users associated with each AP can be calculated at that AP since the user positions, the pathloss parameters, and the positions of all other APs are known. The fine-tuning is designed as a two-step iterative procedure where a subset of APs is first chosen sufficiently far apart from each other so that their positions can be updated without interfering with each other. Then, the positions of the selected APs is updated according to the methods described below. A different subset (with no repetition) is selected each time and their positions are also updated. The whole procedure is repeated until the performance has improved to the required level or a set number of iterations have been concluded.

5.6.1 AP Subset Selection

For the AP subset selection, the k-means++ method [95], used typically for initializing the Lloyd or Lloyd-type algorithms, is applied and is described in Algorithm 16. Note that $d_{\mathrm{SE}}(\mathbf{q}_x, \mathbf{q}_{\mathcal{N}(x)})$ represents the squared Euclidean distance between AP x and its nearest AP, indexed by $\mathcal{N}(x)$, and M_s is the number of APs to be selected in each subset.

Algorithm 16 k-means++

1: Choose one AP uniformly at random.
2: For each AP x not chosen yet, compute $d_{\mathrm{SE}}(\mathbf{q}_x, \mathbf{q}_{\mathcal{N}(x)})$.
3: Choose one new AP at random as the new AP, using a weighted probability distribution where AP x is chosen with probability proportional to $d_{\mathrm{SE}}(\mathbf{q}_x, \mathbf{q}_{\mathcal{N}(x)})$.
4: Repeat steps 2 and 3 until M_s APs have been chosen.

5.6.2 AP Position Update

Here, we describe the various AP position update methods according to whether the sum rate or minimum rate performance of the system is to be improved. A brief description of the two problems can been found in [109].

Sum Rate Maximization

To perform the position update of AP m for the maximization of sum rate, whose optimization problem is defined as

$$\arg\max_{\mathbf{q}_m} \sum_{\mathbf{p}_k \in \mathcal{C}_m} \log\left(1+\psi_k^{\mathrm{ZF}}\right), \tag{5.17}$$

we present two approaches, namely the local max-sum gradient and the local majorization-minimization (MM) method.

To maximize the sum rate for AP m, the gradient update expression with j as the iteration index is

$$\mathbf{q}_m^{(j+1)} = \mathbf{q}_m^{(j)} + \delta \frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \sum_{\mathbf{p}_k \in \mathcal{C}_m} \log\left(1+\rho_r \sum_{m=1}^{M} \beta_{mk}\right) \right\}, \ \forall m, \tag{5.18}$$

where in the objective function in (5.17), shadow fading is ignored and δ is the step size. The gradient in (5.18) is calculated as

$$\frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \sum_{\mathbf{p}_k \in \mathcal{C}_m} \log\left(1+\rho_r \sum_{m=1}^{M} \beta_{mk}\right) \right\} = \frac{\gamma \rho_r}{2} \sum_{\mathbf{p}_k \in \mathcal{C}_m} \frac{1}{1+\psi_k^{\mathrm{ZF}}} \frac{(\mathbf{p}_k - \mathbf{q}_m^{(j)})}{||\mathbf{p}_k - \mathbf{q}_m^{(j)}||^{\gamma+2}}. \tag{5.19}$$

As the alternate approach, the MM iteration update equations to compute the subset AP locations are

$$\begin{aligned} \mathbf{q}_m^{(j+1)} &= \frac{\sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k^{(j)} \mathbf{p}_k}{\sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k^{(j)}}, \\ w_k^{(j+1)} &= \frac{1+||\mathbf{p}_k - \mathbf{q}_m^{(j+1)}||^2 t_{mk}}{||\mathbf{p}_k - \mathbf{q}_m^{(j+1)}||^2 + \epsilon}, \ \forall \mathbf{p}_k \in \mathcal{C}_m, \end{aligned} \tag{5.20}$$

where

$$t_{mk} = \sum_{\substack{s=1 \\ s \neq m}}^{M} \frac{1}{||\mathbf{p}_k - \mathbf{q}_s||^{\gamma}}. \tag{5.21}$$

The above update equations are obtained by assuming a high SNR regime in the objective function in (5.17), followed by the simplification

$$\sum_{\mathbf{p}_k \in \mathcal{C}_m} \log\left(\sum_{m=1}^{M} \beta_{mk}\right) = \sum_{\mathbf{p}_k \in \mathcal{C}_m} \log\left(\frac{1}{||\mathbf{p}_k - \mathbf{q}_m||^{\gamma}} + t_{mk}\right). \tag{5.22}$$

From this expression, the MM optimization approach similar to that described in [80] can be employed to achieve the above iteration equations. This proof is provided in Appendix 5.A.

Minimum Rate Maximization

The optimization problem here corresponding to AP m is

$$\arg\max_{\mathbf{q}_m} \min_{\mathbf{p}_k \in \mathcal{C}_m} \log\left(1+\psi_k^{\mathrm{ZF}}\right). \tag{5.23}$$

Applying gradient ascent to just the worst user alone would make the gradient value vary significantly across the iterations, resulting in convergence issues. Additionally, we are interested in evaluating the system performance utilizing the 95%-likely rate (the best rate among the worst 5% of the users). Hence, to evaluate the minimum rate performance, we use the following surrogate optimization problem

$$\arg\max_{\mathbf{q}_m} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{5\%}} \log\left(1+\psi_k^{\mathrm{ZF}}\right), \tag{5.24}$$

where the set $\mathcal{C}_m^{5\%}$ represents the set of the worst 5% of the users. The update equation then is

$$\mathbf{q}_m^{(j+1)} = \mathbf{q}_m^{(j)} + \delta \frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \sum_{\mathbf{p}_k \in \mathcal{C}_m^{5\%}} \log\left(1+\rho_r \sum_{m=1}^{M} \beta_{mk}\right) \right\}, \ \forall m, \tag{5.25}$$

and the gradient is

$$\frac{\partial}{\partial \mathbf{q}_m^{(j)}} \left\{ \sum_{\mathbf{p}_k \in \mathcal{C}_m^{5\%}} \log\left(1+\rho_r \sum_{m=1}^{M} \beta_{mk}\right) \right\} = \frac{\gamma \rho_r}{2} \sum_{\mathbf{p}_k \in \mathcal{C}_m} \frac{1}{1+\psi_k^{\mathrm{ZF}}} \frac{(\mathbf{p}_k - \mathbf{q}_m^{(j)})}{||\mathbf{p}_k - \mathbf{q}_m^{(j)}||^{\gamma+2}}. \tag{5.26}$$

The complete fine-tuning procedure written as an algorithm is outlined in Algorithm 17 below.

Algorithm 17 AP Fine-Tuning Procedure

1: Select M_s APs (not previously chosen) using k-means++ (Algorithm 16).
2: Update selected AP positions though local max-sum in (5.18), local MM in (5.20), or local max-min in (5.25).
3: Repeat from step 1 until all M APs have been chosen.
4: Repeat from step 1 until required rate improvement has been met or maximum number of iterations have been reached.

