L'amélioration des performances des systèmes sans fil 5G par groupements adaptatifs des utilisateurs

Salah Eddine Hajri

université
PARIS-SACLAY

CentraleSupélec

NNT : 2018SACLC029

L'amélioration des performances des systèmes sans fil 5G par groupements adaptifs des utilisateurs

(Performance improvement of 5G Wireless systems through adaptive grouping of users)

Thèse de doctorat de l'Université Paris-Saclay
Préparée à CENTRALESUPÉLEC

École doctorale n°580 : sciences et technologies de l'information
et de la communication (STIC)

Spécialité de doctorat: réseaux, information et communications

Thèse présentée et soutenue à Gif-sur-Yvette, le 09 Avril 2018, par

M. Salah Eddine HAJRI

Composition du Jury :

M. Michel Kieffer Professeur, Université ParisSud (Paris Saclay)	Président
M. Dirk Slock Professeur, Eurecom	Rapporteur
M. Mehdi Bennis Professeur associé, University of Oulu	Rapporteur
Mme Inbar Fijalkow Professeure, Ensea - Université Cergy-Pontoise	Examinatrice
M. Gerhard Wunder Professeur, Freie Universität Berlin (FUB)	Examinateur
M. Mohamed-Slim Alouini Professeur, King Abdullah University of Science and Technology	Examinateur
M. Mohamad ASSAAD Professeur, CentraleSupélec, TCL Chair on 5G	Directeur de thèse

Thèse de doctorat

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies
de l'information
et de la communication (STIC)

Title: L'amélioration des performances des systèmes sans fil 5G par groupements adaptatifs des utilisateurs

Mots clés : Mise en cache proactive, MIMO massif, apprentissage automatique, réseaux cellulaires, 5G

Résumé: 5G est prévu pour s'attaquer, en plus d'une augmentation considérable du volume de trafic, la tâche de connecter des milliards d'appareils avec des exigences de service hétérogènes. Afin de relever les défis de la 5G, nous préconisons une utilisation plus efficace des informations disponibles, avec plus de sensibilisation par rapport aux services et aux utilisateurs, et une expansion de l'intelligence du RAN. En particulier, nous nous concentrons sur deux activateurs clés de la 5G, à savoir le MIMO massif et la mise en cache proactive.

Dans le troisième chapitre, nous nous concentrons sur la problématique de l'acquisition de CSI dans MIMO massif en TDD. Pour ce faire, nous proposons de nouveaux schémas de regroupement spatial tels que, dans chaque groupe, une couverture maximale de la base spatiale du signal avec un chevauchement minimal entre les signatures spatiales des utilisateurs est obtenue. Ce dernier permet d'augmenter la densité de connexion tout en améliorant l'efficacité spectrale.

MIMO massif en TDD est également au centre du quatrième chapitre. Dans ce cas, en se basant sur les différents taux de vieillissement des canaux sans fil, la périodicité d'estimation de CSI est supplémentaire. Nous le faisons en proposant un exploité comme un degré de liberté supplémentaire. Nous le faisons en proposant une adaptation dynamique de la trame TDD en fonction des temps de cohérence des canaux hétérogènes. Les stations de bases MIMO massif sont capables d'apprendre la meilleure politique d'estimation sur le uplink pour de longues périodes. Comme les changements de canaux résultent principalement de la mobilité de l'appareil, la connaissance de l'emplacement est également incluse dans le processus d'apprentissage. Le problème de planification qui en a résulté a été modélisé comme un POMDP à deux échelles temporelles et des algorithmes efficaces à faible complexité ont été fournis pour le résoudre.

Le cinquième chapitre met l'accent sur la mise en cache proactive. Nous nous concentrons sur l'amélioration de l'efficacité énergétique des réseaux dotes de mise en cache en exploitant la corrélation dans les modèles de trafic en plus de la répartition spatiale des demandes. Nous proposons un cadre qui établit un compromis optimal entre la complexité et la véracité dans la modélisation du comportement des utilisateurs grâce à la classification adaptative basée sur la popularité du contenu. Il simplifie également le problème du placement de contenu, ce qui se traduit par un cadre d'allocation de contenu rapidement adaptable et économe en énergie.

universite PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies
de l'information
et de la communication (STIC)

Title: Performance improvement of 5G Wireless Systems through adaptive grouping of users

Keywords: Massive multiple-input multiple-output 5Massive MIMO), proactive caching, machine learning, cellular networks, 5G

Abstract: 5G is envisioned to tackle, in addition to a considerable increase in traffic volume, the task of connecting billions of devices with heterogeneous service requirements. In order to address the challenges of 5G, we advocate a more efficient use of the available information, with more service and user awareness, and an expansion of the RAN intelligence. In particular, we focus on two key enablers of 5G, namely massive MIMO and proactive caching.

In the third chapter, we focus on addressing the bottleneck of CSI acquisition in TDD Massive MIMO. In order to do so, we propose novel spatial grouping schemes such that, in each group, maximum coverage of the signal's spatial basis with minimum overlapping between user spatial signatures is achieved. The latter enables to increase connection density while improving spectral efficiency.

TDD Massive MIMO is also the focus of the fourth chapter. Therein, based on the different rates of wireless channels aging, CSI estimation periodicity is exploited as an additional DoF. We do so by proposing a dynamic adaptation of the TDD frame based on the heterogeneous channels coherence times. The Massive MIMO BSs are enabled to learn the best uplink training policy for long periods. Since channel changes result primarily from device mobility, location awareness is also included in the learning process. The resulting planning problem was modeled as a two-time scale POMDP and efficient low complexity algorithms were provided to solve it.

The fifth chapter focuses on proactive caching. We focus on improving the energy efficiency of cache-enabled networks by exploiting the correlation in traffic patterns in addition to the spatial repartition of requests. We propose a framework that strikes the optimal trade-off between complexity and truthfulness in user behavior modeling through adaptive content popularity-based clustering. It also simplifies the problem of content placement, which results in a rapidly adaptable and energy efficient content allocation framework.

... this thesis is dedicated to my parents
for their limitless support...

Acknowledgments

This PhD was an opportunity to challenge and surpass myself. I worked hard and I was lucky to be able to count on marvelous people whom were there when I needed help. My sincere thanks go to:

... **Mohamad Assaad**, my supervisor, who always come up with novel ideas, for his support, valuable advice, technical guidance and generosity of time.