5.7 Simulation Methodology and Results

To conduct numerical simulations, we consider that the users are distributed over a geographical area of dimensions 2×2 km^2. $M = 32$ APs serve $K = 4$ users, since $M \gg K$ for a cell-free system. For the purpose of AP placement, however, we assume a total of 2000 users, distributed as a Gaussian mixture model (GMM) of the form

$$f_{\mathbf{P}}(\mathbf{p}) = \sum_{l=1}^{L} p_l \mathcal{N}\left(\mathbf{p}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right), \tag{5.27}$$

where L is the number of mixture components (clusters), and p_l, $\boldsymbol{\mu}_l$, and $\boldsymbol{\Sigma}_l$ are the probability, mean, and covariance matrix, respectively, of mixture component l. The parameters take values $L = 3$, $\boldsymbol{\mu}_1 =$

Table 5.2: Final number of clusters under different linkages

Linkage	Final Number of Clusters
Single	3
Complete	4
Average	4
Centroid	4

$[0.5, -0.5]^T$, $\boldsymbol{\mu}_2 = [0, 0.5]^T$, $\boldsymbol{\mu}_3 = [-0.5, 0]^T$, $p_1 = 0.6$, and $p_2 = p_3 = 0.2$. The covariance matrices of the clusters are $\boldsymbol{\Sigma}_1 = \sigma^2\mathbf{I}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma^2 & 2\sigma^2/3 \\ 2\sigma^2/3 & 2\sigma^2 \end{bmatrix}$, and $\boldsymbol{\Sigma}_3 = \sigma^2\mathbf{I}$ with $\sigma = 100$. The pathloss model from [61, (4.34)] is used with shadow fading ignored for simplicity and the transmit power ρ_r of the users is varied from 5 to 30 dB. The local max-sum and local max-min gradient methods utilize step sizes of $\delta = 10^3$ and $\delta = 3 \times 10^4$, respectively.

5.7.1 Performance Measures

To measure performance, we utilize the per-user achievable rate defined for user k as

$$R_k = \mathbb{E}\left\{\log_2\left(1 + \psi_k^{\mathrm{ZF}}\right)\right\}, \tag{5.28}$$

with ψ_k^{ZF} from (5.5). R_k is used to generate sum rate and 95%-likely rate values for comparison between different AP placements. The rate values are generated through Monte Carlo simulations and the best solution is selected after the algorithm is run multiple times. The relative performance between algorithms (say, algorithm 2 over algorithm 1) can be calculated by using the following measure expressed as percentage

$$\text{Improvement Ratio} = \frac{\mathcal{P}^{\text{Algorithm 2}} - \mathcal{P}^{\text{Algorithm 1}}}{\mathcal{P}^{\text{Algorithm 1}}} \times 100, \tag{5.29}$$

where $\mathcal{P}$ is either the sum rate or 95%-likely rate.

5.7.2 Numerical Results

Experiment 1. We first implement the AP clustering using the different linkage methods discussed in Section 5.4. For this purpose, we initially consider the AP locations determined by the Lloyd algorithm for a GMM user configuration different to that described above. These AP locations represent the initial AP positions available before the entire cell-free AP deployment process is started, i.e., in step 0 in Table 5.1. In Fig. 5.1, we show these AP positions (with the user distribution that resulted these locations) and the result of the clustering procedure when single linkage is used. The figure shows that three clusters are generated and the clusters themselves are aligned to the mixture components of the user GMM. The final average number of clusters generated for all linkage methods are provided in Table 5.2. We will consider that the single linkage clustering is utilized for the remaining simulation experiments.

Experiment 2. In our simulations, we assume knowledge of the pathloss model and hence, the user positions obtained contain no error. However, it is possible that the when the user positions are determined,

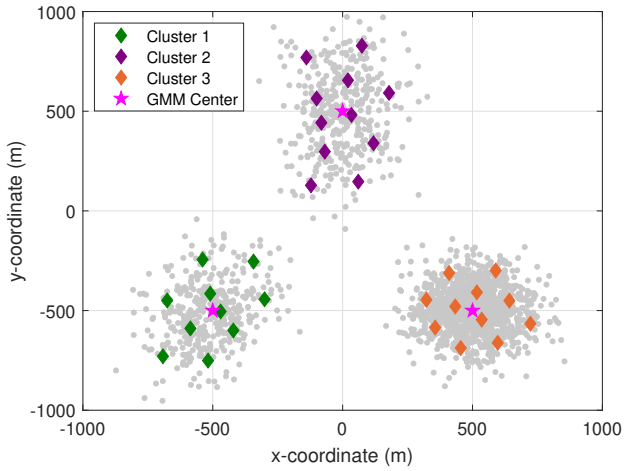

Figure 5.1: AP locations showing the clustering when single linkage is used.

especially when estimates of the pathloss constant and exponent are used, the user positions contain some error. Thus, in this experiment, we wish to observe the effect of such errors in the placement process. For this purpose, we assume that a random error is added to the user positions generated. In particular, we consider an i.i.d. Gaussian distributed error with zero mean and standard deviation σ_e added to each dimension of the user positions. We investigate the sum rate and 95%-likely rate performance of the Lloyd algorithm applied to the user positions with error by changing σ_e and compare them to the case where no error is present. We choose five different standard deviation values and in Fig. 5.2, the sum rates achieved are shown. It is observed that for the considered user configuration, the sum rates when $\sigma_e = 1$, 5, and 10, are all worse than when there is no error. Quantitatively, this value is about 5% as shown in Table 5.3. However, for higher σ_e values, the inferiority of the Lloyd algorithm when user position errors are considered is larger. In the case of the 95%-likely rate performances plotted in Fig. 5.3, the lowest $\sigma_e = 1$ shows the least change in performance while $\sigma_e = 5$, 10, and 25 all show up to about 5% inferior performance while when $\sigma_e = 50$, there is significant performance loss of about 26%. These losses are due to the fact that when the error standard deviation increases, the variance of the GMM mixture components or the spreading out of the user positions increases on average resulting in the Lloyd algorithm having to place APs farther away from the GMM centers while using the same number of APs. This effect is less observed when σ_e is small but becomes more pronounced for larger values.

Experiment 3. In this experiment, we conduct the numerical simulations that show the rate improvements achieved by using the fine-tuning procedure discussed in Section 5.6. We show the results generated for 10 iterations of the fine-tuning procedure for each of the cases where the local max-sum, local MM, and local max-min methods are utilized. In all cases, the AP locations obtained through the Lloyd algorithm are chosen as the initial positions for the fine-tuning algorithms. We start by showing both the initial AP positions obtained through the Lloyd algorithm and the final AP locations obtained when the fine-tuning procedure using all three algorithms, namely the local max-sum, local MM, and local max-min algorithms, is applied. For the sake of clear observation, we have shown the case where the fine-tuning procedure is intended to improve the sum rate in Fig. 5.4 and the case where the 95%-likely rate is to be improved in Fig.

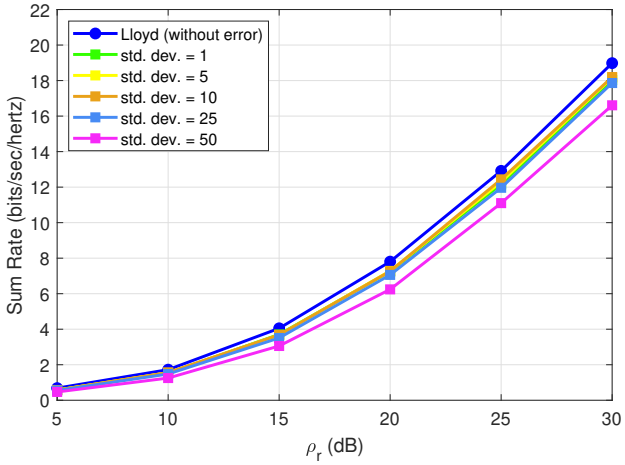

Figure 5.2: Sum rate as a function of ρ_r for the Lloyd algorithm with and without errors in user positions.

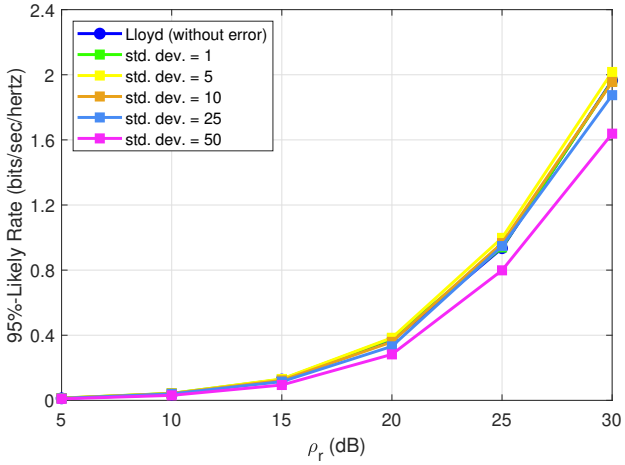

Figure 5.3: 95%-likely rate as a function of ρ_r for the Lloyd algorithm with and without errors in user positions.

Table 5.3: Percentage Rate Improvement of the Lloyd Algorithm with Error in User Positions Relative to the Lloyd Algorithm without Error at $\rho_r = 30$ dB

Error Standard Deviation σ_e	Sum Rate	95%-Likely Rate
1	-5.02%	-0.26%
5	-4.92%	-4.27%
10	-5.38%	-2.93%
25	-7.09%	-5.28%
50	-12.51%	-25.59%

5.5. From these two figures, we observe that as a result of the iterations of the fine-tuning procedure, the final AP locations obtained are distinct from the initial Lloyd positions. In particular, both the local max-sum and local MM algorithms situate the APs closer to the GMM centers compared to the Lloyd algorithm while

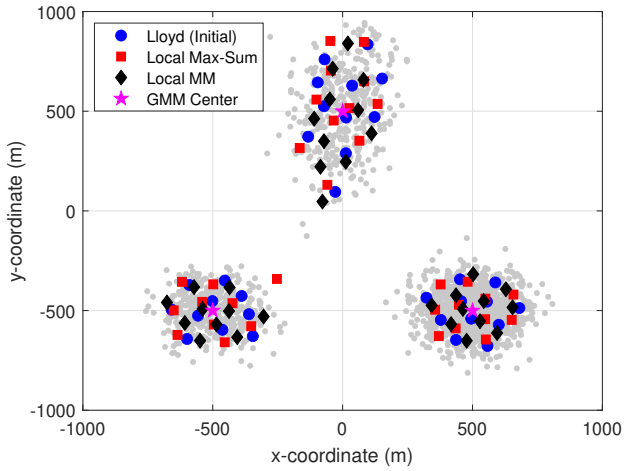

Figure 5.4: Initial AP positions using the Lloyd algorithm and final AP positions using the fine-tuning procedure with the local max-sum and local MM algorithms.

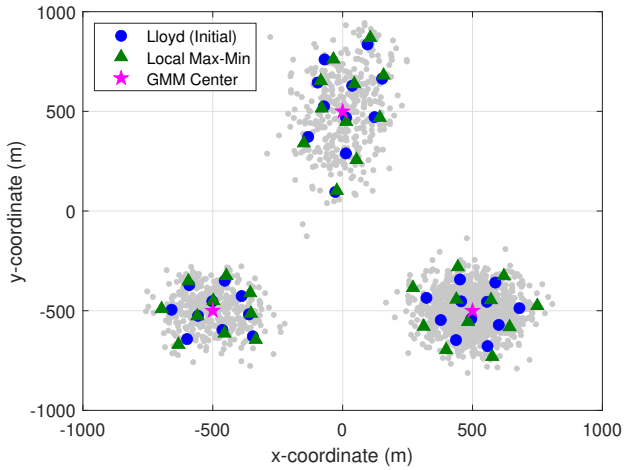

Figure 5.5: Initial AP positions using the Lloyd algorithm and final AP positions using the fine-tuning procedure with the local max-min algorithm.

the local max-min algorithm places the APs closer to the GMM edge users. Fig. 5.6 and Fig. 5.7 show the sum and 95%-likely rates corresponding to the three algorithms used in the fine-tuning procedure. Both the local max-sum and local MM algorithms increase the sum rate with the local MM method providing a 4.18% higher sum rate (at $\rho_r = 30$ dB) than the Lloyd algorithm, as quantified in Table 5.4. This results in a lower 95%-likely rate, where both algorithms achieve about 7.5% lower rate than the Lloyd algorithm. Note that the local MM algorithm is observed to achieve significantly lower 95%-likely rates at other power levels compared to the local max-sum method. In regards to the local max-min method, while a significantly lower sum rate performance is observed, the 95%-likely rate performance is 7.87% higher than that of the Lloyd algorithm. Hence, the fine-tuning procedures do accomplish their intended rate performance improvement.

Experiment 4. Here, we present an example where the fine-tuning procedure alone (i.e., without performing the placement step) can be advantageous in a situation where the user distributed has changed.

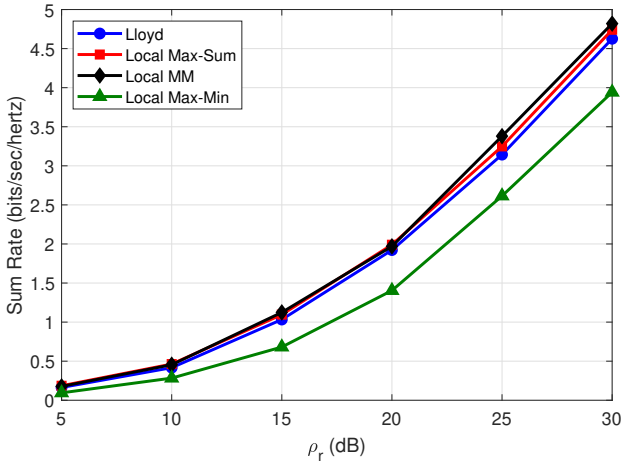

Figure 5.6: Sum rate as a function of ρ_r for the Lloyd algorithm and the fine-tuning procedure with local max-sum, local MM, and local max-min algorithms.

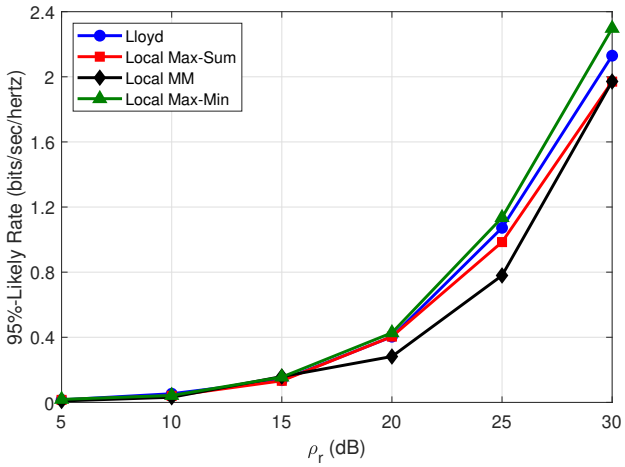

Figure 5.7: 95%-likely rate as a function of ρ_r for the Lloyd algorithm and the fine-tuning procedure with local max-sum, local MM, and local max-min algorithms.