... all the brilliant collaborators from **TCL** whom I had the opportunity to work with, for their valued council and technical guidance, namely, **Pierre BONNARD** , **Salim Umer**, **Joel VARIN**, **Bruno JECHOUX**,

... all the great collaborators I had the opportunity to work with, namely, **Ali Maatouk**, **Hikmet Sari**, **Serdar Sezginer**, **Maialen Larrañaga**, **Thomas Bonald**, **Anthony Ephremides**, **Giuseppe Caire**,

... all my friends and colleagues with whom I was glad to share this experience in CentraleSupélec, especially,

Kenza Hamidouche, **Ejder Baştuğ**, **Gil Katz**, **Matha Deghel**, **Hafiz Tiomoko Ali**, **Jerome Gaveau**, **Asma Ghorbel**, **Chien-Chun Cheng**, **Zheng Chen**, **Bakarime Diomande**, **Chao He**, **Clement FEUTRY**, **Victor EXPOSITO**, **Julien FLOQUET**, **Rita IBRAHIM**, **Richard COMBES**, **Juwendo Denis** , **German Bassi**, **Ali Maatouk**, **Maialen Larrañaga**, **Zheng LI**, **Wenjie LI**, and **Khac Hoang NGO**.

... all my friends and nice people with whom I have interacted in CentraleSupélec during these last three years, especially, **Catherine Magnet** , **Jose Fonseca**, **Huu-Hung Vuong** , **Elodie Ledoux**, and **Anne Batalie**,

Thank you all.

Contents

Glossary

AIC Akaike information criterion. ii, 8, 12, 73

AP access point. 1, 78

AR augmented reality. 2

AS Angular spread. 30

BS base station. ii, 3–7, 10, 15–21, 24, 30, 53, 59, 65, 78

CDF Cumulative density function. 30, 32

CoMP coordinated multi-point transmission/reception. 6

CRAN cloud radio access network. 3, 78

CS Compressed sensing. 6

CSI Channel state information. ii, 4–7, 10, 11, 15, 17–19, 21, 24, 35–43, 45, 47, 49, 53, 55, 56, 77–79, 85, 89

CSIT channel state information at the transmitter side. 6

D2D device to device. 9

DFT discrete Fourier transform. 16, 18, 19, 21, 23–26, 32, 33

DL downlink. 4–7, 17–19, 35, 37, 39, 59, 63, 65, 66

DOA Direction of Arrival. 17

DoF Degrees of freedom. 3, 10, 11, 16, 21

DPC dirty-paper-coding. 7

E2E End to End. 7, 8

EE Energy efficiency. ii, 1, 4, 7, 9, 10, 12, 59, 63–68, 72, 73

eMBB enhanced Mobile Broadband. 1

EVD eigenvalue decomposition. 5

FDD Frequency division duplexing. 4–6, 15, 20, 21, 35

GPS global positioning system. 43, 44

IMT International Mobile Telecommunications. 1

IoT Internet of Things. 1, 2

ITU International Telecommunication Union. 1

KPI Key performance indicator. ii, 10, 39, 78

LTE long term evolution. 1, 57, 77

MAP maximum a-posteriori. 5

Massive MIMO Massive multiple-input multiple-output. ii, 2–7, 10, 11, 15–18, 20, 21, 30, 32, 33, 77, 78

MDP Markov Decision Process. 47, 51

MMSE minimum mean square error. 6, 7, 19, 37, 39, 90

mm-wave Millimeter-Wave. 2, 6, 7

MRC Maximum ratio combining. 6, 7, 38–40, 45, 53

MSE Mean square error. 30

MTC Machine Type Communications. 1, 2

MU-MIMO Multi-user multiple-input multiple-output. 3, 5

NOMA Non-Orthogonal Multiple Access. 2

NR New Radio. 2

OFDM Orthogonal frequency-division multiplexing. 2, 7

OFDMA orthogonal frequency-division multiple access. 58

OTDOA observed time difference of arrival. 43

PAPR peak-to-average-power-ratio. 7

POMDP Partially Observable Markov Decision Process. 12, 46, 47, 51, 78

PPP Poisson point process. 58, 64, 68, 72, 95, 96

QoE Quality of Experience. ii, 8

RAN radio access network. ii, 3, 7, 8, 77

RF radio frequency. 4

SBS small base station. ii, xv, 9, 10, 58, 59, 64–73, 75, 95

SCN small cell network. ii, 8, 12

SE Spectral efficiency. ii, 3–5, 11, 12, 15, 16, 18, 30, 32, 33, 53, 77

SINR Signal-to-interference-plus-noise ratio. 18, 20, 27, 41, 42, 45, 53, 58, 65, 90

SNR Signal-to-noise ratio. 7, 30

SON Self Organizing Network. 3

TDD Time division duplexing. ii, 4, 5, 10, 11, 15–17, 20, 21, 30, 32, 35–39, 42, 47, 52, 53, 55, 77–79

UL uplink. ii, 4–7, 10, 11, 15–21, 23, 24, 28, 30, 35–43, 46, 47, 49, 52, 53, 55, 59, 77–79, 85, 89, 91, 92

ULA Uniform linear array. 16–18

URLLC Ultra-reliable low latency communication. 1

VR virtual reality. 2

ZF Zero Forcing. 6, 7, 38, 39, 41, 45, 53

List of Figures

List of Tables

Chapter 1

Resumé (French)

1.1 Contexte et motivation

Les communications mobiles ont été fondamentales dans la production de nos sociétés connectées contemporaines. Des anciens systèmes mobiles analogiques aux réseaux long term evolution (LTE), plus sophistiqués [1], les progrès dans les systèmes sans fil ont radicalement changé la façon dont les humains accèdent et échangent des informations.

Actuellement, les communications sans fil sont à une croisée de chemins. En effet, la demande de capacité toujours croissante et la prolifération d'appareils intelligents, avec des applications nécessitant des débits élevés, nécessitent de nouvelles générations de réseaux plus efficaces pour permettre une augmentation substantielle des performances. Les systèmes de communication mobile de cinquième génération (5G) émergent à grande vitesse pour répondre à un large éventail de défis apportés par la soif de nos sociétés actuelles et futures pour les communications sans fil. La 5G doit s'attaquer, en plus d'une augmentation du volume de trafic, au défi de connecter des milliards d'appareils avec des besoins de service hétérogènes. Les réseaux de 5 G devraient fournir des améliorations telles que [2]:

- 10 fois plus de débits expérimenté: l'ère des débits de pointe plus uniformes et multi-Gbps.

- 10 fois moins de temps de latence: les niveaux de latence devraient être aussi bas que 1 ms.

- 10 fois plus de densité de connexion: activation de la connectivité Internet of Things (IoT) avec peu de complexité et de surcharge de signalisation.

- Augmentation de 3 fois dans l'efficacité spectrale: une utilization plus efficace de la bande passante.

- 100 fois plus de capacité de trafic: réseaux très densifiés avec plus de points d'accés partout.