Table 5.4: Rate Improvement of the Fine-Tuning Procedure with the Local Max-Sum, Local MM, and Local Max-Min Algorithms Relative to the Lloyd Algorithm at $\rho_r = 30$ dB

Algorithm	Sum Rate	95%-Likely Rate
Local Max-Sum	2.48%	−7.49%
Local MM	4.18%	−7.44%
Local Max-Min	−14.79%	7.87%

Such a situation was discussed in the remarks in Section 5.3. Consider the same distribution as been considered in the previous Experiment 4. Then, we determine the Lloyd AP positions for this density and then fine-tune the positions based on the local MM algorithm. We then consider a new distribution, which for the purpose of this examples has the covariance of the second mixture component changed to

Table 5.5: Rate Improvement of the Local MM Fine-Tuning Procedure Applied to the Lloyd Algorithm Positions and Applied to the Fine-Tuned AP Positions of the Previous Distribution Relative to the Lloyd Algorithm at $\rho_r = 30$ dB

Algorithm	Sum Rate	95%-Likely Rate
Local MM (with Lloyd)	3.75%	−13.34%
Local MM (without Lloyd)	8.16%	−6.53%

$\mathbf{\Sigma}_2 = \begin{bmatrix} 1.14\sigma^2 & 2\sigma^2/3 \\ 2\sigma^2/3 & 2.67\sigma^2 \end{bmatrix}$, with $\sigma = 100$, which causes the users in the mixture component to spread out slightly more than in the original distribution. Using this distribution, we implement the Lloyd algorithm and again, the fine-tuning procedure utilizing the local MM algorithm. However, in addition, we also implement the local MM algorithm with the AP positions corresponding to the old distribution as the starting point, i.e., without the Lloyd algorithm. We plot the sum rate and 95%-likely rate performances in Fig. 5.8 and Fig. 5.9, respectively, where we refer to the result in which the local MM fine-tuning procedure is applied to the Lloyd algorithm as 'Local MM (with Lloyd)', and the result in which the local MM fine-tuning procedure is applied to the fine-tuned AP positions of the previous (original) distribution as 'Local MM (without Lloyd)'. In the case of sum rate, it is observed that while Local MM (with Lloyd) performs better than the Lloyd algorithm as expected, the Local MM (without Lloyd) achieves the highest sum rate among the three curves. This performance can be explained by the fact that as the original user density in the second mixture component is expanded slight to the new density, the fine-tuned AP positions (using which the subsequent fine-tuning is performed) are closer to the center of the mixture component than the Lloyd algorithm. Since the local MM procedure now starts with these AP positions which are closer to the majority of the users, which by themselves exhibit a higher sum rate, applying the local MM algorithm to these positionss result in an even higher sum rate performance. From Table 5.5, Local MM (without Lloyd) is shown to have a 8.16% increase in sum rate at $\rho_r = 30$ dB over the Lloyd algorithm, over double the 3.75% increase of the Local MM (with Lloyd). The 95%-likely rate performances of both the local MM algorithms implemented are anticipatedly lower than the Lloyd algorithm. The results show that in this situation when the user density is changed to a small degree, the fine-tuning procedure can be directly applied to the existing AP positions as opposed to repeating the AP placement process followed by the fine-tuning procedure. The advantage in this case is that the local (and limited) information exchange alone is necessary instead of the complete information exchange required by the AP placement algorithms. However, it is important to note that the difference in the the user distribution is byond a certain degree, then the AP placement process must be executed once again.

5.8 Conclusion

In this paper, we have addressed cell-free access point (AP) deployment in uplink multiple-input-multiple-output (MIMO) systems for optimal throughput. To the AP placement problem, we added user position determination, limited AP cooperation, and fine-tuning of the AP positions. Accordingly, a multi-step deployment process was proposed where agglomeration clustering of the APs first allowed limited coop-

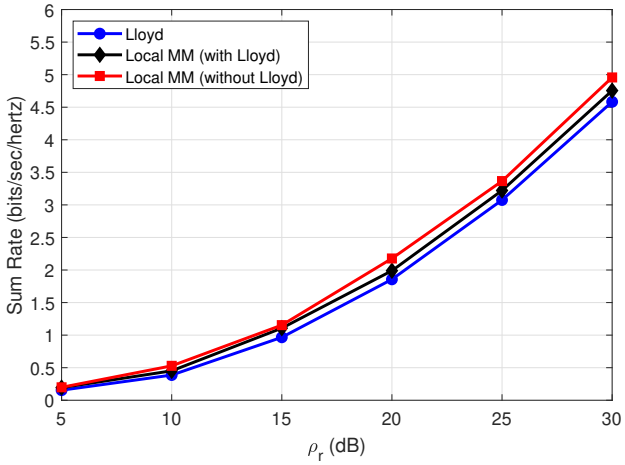

Figure 5.8: Sum rate as a function of ρ_r for the Lloyd algorithm and the local MM fine-tuning procedure applied to the Lloyd algorithm positions (with Lloyd) and applied to the fine-tuned AP positions of the previous distribution (without Lloyd).

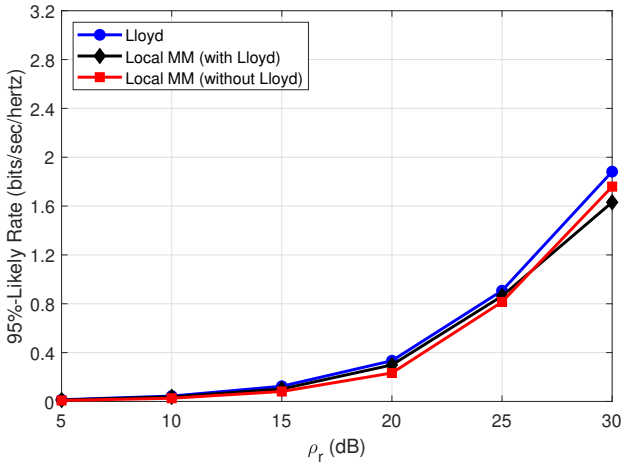

Figure 5.9: 95%-likely rate as a function of ρ_r for the Lloyd algorithm and the local MM fine-tuning procedure applied to the Lloyd algorithm positions (with Lloyd) and applied to the fine-tuned AP positions of the previous distribution (without Lloyd).

eration to calculate the user positions through multilateration, necessary for the placement process. Then, an AP placement solution is implemented. Finally, to further improve the system throughput, fine-tuning procedures based on local max-sum and local majorization-minimization methods (for sum rate maximization), and local max-min (for minimum rate maximization) were proposed. The entire procedure presents a practical and a more complete approach to cell-free AP deployment compared to prior works. Accordingly, in addition to the Lloyd algorithm providing a good sum and minimum rate performance for the AP placement step, the fine-tuning procedures are able to generate up to around 4% increase in sum rate and 8% improvement in the 95%-likely rate of the network.

Chapter 5, in part, is a reprint with permission of the material as it appears in the paper: Govind Ravikumar Gopal and Bhaskar D. Rao, "Throughput oriented access point deployment in cell-free massive

MIMO systems," which is submitted to *IEEE Transactions on Wireless Communications*. The dissertation author was the primary investigator and author of these papers. These works were supported in part by National Science Foundation (NSF) under Grant CCF-2124929 and Grant CCF-2225617, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the Center for Wireless Communications (CWC), University of California San Diego.

Appendices

5.A Proof of Solution for the Local MM Method

The objective function in (5.17) can be simplified as follows

$$\begin{aligned}\sum_{\mathbf{p}_k\in\mathcal{C}_m}\log\left(1+\psi_k^{\mathrm{ZF}}\right) &\overset{(a)}{\approx} \sum_{\mathbf{p}_k\in\mathcal{C}_m}\log\left(\sum_{m=1}^{M}\beta_{mk}\right),\\ &= \sum_{\mathbf{p}_k\in\mathcal{C}_m}\log\left(\beta_{mk}+\sum_{\substack{s=1\\s\neq m}}^{M}\beta_{sk}\right),\\ &\overset{(b)}{=} \sum_{\mathbf{p}_k\in\mathcal{C}_m}\log\left(\frac{1}{||\mathbf{p}_k-\mathbf{q}_m||^{\gamma}}+t_{mk}\right),\end{aligned} \tag{5.30}$$

where in (a) we have assumed a high SNR regime and in (b), we have used the LSFC model from (5.2), $t_{mk}=\sum\limits_{\substack{s=1\\s\neq m}}^{M}\frac{1}{||\mathbf{p}_k-\mathbf{q}_s||^{\gamma}}$, and the pathloss constant c and shadow fading are neglected. In order to align the optimization problem in (5.17) with the MM framework, it is converted to a minimization problem by negating the objective function. Thus, letting J be the new objective function, we have

$$J=-\sum_{\mathbf{p}_k\in\mathcal{C}_m}\log\left(\frac{1}{||\mathbf{p}_k-\mathbf{q}_m||^{\gamma}+\epsilon}+t_{mk}\right), \tag{5.31}$$

which can be majorized. Thus, the new optimization problem is

$$\arg\min_{\mathbf{q}_m}\sum_{\mathbf{p}_k\in\mathcal{C}_m}\log\left(\frac{||\mathbf{p}_k-\mathbf{q}_m||^{\gamma}+\epsilon}{1+||\mathbf{p}_k-\mathbf{q}_m||^{\gamma}t_{mk}}\right). \tag{5.32}$$

Note that we have added a constant $\epsilon>0$ which is typically a very small quantity and is used to prevent the logarithm from approaching negative infinity if the user position were to overlap with the AP position. Conforming to the MM framework, a concave function can be upper bounded by its first-order Taylor expansion [66]

$$h(z)\leq h'(z_l)(z-z_l)+h(z_l), \tag{5.33}$$

where $h(\cdot)$ is concave on $\mathbb{R}^+$, z is the variable, z_l is the point around which the expansion is carried out, and $h'(\cdot)$ is the first derivative. In the objective function of (5.32), we can take $h(z_k)=\log(z_k)$ and

$z_k = \frac{||\mathbf{p}_k - \mathbf{q}_m||^\gamma + \epsilon}{1 + ||\mathbf{p}_k - \mathbf{q}_m||^\gamma t_{mk}}$. Accordingly, applying (5.33) to (5.32), the objective function is

$$J_1 = \sum_{\mathbf{p}_k \in \mathcal{C}_m} h(z_k) \leq \sum_{\mathbf{p}_k \in \mathcal{C}_m} [h'(z_{k,l})(z_k - z_{k,l}) + h(z_{k,l})]. \tag{5.34}$$

Removing the irrelevant terms in the optimization, we get

$$\begin{aligned} J_2 &= \sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k z_k, \\ &= \sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k \left(\frac{||\mathbf{p}_k - \mathbf{q}_m||^\gamma + \epsilon}{1 + ||\mathbf{p}_k - \mathbf{q}_m||^\gamma t_{mk}} \right), \\ &< \sum_{\mathbf{p}_k \in \mathcal{C}_m} w_k \left(||\mathbf{p}_k - \mathbf{q}_m||^\gamma + \epsilon \right), \end{aligned} \tag{5.35}$$

where the strict inequality in the last step is due to the fact that the denominator in the fraction is greater than 1 as a consequence of the Voronoi regions $\mathcal{C}_m$. In the above expression, the weight w_k is defined as

$$\begin{aligned} w_k = h'(z_{k,l}) &= \left. \frac{\partial h(z_{k,l})}{\partial z_{k,l}} \right|_{z_{k,l} = \frac{||\mathbf{p}_k - \mathbf{q}_m||^\gamma + \epsilon}{1 + ||\mathbf{p}_k - \mathbf{q}_m||^\gamma t_{mk}}}, \\ &= \left. \frac{1}{z_{k,l}} \right|_{z_{k,l} = \frac{||\mathbf{p}_k - \mathbf{q}_m||^\gamma + \epsilon}{1 + ||\mathbf{p}_k - \mathbf{q}_m||^\gamma t_{mk}}} = \frac{1 + ||\mathbf{p}_k - \mathbf{q}_m||^\gamma t_{mk}}{||\mathbf{p}_k - \mathbf{q}_m||^\gamma + \epsilon}. \end{aligned} \tag{5.36}$$

This is the weight update equation. It should be noted that we are interested in minimizing the surrogate objective function J_2 defined in (5.35). Given the weights w_k, the derivative of the objective function J_2 can be taken and equated to zero in order to find the AP update equation. However, the pathloss exponent γ prevents us from obtaining a closed-form expression in this case. Fortuitously, the theory of MM states that it suffices to have the objective function reduce or remain constant from one iteration to the next. Thus, we can assume that the pathloss exponent in J_2 is two, and by taking the gradient, i.e., $\partial J_2 / \partial \mathbf{q}_m = 0$, we have the update equation for the AP position as

$$\mathbf{q}_m = \frac{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m} w_k \mathbf{p}_k}{\sum\limits_{\mathbf{p}_k \in \mathcal{C}_m} w_k}. \tag{5.37}$$

Chapter 6

Conclusions

In this dissertation, we have broadly discussed access point (AP) placement in distributed massive multiple-input-multiple-output (MIMO) systems in the context of optimal throughput. We considered two distinct types of networks, namely the non-cooperative small-cell and the cooperative cell-free systems.

First, the small-cell AP placement problem that takes into account inter-cell interference (ICI) was investigated and its connection to the vector quantization (VQ) framework carefully explored. It was established that while VQ is not the same as AP placement, the solution framework, specifically, the Lloyd algorithm can be used as a stepping stone for throughput optimization and other related problems. Consequently, the inter-AP Lloyd algorithm was proposed to place APs while accounting for ICI and increasing the minimum rate of the system.

Second, the placement solution was extended to hybrid terrestrial and unmanned aerial vehicle (UAV) AP systems, where an alternate measure, that is the signal-to-generated-interference-plus-noise ratio was maximized to place UAV-APs according to the changing user density. Further, load balancing (of cell occupancies) was considered as another extension problem to address user-AP access delay and two Lloyd-type algorithms were designed to balance cell occupancies while minimizing the associated throughput loss.