- 100 fois plus d'efficacité du réseau: réseaux énergétiquement efficace avec traitement de signal et matériel efficaces.

Ces objectifs de performance de haut niveau pour la 5G ont été développés dans le cadre de International Mobile Telecommunications (IMT)-2020, l'initiative International Telecommunication Union (ITU) pour définir la base de 5G. Ces exigences sont associées à trois cas d'utilisation majeurs, à savoir Haut débit mobile amélioré (enhanced Mobile Broadband (eMBB)), Communications massives de type machine (massifs Machine Type Communications (MTC)) et Communications ultra-fiables et à faible latence (Ultra-reliable low latency communication (URLLC)) [3], [4], [5]:

- Communications ultra-fiables et à faible latence (URLLC): Ce cas d'utilisation se concentre sur les services exigeants en termes de latence et de fiabilité. Il répond aux attentes de la fabrication industrielle contrôlée à distance, la chirurgie médicale à distance, des réseaux intelligents et de la conduite automatisée, etc.

- Haut débit mobile amélioré: Ce cas d'utilisation s'accompagne de nouvelles applications et exigences, en plus de la communication cellulaire conventionelle. Il vise à répondre aux exigences d'un mode de vie humain de plus en plus numérisé. Il se concentre sur les services à haut débit tels que la réalité virtuelle, réalité augmentée et le streaming vidéo.

- Communications massives de type machine: massive MTC se concentre sur les exigences d'un grand nombre d'appareils connectés avec une faible capacité de données et de faibles exigences en latence. Cela inclut les applications telles que les villes intelligentes, IoT, etc.

Les différents cas d'utilisation mentionnés ont des caractéristiques variables et parfois contradictoires. En effet, à titre d'exemple, 5G doit passer de la prise en charge de capteurs à faible débit (10 kbps) à de nouvelles expériences mobiles immersives à plusieurs Gbps. Cela signifie que les réseaux 5 G doivent pouvoir évoluer à travers divers services avec des appareils mobiles également divers. Ce faisant vient avec sa charge de difficultés.

1.2 5G: Une concentration de nouveaux paradigmes et de technologies innovantes

Répondre aux exigences mentionnées nécessite des changements radicaux dans le paradigme du réseau en plus d'innovations perturbatrices. Dans ce contexte, les réseaux 5 G peuvent faire appel à un large éventail de nouvelles technologies. Cela permet un saut dans les performances qui éclipse ses prédécesseurs. Ces innovations toucheront la transmission et la conception de la couche physique en plus d'introduire des bouleversements dans les couches supérieures du réseau. En fait, 5 G New Radio (NR) utilisera de nombreuses technologies clés afin d'atteindre de nouveaux niveaux de performance et d'efficacité. Les combinaisons de ces dernières étendront l'importance des communications mobiles et leurs permettront de jouer un rôle central dans un monde de cas d'utilisation changeants. Parmi les innovations potentielles dans la couche physique 5G, on peut citer [3]:

- Communications dans la plage des ondes millimétriques.
- Entrée multiple sortie multiple massif (mMIMO).
- Accès multiple non orthogonal (NOMA).
- Communications sans fil full-duplex.
- Agrégation de porteuse et modulations Multicarrier.
- Plus grand spectre.
- Communication Sidelink.
- Nouvelle forme d'onde et numérologie Orthogonal frequency-division multiplexing (OFDM) hétérogène.

En plus de ces améliorations dans la couche physique, les innovations de 5 G changeront la façon dont le réseautage est effectué. En fait, 5 G concentre un certain nombre de nouveaux paradigmes qui visent à permettre un réseau plus agile, automatique et intelligent dans chacune de ses opérations [5]. Parmi les principales innovations de réseautage à 5 G, on peut citer [3]:

- Réseau d'accès radio cloud (nuage).
- Récolte d'énergie.
- Accès sans fil vert.
- Réseau auto-organisé (SON).
- Réseau sans fil centré sur l'utilisateur (mise en cache proactive, etc.).

Dans cette thèse, nous nous sommes concentrés principalement sur deux technologies majeures qui vont permettre d'atteindre l'âge de 5 G, à savoir MIMO massif et la mise en cache proactive. L'objectif principal de la présente thèse est d'améliorer les performances des deux technologies en utilisant une optimisation intelligente basée sur la sensibilisation aux services et utilisateurs. Cet intérêt était basé sur une observation fondamentale. Les deux technologies peuvent être améliorées en optimisant les opérations du réseau en se basant sur des informations secondaires facilement accessibles dans les réseaux mobiles.

Tirant parti des connaissances sous-exploitées du côté de radio access network (RAN), telles que les statistiques spatiales, la propagation Doppler et la modélisation efficace de la popularité du contenu, les performances du réseau peuvent être considérablement améliorées. Ceci est réalisé grâce à l'incorporation d'algorithmes efficaces dans les procédures de réseaux.

Cette observation est la pierre angulaire de cette thèse et un soin particulier a été pris afin de développer des schémas permettant d'améliorer les performances du réseau avec le minimum possible de signalisation.

Dans les sections suivantes, nous donnons d'abord un aperçu des progrès passés et récents dans les deux technologies choisies. Nous présentons ensuite les grandes lignes de la thèse et de ses contributions.

1.3 MIMO massif pour 5G: Un bref historique et travaux connexes

Toute évolution dans les réseaux doit apporter une amélioration substantielle d'efficacité spectrale. Massive MIMO est une technologie qui rend cela possible pour 5 G car elle peut apporter une amélioration de dix fois dans l'efficacité spectrale [3]. Ce gain impressionnant est obtenu en utilisant une centaine d'éléments d'antenne bon marché dans les stations de bases. Cela permet le multiplexage spatial d'un nombre considérable de dispositifs mobiles.

L'idée originale est basée sur des effets statistiques à grande échelle qui résultent de l'augmentation drastique du nombre d'antennes base station (BS) (de l'ordre de centaines) [6]. Il en résulte, pratiquement, une réduction des impacts de l'évanouissement rapide, des interférences et du bruit additif. Plus important encore, il permet de concentrer l'énergie rayonnée sur les cibles prévues.

Par conséquent, par un traitement cohérent des signaux sur le réseau d'antennes BS, le précodage d'émission peut être utilisé pour concentrer chaque signal sur sa borne prévue et la combinaison de réception peut être utilisée pour discriminer les signaux de différents utilisateurs. Plus il y a d'antennes à la BS, plus la focalisation spatiale peut être fine. Cela permet de planifier beaucoup plus d'utilisateurs que ce qui est possible aujourd'hui, ce qui augmente énormément l'efficacité spectrale et la densité de connexion.