Third, the sum rate and minimum rate maximization AP placement problems in cell-free networks were analyzed in terms of their solutions structures. The VQ framework was found to be a good candidate for obtaining AP locations. Consequently, in addition to the Lloyd algorithm, two other matched techniques, namely the tree-structured VQ (TSVQ) and the probability density function optimized VQ (PDFVQ) were proposed. In addition to advantages of a distributed solution, promoting cooperation among closely-spaced APs, and scalability, these techniques were shown to provide good starting points for gradient-based throughput optimizations.

Finally, to address major challenges in cell-free AP placement, namely user position determination and limited cooperation in order to reduce backhaul, a multi-step AP deployment procedure was proposed. This procedure also included a fine-tuning step in which APs locally adjust their positions in order to further improve the sum or minimum rates.

In terms of future directions, there are multiple problems that can be addressed. The current

VQ framework is an unsupervised learning method and thus, data driven models can be designed for the placement of APs. Reinforcement learning schemes, such as a multi-agent deep Q-learning framework can also be set up for the throughput oriented AP placement problems discussed in this dissertation. Additionally, the consideration of multiple antennas at both the AP and the user sides must be undertaken to reflect a true MIMO system. Further, power control can be added as an additional optimization problem to further improve system throughput.

Bibliography

[1] W. Belaoura, K. Ghanem, M. Z. Shakir, and K. Qaraqe, "Impact of hardware impairments on the performance of millimeter-wave massive MU-MIMO systems with distributed antennas," in *2020 IEEE Eighth Int. Conf. Commun. Netw. (ComNet)*, Oct. 2020, pp. 1–4.

[2] W. S. H. M. W. Ahmad, N. A. M. Radzi, F. S. Samidi, A. Ismail, F. Abdullah, M. Z. Jamaludin, and M. N. Zakaria, "5G technology: Towards dynamic spectrum sharing using cognitive radio networks," *IEEE Access*, vol. 8, pp. 14 460–14 488, Jan. 2020.

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

[4] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[5] R. Chataut and R. Akl, "Massive MIMO systems for 5G and beyond networks — overview, recent trends, challenges, and future research direction," in *Sensors*, vol. 20, no. 10, May 2020, Art. ID 2753.

[6] J. Jeon, G. Lee, A. A. Ibrahim, J. Yuan, G. Xu, J. Cho, E. Onggosanusi, Y. Kim, J. Lee, and J. C. Zhang, "MIMO evolution toward 6G: Modular massive MIMO in low-frequency bands," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 52–58, Nov. 2021.

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

[8] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[9] Y. Huang, G. Zheng, M. Bengtsson, K.-K. Wong, L. Yang, and B. Ottersten, "Distributed multicell beamforming with limited intercell coordination," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 728–738, Feb. 2011.

[10] K. T. Truong and R. W. Heath, "The viability of distributed antennas for massive MIMO systems," in *Proc. 2013 47th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2013, pp. 1318–1323.

[11] R. Rogalin, O. Y. Bursalioglu, H. Papadopoulos, G. Caire, A. F. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1815–1831, Apr. 2014.

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

[13] X. Wang, P. Zhu, and M. Chen, "Antenna location design for generalized distributed antenna systems," *IEEE Commun. Lett.*, vol. 13, no. 5, pp. 315–317, May 2009.

[14] E. Park, S. Lee, and I. Lee, "Antenna placement optimization for distributed antenna systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 7, pp. 2468–2477, July 2012.

[15] A. Yang, Y. Jing, C. Xing, Z. Fei, and J. Kuang, "Performance analysis and location optimization for massive MIMO systems with circularly distributed antennas," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5659–5671, Oct. 2015.

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

[17] J. Wang, H. Zhu, and N. J. Gomes, "Distributed antenna systems for mobile communications in high speed trains," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 675–683, May 2012.

[18] E. Koyuncu, "Performance gains of optimal antenna deployment for massive MIMO systems," in *Proc. 2017 IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[19] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and B. D. Rao, "Performance of cell-free massive MIMO systems with MMSE and LSFD receivers," in *Proc. 2016 50th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2016, pp. 203–207.

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

[21] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, 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.

[22] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[23] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/June 2020.

[24] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: A view on 5G cellular technology beyond 3GPP release 15," *IEEE Access*, vol. 7, pp. 127 639–127 651, Sept. 2019.

[25] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, "A tutorial on IEEE 802.11ax high efficiency WLANs," *IEEE Commun. Surveys Tut.*, vol. 21, no. 1, pp. 197–216, Firstquarter 2019.

[26] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. and Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[27] M. Alonzo, S. Buzzi, A. Zappone, and C. D'Elia, "Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 651–663, Sept. 2019.

[28] N. Rajapaksha, K. B. S. Manosha, N. Rajatheva, and M. Latva-Aho, "Deep learning-based power control for cell-free massive MIMO networks," in *Proc. 2021 IEEE Int. Conf. Commun. (ICC)*, June 2021, pp. 1–7.

[29] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max–min SINR of cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2021–2036, Apr. 2019.

[30] M. Farooq, H. Q. Ngo, E.-K. Hong, and L.-N. Tran, "Utility maximization for large-scale cell-free massive MIMO downlink," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 7050–7062, Oct. 2021.

[31] M. Alonzo and S. Buzzi, "Cell-free and user-centric massive MIMO at millimeter wave frequencies," in *Proc. 2017 IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.

[32] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[33] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. 2019 IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[34] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.

[35] Y. Al-Eryani, M. Akrout, and E. Hossain, "Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1028–1042, Apr. 2021.

[36] M. Bashar, H. Q. Ngo, K. Cumanan, A. G. Burr, P. Xiao, E. Björnson, and E. G. Larsson, "Uplink spectral and energy efficiency of cell-free massive MIMO with optimal uniform quantization," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 223–245, Jan. 2021.

[37] S. Buzzi, C. D'Andrea, M. Fresia, Y.-P. Zhang, and S. Feng, "Pilot assignment in cell-free massive MIMO based on the hungarian algorithm," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 34–37, Jan. 2021.

[38] Y. Zhang, B. Di, H. Zhang, J. Lin, C. Xu, D. Zhang, Y. Li, and L. Song, "Beyond cell-free MIMO: Energy efficient reconfigurable intelligent surface aided cell-free MIMO communications," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 2, pp. 412–426, June 2021.

[39] T. T. Vu, H. Quoc Ngo, T. L. Marzetta, and M. Matthaiou, "How does cell-free massive MIMO support multiple federated learning groups?" in *Proc. 2021 IEEE 22nd Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Sept. 2021, pp. 401–405.

[40] G. P. Villardi, G. Thadeu Freitas de Abreu, and H. Harada, "TV white space technology: Interference in portable cognitive emergency network," *IEEE Veh. Technol. Mag.*, vol. 7, no. 2, pp. 47–53, June 2012.

[41] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[42] T. Ding, M. Ding, G. Mao, Z. Lin, A. Y. Zomaya, and D. López-Pérez, "Performance analysis of dense small cell networks with dynamic TDD," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9816–9830, Oct. 2018.

[43] G. P. Villardi, C. Sum, C. Sun, Y. Alemseged, Z. Lan, and H. Harada, "Efficiency of dynamic frequency selection based coexistence mechanisms for TV white space enabled cognitive wireless access points," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 69–75, Dec. 2012.

[44] G. P. Villardi, Y. D. Alemseged, C. Sun, C. Sum, T. H. Nguyen, T. Baykas, and H. Harada, "Enabling coexistence of multiple cognitive networks in TV white space," *IEEE Wireless Commun.*, vol. 18, no. 4, pp. 32–40, Aug. 2011.

[45] X. Chen, D. Guo, and J. Grosspietsch, "The public safety broadband network: A novel architecture with mobile base stations," in *Proc. 2013 IEEE Int. Conf. Commun. (ICC)*, June 2013, pp. 3328–3332.

[46] S. T. Abraha, D. F. Castellana, X. Liang, A. Ng'oma, and A. Kobyakov, "Experimental study of distributed massive MIMO (DM-MIMO) in in-building fiber-wireless networks," in *Proc. Opt. Fiber Commun. Conf. Expo. (OFC)*, 2018.

[47] J. Guo and H. Jafarkhani, "Sensor deployment with limited communication range in homogeneous and heterogeneous wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6771–6784, Oct. 2016.

[48] E. Koyuncu and H. Jafarkhani, "On the minimum average distortion of quantizers with index-dependent distortion measures," *IEEE Trans. on Signal Process.*, vol. 65, no. 17, pp. 4655–4669, Sep. 2017.

[49] J. Guo, E. Koyuncu, and H. Jafarkhani, "A source coding perspective on node deployment in two-tier networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3035–3049, July 2018.

[50] S. Karimi-Bidhendi, J. Guo, and H. Jafarkhani, "Using quantization to deploy heterogeneous nodes in two-tier wireless sensor networks," in *Proc. 2019 IEEE Int. Symp. Inf. Theory (ISIT)*, July 2019, pp. 1502–1506.

[51] S. Karimi-Bidhendi, J. Guo, and H. Jafarkhani, "Energy-efficient node deployment in heterogeneous two-tier wireless sensor networks with limited communication range," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 40–55, Jan. 2021.

[52] B. Galkin, J. Kibilda, and L. A. DaSilva, "Deployment of UAV-mounted access points according to spatial user locations in two-tier cellular networks," in *Proc. 2016 Wireless Days (WD)*, Mar. 2016, pp. 1–6.

[53] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," in *Proc. 2016 IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–5.

[54] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.

[55] L. Zhang, Q. Fan, and N. Ansari, "3-D drone-base-station placement with in-band full-duplex communications," *IEEE Commun. Lett.*, vol. 22, no. 9, pp. 1902–1905, Sept. 2018.

[56] C. Lai, C. Chen, and L. Wang, "On-demand density-aware UAV base station 3D placement for arbitrarily distributed users with guaranteed data rates," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 913–916, June 2019.

[57] L. Xie, J. Xu, and R. Zhang, "Throughput maximization for UAV-enabled wireless powered communication networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1690–1703, Apr. 2019.

[58] J. Guo, P. Walk, and H. Jafarkhani, "Optimal deployments of UAVs with directional antennas for a power-efficient coverage," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5159–5174, Aug. 2020.

[59] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.

[60] E. Nayebi and B. D. Rao, "Access point location design in cell-free massive MIMO systems," in *Proc. 2018 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 985–989.

[61] E. Nayebi, "TDD massive MIMO systems: Channel estimation, power optimization, and access point location design," Ph.D. dissertation, University of California San Diego, 2018.

[62] Z. Yun and M. F. Iskander, "Ray tracing for radio propagation modeling: Principles and applications," *IEEE Access*, vol. 3, pp. 1089–1100, 2015.

[63] E. Ostlin, H. Zepernick, and H. Suzuki, "Macrocell path-loss prediction using artificial neural networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 2735–2747, July 2010.

[64] E. Weiszfeld, "Sur le point pour lequel la somme des distances de n points donnes est minimum," *Tohoku Math. J.*, vol. 43, pp. 355–386, 1937.

[65] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York, NY, USA: Academic, 1970.

[66] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[67] A. Garcia-Rodriguez, G. Geraci, D. Lopez-Perez, L. G. Giordano, M. Ding, and E. Bjornson, "The essential guide to realizing 5G-connected UAVs with massive MIMO," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 84–90, Oct. 2019.

[68] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2125–2140, Apr. 2019.

[69] W. Mei and R. Zhang, "Aerial-ground interference mitigation for cellular-connected UAV," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 167–173, Feb. 2021.

[70] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1647–1650, Aug. 2016.

[71] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, Aug. 2017.

[72] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[73] J. Lyu, Y. Zeng, and R. Zhang, "UAV-aided offloading for cellular hotspot," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3988–4001, June 2018.

[74] Y. Huang, W. Mei, J. Xu, L. Qiu, and R. Zhang, "Cognitive UAV communication via joint maneuver and power control," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7872–7888, Nov. 2019.

[75] J. Qin, Z. Wei, C. Qiu, and Z. Feng, "Edge-prior placement algorithm for UAV-mounted base stations," in *Proc. 2019 IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[76] X. Guo, B. Li, K. Liu, and R. Zhang, "Joint placement optimization and RNC in UAV-based wireless multicast networks," in *Proc. 2020 IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, June 2020, pp. 1–6.

[77] Y. Zhang and L. Dai, "Joint optimization of placement and coverage of access points for IEEE 802.11 networks," in *Proc. 2020 IEEE Int. Conf. Commun. (ICC)*, June 2020, pp. 1–7.

[78] J. Guo, P. Walk, and H. Jafarkhani, "Optimal deployments of UAVs with directional antennas for a power-efficient coverage," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5159–5174, Aug. 2020.

[79] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," 2020, arXiv:2012.03158.

[80] G. R. Gopal, E. Nayebi, G. P. Villardi, and B. D. Rao, "Modified vector quantization for small-cell access point placement with inter-cell interference," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6387–6401, Aug. 2022.

[81] H. Yamamura, K. Kaneda, and Y. Mizobata, "Communication problems after the great east japan earthquake of 2011," *Disaster Med Public Health Prep.*, vol. 8, no. 4, pp. 293–296, Aug. 2014.

[82] K. Hill. (2019, Oct.) Wildfires, power shut-offs impact cellular networks in california. [Online]. Available: https://www.rcrwireless.com/20191031/network-infrastructure/wildfires-power-shut-offs-impact-cellular-networks-in-california

[83] A. Jalali, "On cell breathing in CDMA networks," in *Proc. 1998 IEEE Int. Conf. Commun. (ICC)*, vol. 2, June 1998, pp. 985–988.

[84] T.-C. Tsai and C.-F. Lien, "IEEE 802.11 hot spot load balance and QoS-maintained seamless roaming," in *Proc. Nat. Comput. Symp. (NCS)*, Jan. 2003.

[85] Y. Bejerano and S. Han, "Cell breathing techniques for load balancing in wireless LANs," *IEEE Trans. Mobile Comput.*, vol. 8, no. 6, pp. 735–749, June 2009.

[86] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[87] J. Lyu and R. Zhang, "Network-connected UAV: 3-D system modeling and coverage performance analysis," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7048–7060, Aug. 2019.