L'excès d'antennes BS entraîne une augmentation du nombre de flux de données qui peuvent être exploités pour desservir plus de terminaux, réduisant la puissance rayonnée, tout en augmentant le débit de données.

Massive MIMO peut aussi améliorer la fiabilité des liens grâce à la diversité spatiale et fournir plus de Degrees of freedom (DoF) dans le domaine spatial, ce qui améliore les performances, quel que soit le bruit des mesures. En raison du multiplexage spatial agressif qui en résulte, Massive MIMO peut fournir un gain impressionnant dans les performances du réseau en dirigeant simplement les ondes rayonnées dans les bonnes directions. Puisque l'énergie rayonnée est fortement concentrée sur des zones centrées sur l'utilisateur, Massive MIMO fournit des gains considérables dans EE [7], [8].

Les principaux avantages des systèmes Massive MIMO peuvent être résumés comme suit:

- **Gain d'efficacité spectrale élevé** Massive MIMO hérite des gains de Multi-user multiple-input multiple-output (MIMO) conventionnels mais à grande échelle, comme son nom l'indique. En effet, avec M antennes à la station de base, desservant K utilisateurs à une seule antenne, on obtient une diversité d'ordre M avec un gain de multiplexage de $min(M, K)$. Ces paramètres peuvent être ajustés afin d'améliorer l'efficacité spectrale de la communication.

- **Gain d'efficacité énergétique élevé**

  Massive MIMO atteint ses performances grâce à l'excès d'antennes BS combinées à un traitement cohérent. Cela permet de réduire considérablement la puissance d'émission. Par conséquent, grâce à une combinaison cohérente à la réception et à la formation de faisceau à la transmission, Energy efficiency (EE) peut être considérablement améliorée.

- **Traitement simple**

  Massive MIMO utilise des schémas de traitement de signaux simples mais efficaces (précodage et décodage linéaires dans les downlink (DL) et uplink (UL), respectivement). De plus, lorsque le nombre d'antennes est suffisamment grand, le durcissement du canal résultant simplifie encore le traitement du signal.

- **Robustesse et fiabilité accrues**

  Le grand nombre d'antennes BS procure plus de diversité. Cela se traduit par une meilleure fiabilité de liaison. De plus, à mesure que le nombre d'antennes augmente, le bruit additif, l'évanouissement à petite échelle et les interférences cellulaires sont vouées à disparaître.

- **Réduction des coûts dans les composants RF** Massive MIMO utilise un traitement cohérent qui permet de réduire la puissance rayonnée. Cela permet d'utiliser des amplificateurs peu-cher dans la gamme milli-Watt.

Cependant, il existe toujours un compromis entre performance réalisable et complexité. Les gains intéressants de Massive MIMO viennent avec leur part de défis:

- **Gestion des interférences multi-utilisateurs** Massive MIMO offre des gains considérables en termes de performances réseau. Cependant, certains utilisateurs peuvent voir leur canal souffrir d'un impact inégal d'interférence. Par conséquent, il peut être nécessaire de mettre en œuvre des schémas d'annulation d'interférence. L'alignement d'interférence [9], détection de multi-utilisateurs Maximum de vraisemblance [10] et le codage Dirty paper coding[11] peuvent être utilisé. Ces schémas souffrent d'un défaut majeur, à savoir une complexité de calcul élevée.
- **Acquisition de CSI**

  Le traitement cohérent est la pierre angulaire des systèmes Massive MIMO. Par conséquent, une estimation Channel state information (CSI) précise est requise. Cela peut être très difficile dans les modes Frequency division duplexing (FDD) et Time division duplexing (TDD), étant donné l'échelle du système (nombre d'utilisateurs, nombre d'antennes BS). La mobilité des utilisateurs a également un impact important car elle définit la corrélation entre le CSI et la réalisation réelle du canal.
- **planification (Scheduling)** Massive MIMO est prévu pour gérer un grand nombre de périphériques connectés. Avec les exigences de 5 G de connexion haute densité, la planification des utilisateurs est d'une importance primordiale. De plus, lorsque les mêmes ressources temps-fréquence sont partagées, la sélection des utilisateurs qui peuvent être actifs simultanément peut considérablement modifier les performances du système.

En raison ces avantages, Massive MIMO a fait l'objet d'une attention de plus en plus importante de la part de la communauté scientifique. Cela a abouti à une littérature riche qui traite différents aspects de ce concept. Dans ce qui suit, nous résumons certains des travaux sur Massive MIMO en fonction de leurs similitudes et de leurs directions.

1.3.1 Méthodes d'estimation CSI:

La pierre angulaire de Massive MIMO est un traitement de signal cohérent qui se base sur des connaissances précises et opportunes du CSI. Dans ce qui suit, nous donnons un aperçu des méthodes d'estimation du CSI dans les systèmes Massive MIMO en plus de la caractérisation des défis les plus importants liés à l'acquisition de CSI.

Systèmes TDD et contamination pilote

Dans les systèmes TDD Massive MIMO, les estimations de CSI sont obtenues en utilisant la réciprocité des canaux et la formation (training) UL. Dans ces systèmes, seuls les BSs doivent avoir une connaissance CSI afin de précoder et de décoder de façon cohérente les signaux multi-utilisateurs. La quantité de ressources d'entraînement temps-fréquence dépend du nombre d'antennes des utilisateurs.

TDD est considéré comme plus approprié pour les opérations Massive MIMO car cela implique que l'estimation du canal doit être effectuée dans une seule direction, et peut ensuite être utilisée dans les deux directions. Cet avantage ne peut pas être négligé, car cela signifie que les frais généraux de formation ne sont fonction que du nombre d'utilisateurs.

Néanmoins, en raison de l'intervalle de cohérence limité, la dimension d'apprentissage est restreinte et les mêmes séquences pilotes doivent être réutilisées, ce qui entraîne une contamination pilote [6], [13]. Ce phénomène a été identifié comme un facteur limitant majeur des performances de Massive MIMO et a attiré une attention considérable dans les travaux précédents [93]. Plusieurs méthodes ont été proposées afin de réduire ou, mieux encore, d'éliminer l'impact de la contamination pilote dans les systèmes TDD Massive MIMO. Ces méthodes sont basées sur les pilotes ou sur les sous-espace spatials des canaux [93].

Le changement de temps pour la transmission pilote a été proposé dans [14], [15], comme moyen de réduire la contamination pilote. L'idée principale était de décaler la transmission du pilote dans le temps afin que les utilisateurs dans différentes cellules transmettent à des heures qui ne se chevauchent pas. Les résultats montrent que le résultat du protocole est une élimination efficace de la contamination du pilote.

Dans [16], Ashikhmin et Marzetta ont proposé une méthode de précodage de contamination pilote (PCP) basée sur les coefficients d'évanouissement lent. La méthode proposée nécessite un certain niveau de coopération entre les BSs pour construire les matrices PCP. Les résultats montrent que cette méthode peut fournir des gains non négligeables en Spectral efficiency (SE). Ce travail a été étendu dans [17] en proposant un pré-codage basée sur l'évanouissement à grande échelle (LSFP) et un décodage à évanouissement à grande échelle (LSFD) dans le régime d'un nombre fini d'antennes BS. Les résultats ont montré un gain intéressant dans le taux d'interruption de 5 $\%$. D'un autre côté, les approches basées sur le sous-espace améliorent la précision de l'estimation de CSI en exploitant les statistiques d'ordre supérieur du signal. Dans [18], les auteurs ont montré que, en utilisant la corrélation spatiale des canaux et un précodage / combinaison adéquat, la capacité du système Massive MIMO peut augmenter sans limite en fonction du nombre de BS antennes. Ceci est rendu possible en exploitant l'indépendance linéaire entre les matrices de covariance des canaux des utilisateurs copilotes. Dans [19], la décomposition en valeurs propres (eigenvalue decomposition (EVD)) des matrices de covariance est implémenté afin d'obtenir les estimations de CSI. Afin de diminuer les erreurs, EVD est combiné avec le moindre carré itératif. Les auteurs ont montré que la méthode EVD permet d'atténuer l'impact de la contamination des pilotes et surpasse les techniques d'estimation classiques du CSI. Dans [20, 21], une détection aveugle a été proposée. L'idée principale était d'exploiter la décomposition en valeurs singulières afin de discriminer les signaux des utilisateurs. Les résultats ont montré que la connaissance du sous-espace engendré par chaque vecteur de canal est suffisante pour obtenir des estimations précises de CSI par simple projection. Néanmoins, cette approche suppose que tous les canaux souhaités sont plus fort que tous les canaux interférents, ce qui ne tiennent pas toujours dans la pratique. Neumann et al. [22] a proposé un critère maximum a-posteriori (MAP) pour l'estimation du canal subspatial afin de résoudre ce problème. Bien qu'une amélioration des performances ait été remarquée, l'utilisation de MAP se fait au prix d'une complexité accrue. Dans [23], une projection itérative des moindres carrés avec une estimation en diagonale a été proposée afin de résoudre le problème de la contamination des pilotes.

Systèmes FDD et retour d'information

Dans les systèmes FDD, puisque UL et DL utilisent différentes bandes de fréquence, CSI des deux liens doivent être estimés. UL CSI est obtenu en permettant aux utilisateurs d'envoyer différentes séquences pilotes. Dans les DL, les estimations de CSI sont obtenues en utilisant la formation DL suivie d'un retour d'information explicites ou implicites des CSI. A mesure que le nombre d'antennes BS augmente, l'estimation du canal FDD devient très problématique puisque le surdébit de retour CSI évolue linéairement avec le nombre d'antennes du système [24].

L'activation des systèmes FDD Massive MIMO ne peut être effectuée que si ce problème est corrigé. Dans [24, 25], JSDM pour MU-MIMO DL a été étudié. JSDM est un système qui vise à servir les utilisateurs en les regroupant, de sorte que les utilisateurs d'un groupe aient des covariances de canaux à peu près similaires, alors que les utilisateurs de differents groupes ont des espaces propres de covariance orthogonaux. JSDM a été conçu, à l'origine, pour les systèmes FDD Massive MIMO, sans tenir compte des interférences entre cellules. Elle permet de réduire le surcoût de retour CSI dans les systèmes FDD tout en n'encaissant aucune perte d'optimalité par rapport au cas complet channel state information at the transmitter side (CSIT). Exploiter le sous-espace de covariance est définitivement approprié pour Massive MIMO puisque, pratiquement, le rang de la matrice de covariance du canal est probablement plus petit que le nombre d'antennes BS. La performance de JSDM est basée sur le regroupement des utilisateurs en fonction de leurs eigenspaces de covariance. Par conséquent, la méthode de regroupement mise en œuvre est d'une importance primordiale. Dans [25], un clustering K -mean basé sur la distance chordale a été proposé pour les systèmes FDD avec JSDM. Dans [26], les auteurs ont étudié un large éventail de mesures de similarité telles que la vraisemblance pondérée, la projection sous-spatiale et les mesures de similarité basées sur Fubini. Dans [26], deux méthodes de clustering, à savoir, le clustering hiérarchique et K -medoids, étaient considérées pour le groupement d'utilisateurs avec les mesures de proximité mentionnées ci-dessus. Une comparaison des méthodes de regroupement proposées a été effectuée et la combinaison qui atteint la plus grande capacité a été dérivée. Dans [27], nous avons proposé une nouvelle mesure de similarité couplée à une nouvelle méthode de clustering afin d'obtenir un groupement d'utilisateurs approprié basé sur les statistiques de second ordre des canaux. En utilisant le même principe de formation de faisceau en deux étapes de JSDM, nous avons développé une approche de regroupement d'utilisateurs basée sur la théorie des graphes qui pallie les lacunes des méthodes de classification d'utilisateurs proposées précédemment. Dans [28], les propriétés spatiales du canal ont été exploitées afin d'obtenir des estimations CSI dans la plage Millimeter-Wave (mm-wave). Les auteurs ont proposé d'impliquer la corrélation temporelle entre deux blocs séquentiels dans la procédure en raison de l'ensemble de défis spécials que mm-waves impose. La réduction de la rétroaction peut également être obtenue à l'aide de la détection compressée et de la répartition des canaux [29,30]. Dans [31], la modélisation des canaux clairsemés (sparse) a été utilisée pour montrer que Compressed sensing (CS) peut réduire efficacement les ressources de formation temps-fréquence. Le même principe a également été utilisé dans [30, 32].