[88] E. Yanmaz, S. Hayat, J. Scherer, and C. Bettstetter, "Experimental performance analysis of two-hop aerial 802.11 networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2014, pp. 3118–3123.

[89] N. Ahmed, S. S. Kanhere, and S. Jha, "On the importance of link characterization for aerial wireless sensor networks," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 52–57, May 2016.

[90] M. Sadek, A. Tarighat, and A. H. Sayed, "Active antenna selection in multiuser MIMO communications," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1498–1510, Apr. 2007.

[91] B. O. Lee, H. W. Je, O. Shin, and K. B. Lee, "A novel uplink MIMO transmission scheme in a multicell environment," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 4981–4987, October 2009.

[92] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1711–1721, May 2007.

[93] S. Hur, B. C. Jung, and B. D. Rao, "Sum rate enhancement by maximizing SGINR in an opportunistic interference alignment scheme," in *Proc. 2011 45th Asilomar Conf. Signals, Syst., and Comput.*, Nov. 2011, pp. 354–358.

[94] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

[95] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, ser. SODA '07. USA: Society for Industrial and Applied Mathematics, Jan. 2007, p. 1027–1035.

[96] C. Diaz-Vilor, A. Lozano, and H. Jafarkhani, "On the deployment problem in cell-free UAV networks," in *Proc. 2021 IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[97] Y.-H. Zhu, G. Callebaut, H. Çalık, L. Van der Perre, and F. Rottenberg, "Energy efficient access point placement for distributed massive MIMO," *Netw.*, vol. 2, no. 2, pp. 288–310, May 2022. [Online]. Available: https://www.mdpi.com/2673-8732/2/2/19

[98] L. Wang and Q. Zhang, "Cell-free massive MIMO with UAV access points: UAV location optimization," in *Proc. 2022 IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2022, pp. 262–267.

[99] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[100] A. Subramaniam and B. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 130–142, Mar. 2003.

[101] A. K. Jain, *Fundamentals of Digital Image Processing*. USA: Prentice-Hall, Inc., 1989.

[102] H. Sun and S. Wang, "Measuring the component overlapping in the Gaussian mixture model," *Data Mining Knowl. Discovery*, vol. 23, pp. 479–502, Nov. 2011.

[103] M. Sabin and R. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inf. Theory*, vol. 32, no. 2, pp. 148–155, Mar. 1986.

[104] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[105] S. Venkatesan, A. Lozano, and R. Valenzuela, "Network MIMO: Overcoming intercell interference in indoor wireless systems," in *Proc. 2007 41th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2007, pp. 83–87.

[106] M. Bashar, H. Q. Ngo, A. G. Burr, D. Maryopi, K. Cumanan, and E. G. Larsson, "On the performance of backhaul constrained cell-free massive MIMO with linear receivers," in *Proc. 2018 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 624–628.

[107] H. He, X. Yu, J. Zhang, S. H. Song, and K. B. Letaief, "Cell-free massive MIMO for 6G wireless communication networks," 2021. [Online]. Available: arXiv:2110.07309.

[108] J. Košmerl and A. Vilhar, "Base stations placement optimization in wireless networks for emergency communications," in *Proc. 2014 IEEE Int. Conf. Commun. (ICC) Workshops*, June 2014, pp. 200–205.

[109] G. R. Gopal and B. D. Rao, "Vector quantization methods for access point placement in cell-free massive MIMO systems," 2022. [Online]. Available: arXiv:2211.13356.

[110] A. Oka and L. Lampe, "Distributed target tracking using signal strength measurements by a wireless sensor network," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 7, pp. 1006–1015, Aug. 2010.

[111] M. Z. Karakusak, H. Kivrak, H. F. Ates, and M. K. Ozdemir, "RSS-based wireless LAN indoor localization and tracking using deep architectures," *Big Data Cogn. Comput.*, vol. 6, no. 3, Aug. 2022.

[112] Y. Yuan, J. Zhao, C. Qiu, and W. Xi, "Estimating crowd density in an RF-based dynamic environment," *IEEE Sensors J.*, vol. 13, no. 10, pp. 3837–3845, Oct. 2013.

[113] S. Di Domenico, M. De Sanctis, E. Cianca, P. Colucci, and G. Bianchi, "LTE-based passive device-free crowd density estimation," in *Proc. 2017 IEEE Int. Conf. Commun. (ICC)*, July 2017, pp. 1–6.

[114] S. Denis, R. Berkvens, B. Bellekens, and M. Weyn, "Large scale crowd density estimation using a sub-Ghz wireless sensor network," in *Proc. 2018 IEEE 29th Ann. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sept. 2018, pp. 849–855.

[115] S. Georgievska, P. Rutten, J. Amoraal, E. Ranguelova, R. Bakhshi, B. L. de Vries, M. Lees, and S. Klous, "Detecting high indoor crowd density with Wi-Fi localization: A statistical mechanics approach," *J. Big Data*, vol. 6, no. 31, Mar. 2019.

[116] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, June 2009, pp. 107–114.

[117] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016, pp. 589–597.

[118] K.-W. Yang, Y.-Y. Huang, J.-W. Huang, Y.-R. Hsu, C.-L. Wan, H.-H. Shuai, L.-C. Wang, and W.-H. Cheng, "Improving crowd density estimation by fusing aerial images and radio signals," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3, Mar. 2022.

[119] G. Charan and A. Alkhateeb, "User identification: The key enabler for multi-user vision-aided wireless communications," 2022. [Online]. Available: arxiv:2210.15652.

[120] F. Thomas and L. Ros, "Revisiting trilateration for robot localization," *IEEE Trans. Robot.*, vol. 21, no. 1, pp. 93–101, Feb. 2005.

[121] Y. Zhou, "An efficient least-squares trilateration algorithm for mobile robot localization," in *Proc. 2009 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 3474–3479.

[122] B. Yang, L. Guo, R. Guo, M. Zhao, and T. Zhao, "A novel trilateration algorithm for RSSI-based indoor localization," *IEEE Sensors J.*, vol. 20, no. 14, pp. 8164–8172, July 2020.

[123] H. M. Le, J.-P. Rossi, and D. Slock, "A geometric interpretation of trilateration for RSS-based localization," in *Proc. 2020 28th European Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 1797–1801.

[124] H. M. Le, D. Slock, and J.-P. Rossi, "Deep learning to robustify a geometric interpretation of trilateration for 3D RSS-based localization," in *Proc. 2022 IEEE 21st Mediterranean Electrotechnical Conf. (MELECON)*, June 2022, pp. 307–312.

[125] E. Björnson and L. Sanguinetti, "A new look at cell-free massive MIMO: Making it practical with dynamic cooperation," in *Proc. 2019 IEEE 30th Ann. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sept. 2019, pp. 1–6.

[126] J. Wang, L. Dai, L. Yang, and B. Bai, "Clustered cell-free networking: A graph partitioning approach," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[127] M. Yemini and A. J. Goldsmith, "Virtual cell clustering with optimal resource allocation to maximize capacity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5099–5114, Aug. 2021.

[128] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. USA: Prentice-Hall, Inc., 1988.

[129] M. Yemini, E. Erkip, and A. J. Goldsmith, "Interference reduction in virtual cell optimization," in *Proc. 2021 55th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2021, pp. 18–25.

[130] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," 2011. [Online]. Available: arXiv:1109.2378.