Vieillissement des canaux

Une autre raison de l'inexactitude de CSI est le vieillissement des canaux (*channel aging*). Ce phénomène résulte de la variation du canal entre l'instant où il est appris et l'instant où il est utilisé pour le décodage du signal dans le UL et le beamforming dans le DL. Cette variation temporelle est due à la mobilité des utilisateurs et aux retards de traitement dans les BS. La dégradation des performances due au vieillissement des canaux a été étudiée dans un système MIMO avec coordinated multi-point transmission/reception (CoMP) dans [33]. Les auteurs ont montré que l'impact du vieillissement des canaux est atténué lors de l'utilisation de filtres de prédiction de canal dans le régime de faible mobilité. Les auteurs de Truong et al. [34] ont fournit une analyse des performances de débit réalisables sur le UL et le DL, en présence de vieillissement des canaux et de prédiction de canal. Ils ont montré que, bien que le vieillissement des canaux entraîne une dégradation des performances des systèmes Massive MIMO, la prédiction des canaux fournit les moyens de surmonter ce problème. Dans Papazafeiropoulos et al. [35, 36], l'effet du vieillissement des canaux combiné à la prédiction a été étudié dans des scénarios avec des précodeurs Zero Forcing (ZF) régularisés (DL) et des récepteurs minimum mean square error (MMSE) , respectivement. Dans Kong et al. [37], les limites inférieures de SE pour les récepteurs Maximum ratio combining (MRC) et ZF avec et sans prédiction de canal ont été dérivées avec un nombre arbitraire d'antennes et d'utilisateurs. L'impact du vieillissement et de la prédiction des canaux sur la loi d'échelle de la puissance a été étudié. Les auteurs ont démontré que, dans le scénario monocellulaire et multicellulaire, l'échelle d'évolution de la puissance d'émission n'est pas affectée ni par un CSI retardé, ni par la prédiction de canal.

1.3.2 Détection de signaux codés

La détection du signal implique une estimation précise des données transmises connaissant le signal reçu. La détection de signal peut également exploiter la connaissance du CSI quand elle est disponible. Une large gamme d'algorithmes de détection est disponible pour les systèmes Massive MIMO. Ces algorithmes peuvent être classés comme linéaires ou non linéaires. La détection linéaire a l'avantage de la faible complexité qui vient avec le prix de la performance inférieure. En fait, la sortie des détecteurs linéaires se détériore rapidement à mesure que le nombre d'utilisateurs émetteurs augmente [38]. Lorsque le système devient limité par les interférences, des algorithmes de détection non linéaires peuvent être utilisés afin d'améliorer les performances. De tels systèmes implémentent une annulation d'interférence à plusieurs étapes. Ces systèmes comprennent des récepteurs d'annulation d'interférence successifs et parallèles.

1.3.3 Précodage et décodage

Massive MIMO exploite les connaissances de CSI afin de discriminer spatialement les signaux des utilisateurs. Le précodage (ou multiplexage) fait référence aux techniques qui permettent de focaliser le signal transmis sur un récepteur donné, minimisant ainsi la perte d'énergie dans les lobes latéraux. Le décodage (ou démultiplexage) fait référence à une combinaison cohérente à la réception de sorte que le signal reçu est détecté dans une direction donnée. Les deux sont réalisés en ajustant les phases et les amplitudes des signaux sur les différentes antennes des BS. De plus, le précodage a l'avantage de réduire peak-to-average-power-ratio (PAPR), un phénoméne très problématique pour les systèmes OFDM. Les techniques de précodage peuvent être non linéaires ou linéaires. Les méthodes non linéaires, telles que dirty-paper-coding (DPC) et les méthodes assistées par treillis, ont des performances plus élevées qui accompagnent une implémentation plus complexe. D'autre part, les précodeurs linéaires ont l'avantage d'une mise en œuvre simple. De tels précodeurs incluent MRC, MMSE, et ZF [12]. MRC maximise les Signal-to-noise ratio (SNR) en ajoutant les composantes du signal de manière cohérente sur les éléments d'antenne. MRC est particulièrement adapté à Massive MIMO qui utilise généralement une puissance rayonnée inférieure à partir de BS. Le précodage ZF est plus adapté aux scénarios à forte interférence et fonctionne plutôt bien avec un SNR élevé. ZF vise à annuler l'interférence multi-utilisateur. Bien qu'efficace dans la réduction des interférences, ZF s'accompagne d'un coût de calcul plus élevé et d'un gain de matrice réduit [39]. Le précodage MMSE est le précodage linéaire optimal dans un système Massive MIMO DL. Il s'agit d'un compromis entre l'amplification de la puissance du signal utile et la suppression des interférences multi-utilisateur. Par conséquent, le précodage MMSE fonctionne bien dans les deux conditions SNR haute et basse. Comme les fréquences mm-wave sont considérées pour 5 G, des techniques de précodage plus spécialisées ont été proposées pour ces fréquences. Dans [40], un schéma de précodage hybride combinant à la fois le précodage analogique et numérique a été proposé pour traiter l'atténuation élevée du signal qui se produit dans les fréquences mm-wave.

1.3.4 Efficacité énergétique

L'énorme potentiel de Massive MIMO dans la réduction de la consommation d'énergie et, par conséquent, l'augmentation de EE est maintenant très bien établi [7], [8]. En fait, puisque la puissance d'émission peut être réduite de manière significative, Massive MIMO peut produire un gain non négligeable en EE. De plus, l'utilisation d'un grand nombre d'antennes permet d'utiliser des composants peu coûteux sans perte de performance notable[117]. Dans [7], la loi d'échelle de puissance pour UL Massive MIMO a été dérivée et le grand potentiel d'amélioration de EE a été étudié. Dans [118], EE de Massive MIMO avec du matériel non idéal a été analysée. Dans [8], les auteurs ont montré, en tenant compte des dégradations matérielles, que EE est maximisé pour un nombre fini d'antennes déployées.

5 G ne peut pas être activé simplement en augmentant les performances de la couche physique. En effet, un changement de paradigme du réseau est nécessaire. Permettre au réseau de passer d'un paradigme agnostique réactif, de service et d'utilisateur à un paradigme plus proactif et intelligent peut produire une augmentation substantielle de ses performances. De plus, on s'attend à ce que 5 G gèrent divers scénarios de déploiement avec des exigences différentes. Cela peut devenir assez compliqué avec une topologie de réseau fixe.

Afin de résoudre ces problèmes, les réseaux 5 G centré sur l'utilisateur sont envisagés [3]. Cela signifie que les futurs réseaux 5 G seront caractérisés par une architecture plus plate avec une partie de l'intelligence déplacée vers le RAN, à la periphérie du réseau. L'un des principaux facilitateurs de 5 G centré sur l'utilisateur est, de toute évidence, la mise en cache proactive (Proactive Caching). Cela fait référence à la capacité de provisionnement local de contenu personnalisé. Ceci peut être réalisé en permettant au RAN d'obtenir les informations de contexte des utilisateurs et de prédire le trafic en utilisant des algorithmes d'analyse et de recommandation. Avoir un système plus dynamique RAN proactif permet de stocker localement du contenu populaire qui décharge le back-haul et réduit la latence End to End (E2E) tout en améliorant l'expérience utilisateur.

1.4 Mise en cache proactive pour les réseaux 5G centrés sur l'utilisateur: un bref historique et travaux connexes

L'idée de mise en cache proactive dans les réseaux sans fil trouve ses racines dans un principe plutôt ancien qui a d'abord été considéré dans le domaine des systèmes d'exploitation [41]. Le principe s'est ensuite étendu au web où il a été constaté que la mise en cache des contenus dans les serveurs proxy et autres nœuds du réseau permet d'améliorer l'évolutivité du World Wide Web et de décharger l'infrastructure réseaux [42]. La mise en cache proactive à la periphérie des réseaux sans fil est un concept plutôt récent. En se basant sur l'observation que les comportements humains sont corrélés et plutôt prévisibles [43], doter le RAN de la capacité d'analyser le trafic utilisateur et prédire le contenu le plus probable peut considérablement décharger le back-haul, améliorer Quality of Experience (QoE) et la latence E2E [3]. La mise en cache proactive permet de révolutionner les RAN. En effet, au lieu du paradigme conventionnel du tube de données réactif et agnostique de l'utilisateur, RAN sera doté de capacités d'analyse et de prédiction qui lui permettent de jouer un rôle important dans le provisionnement et la gestion de contenu. Les principaux avantages de la mise en cache proactive du côté RAN peuvent être résumés comme suit:

- **Réduire la latence de bout en bout (E2E)**
- **Améliorer l'efficacité énergétique**
- **Gérer la charge de trafic**
- **Augmenter le débit du réseau et améliorer la qualité d'expérience (QoE)**

En raison de son énorme potentiel pour répondre aux besoins de 5 G, la mise en cache proactive a attiré beaucoup d'attention dans les milieux de la recherche universitaire et industrielle. Cela a abouti à une littérature riche qui traite différents aspects de ce concept. Dans la suite, nous résumons certains des travaux sur la mise en cache proactive en fonction de leurs similitudes et directions.

Mise en cache proactive et l'estimation de la popularité du contenu

La mise en cache proactive se base sur la connaissance de l'utilisateur et du trafic, en particulier les contenus susceptibles d'intéresser l'utilisateur en prenant en compte les étiquettes de trafic, les attributs utilisateur, les types de terminaux, etc.

Dans [44], l'apprentissage automatique supervisé, en particulier, le filtrage collaboratif est utilisé pour estimer la popularité du contenu. Mettant à profit la connaissance du contexte, les réseaux sociaux et la corrélation des comportements humains, la popularité du contenu peut être estimée efficacement, ce qui entraîne des gains de déchargement considérables.

Dans [46], l'apprentissage à l'aide de transfert est étudié pour la mise en cache dans small cell network (SCN). Les résultats ont montré que le transfert de connaissances d'un domaine source d'informations contextuelles vers un domaine cible peut considérablement améliorer le gain de déchargement.

Dans [47], les mesures de centralité pour le placement de contenu sont exploitées. Les auteurs ont proposé un processus de diffusion de contenu basé sur la centralité où l'information complète sur la diffusion du contenu dans les réseaux sociaux n'est pas parfaitement connue. Les résultats ont montré que des gains de déchargement raisonnables peuvent être obtenus.

Dans [48], une formulation théorique alternative du problème de mise en cache proactive a été proposée. Après avoir modélisé le problème comme un jeu d'appariement plusieurs-à-plusieurs, les auteurs ont proposé un algorithme d'appariement qui aboutit à un résultat stable par paire avec un gain considérable dans le rapport des demandes satisfaites par le cache.

Alternativement, au lieu de généraliser la distribution de popularité sur tous les profils d'utilisateurs, nous avons proposé une approche de clustering d'utilisateurs basée sur la popularité du contenu dans [49]. Nous avons étudié l'impact d'une analyse plus détaillée du comportement des utilisateurs et justifié cette approche en utilisant un outil de sélection de modèles statistiques, à savoir Akaike information criterion (AIC). Les résultats ont montré que le regroupement des utilisateurs en fonction de leurs préférences peut augmenter considérablement les performances du système. Une extension de ce travail mettant l'accent sur l'impact d'une telle approche sur EE est donnée dans [50]. Dans [51], le regroupement des utilisateurs selon leur modèle de demande a également été étudié dans le but de réduire les délais de service. Les auteurs ont montré que le schéma de regroupement surpasse l'approche de mise en cache non groupée et aléatoire.

Gains de mise en cache codés

Traiter le problème de mise en cache proactive par les outils de théorie d'information est assez naturelle. En effet, la mise en cache proactive est un paradigme entièrement défini par les informations pouvant être collectées sur le comportement des utilisateurs. Dans [52], une approche théorique de la mise en cache proactive est donnée. Les auteurs ont dérivé des gains de mise en cache locaux et globaux basés sur un schéma codé qui exploite les deux. Dans ce cas, les résultats ont montré que l'approche codée conduit à une amélioration multiplicative du débit de pointe par rapport aux schémas précédemment connus. Ces résultats sont ensuite étendus aux popularités non uniformes du contenu dans [53, 54], l'accès au cache non uniforme dans [55], les tailles de cache hétérogènes dans [56], les systèmes de cache en ligne dans [57], les réseaux hiérarchiques de cache dans [58] et le cas multi-serveur dans [59]. Dans [60, 61], la mise en cache codée aléatoire est étudiée dans les réseaux sans fil device to device (D2D). Dans ce cas, il a été trouvé que le schéma D2D avec réutilisation spatiale et simple mise en cache aléatoire décentralisée permet d'obtenir la même loi de mise à l'échelle du débit quasi-optimal que la multidiffusion codée. Dans la même ligne de pensée, la performance du placement de cache aléatoire décentralisé avec un schéma de livraison codé est donnée dans [62, 63]. Une communauté de stockage D2D est étudiée dans [64]. En utilisant des codes de régénération et de la redondance dans le contexte de la mise en cache codée distribuée, les auteurs ont montré qu'une simple redondance peut entraîner des gains considérables de consommation d'énergie. Dans [65], les effets du codage réseau sur l'augmentation de la quantité de données disponibles pour les utilisateurs à travers les nœuds de cache sont étudiés. Dans ce cadre, une méthode de placement de contenu basée sur le codage réseau a été proposée, ce qui a permis d'accroître l'équité et le gain de déchargement. De plus, dans [66], la mise en cache codée a été étudiée dans des réseaux sans fil avec canal d'évanouissement. Dans ce document, il a été montré qu'avec une allocation de puissance ou de bande passante, les performances de débit de la mise en cache codée dans le mode de transport à division de fréquence sont nettement meilleures que dans le mode de répartition dans le temps.

Aspects de déploiement

Les aspects de déploiement sont d'une importance primordiale dans les réseaux dotés de mise en cache proactive. L'allocation optimale du contenu a été étudiée dans [45] où SBS (helpers) sont chargés de fournir le contenu aux utilisateurs via des transmissions à courte portée.

Les cas codés et non codés ont été étudiés. Les auteurs ont fourni des algorithmes d'affectation de contenu afin d'améliorer le délai de téléchargement total. Les extensions de ce travail, y compris pour D2D, sont données dans [67, 68].

Dans [69], le placement optimal du contenu dans SBS avec une capacité de backhaul limitée est étudié. Là, conditionnée par la connaissance de la distribution de popularité, le placement de contenu a été formulé comme un problème de sac à dos. Lorsque les profils des contenus disponibles ne sont pas connus à l'avance, le problème est formulé comme un problème "bandit manchot" qui permet d'apprendre la distribution de la popularité et de procéder au placement du contenu. Les auteurs ont fourni trois algorithmes pour identifier les compromis importants entre l'exploitation et l'exploration. Ce travail a été étendu avec une analyse plus approfondie dans [70].

La relation entre distance de collaboration et interférence a été étudiée dans [71] pour les réseaux D2D. Les auteurs ont montré qu'avec suffisamment de réutilisation de contenu, un débit non nul par utilisateur peut être atteint, même avec un stockage et un délai limités.

Dans [72], les différents compromis, pour les SBS avec backhaul limité, ont été étudiés. Dans ce travail, les auteurs ont étudié l'impact des différents paramètres du système sur la probabilité d'interruption et le taux moyen de distribution de contenu. Dans [73], le problème d'un placement géographique optimal du contenu est considéré. Là, il a été montré que le stockage des contenus les plus populaires n'est bénéfique que dans certains scénarios de déploiement particuliers. Particulièrement, lorsque les zones de couverture multiple sont importantes, il est plus avantageux d'introduire plus de diversité dans le contenu mis en cache. Des conclusions similaires ont été données dans [50]. Dans [74], le front de Pareto du coût de déploiement des caches et le coût attendu pour un client afin de récupérer un grand fichier à partir du cache sont étudiés.

Mise en cache proactive pour une communication verte

L'efficacité énergétique est considérée comme un Key performance indicator (KPI) primordiale pour les réseaux 5 G . Comme la mise en cache proactive échange un débit réseau avec la capacité de stockage, elle peut apporter un gain considérable en termes d'efficacité énergétique des réseaux. Cet impact de la mise en cache proactive a été étudié dans [75]. Là, les auteurs ont optimisé EE en ce qui concerne la puissance de transmission et ont montré qu'activer la mise en cache dans les BSs peut améliorer de manière significative l'EE. L'impact de la mise en cache proactive sur EE a également été étudié dans [76]. Les facteurs clés qui affectent les EE des réseaux activés en cache ont été étudiés. Dans [77], les auteurs ont fourni un cadre GreenDelivery pour la mise en cache. Il en résulte une réduction des activités BS et donc une consommation d'énergie réduite. Dans [78], une mise en cache conjointe et un cadre d'activation de BS pour les réseaux cellulaires verts est proposé. Les résultats démontrent jusqu'à 45 $\%$ d'économie énergitique. Dans [50], l'impact de la modélisation de la popularité du contenu sur EE a été étudié. Les auteurs ont optimisé EE par rapport à la densité optimale des SBSs actifs et au placement de contenu, en fonction de la classification des utilisateurs basée sur la popularité.

1.5 Plan de la thèse et contributions

Dans le chapitre 3 (TDD Massive MIMO systems: Enhancing CSI estimation through Spatial Division based training), nous nous attaquons au problème d'estimation du CSI dans les systèmes Massive MIMO TDD . Nous adoptons la division spatiale comme moyen de planifier plus d'utilisateurs tout en gardant les frais généraux d'entraînement sous contrôle.

L'idée principale est basée sur le fait que le planning d'un plus grand nombre d'utilisateurs, par cellule, peut être réalisée sans avoir à augmenter les ressources de formation. Ceci est réalisé en permettant la réutilisation des pilotes dans les cellules tout en séparant les signaux des utilisateurs copilotes en fonction de leurs informations spatiales. Nous proposons une approche alternative pour le clustering spatial d'utilisateurs afin de pallier les lacunes des méthodes de regroupement précédemment utilisées [25, 26]. Comme un système multicellulaire est pris en compte, les interférences intra-cellulaires et inter-cellulaires sont traitées. Dans ce chapitre, ces deux problèmes sont découplés et traités successivement.

Afin de gérer les interférences copilotes intra-cellulaire, au lieu de regrouper les utilisateurs en fonction de la similarité de leurs signatures spatiales [25, 26], nous adoptons une approche différente qui vise à construire des groupes d'utilisateurs copilotes. Dans chaque cellule, un groupe copilote donné est formé de sorte qu'il contienne des utilisateurs avec un chevauchement minimal dans leurs signatures spatiales et qui fournissent une couverture maximale des DoFs du système. L'idée est d'associer chaque utilisateur à un ensemble de faisceaux qui concentrent une grande partie de la puissance de son canal. Après avoir obtenu les matrices de décodage spécifiques à l'utilisateur, les BSs dérivent des groupes d'utilisateurs copilotes. Chaque groupe fournit une couverture maximale de tous les flux indépendants disponibles avec un chevauchement minimal entre les matrices spécifiques aux utilisateurs. Contrairement aux schémas de regroupement spatial antérieurs, où chaque signal d'utilisateur est traité avec une matrice spécifique au groupe [24–26, 82], l'approche proposée permet de prendre en compte les informations spatiales réelles de chaque utilisateur. Cette approche permet également de coupler les problèmes de regroupement d'utilisateurs et d'ordonnancement, ce qui réduit la complexité de la gestion du réseau.

Nous proposons deux formulations du problème de regroupement spatial et nous proposons deux algorithmes en conséquence. Tout d'abord, le problème de la sélection des utilisateurs copilotes basé sur la couverture spatiale est formulé comme un *problème de couverture maximale* [79]. Dans le second cas, le problème de la génération des groupes copilotes est formulé comme un *Problème de couverture maximale généralisée* [80]. Nous fournissons deux algorithmes efficaces pour résoudre les deux problèmes formulés et nous évaluons leur performance en dérivant leurs rapports d'approximation respectifs.

Une fois les groupes d'utilisateurs copilotes formés, nous abordons le problème de l'interférence copilots inter-cellulaires grâce à un système efficace d'attribution de séquences d'apprentissage. Pour ce faire, nous formulons un problème d'optimisation combinatoire qui exploite l'information spatiale des liaisons interférentielles. Le réseau est par conséquent capable d'allouer des séquences d'apprentissage UL spécifiques à des groupes d'utilisateurs copilotes dans des cellules différentes, de sorte que les interférences résultantes peuvent être gérées efficacement en utilisant les récepteurs spatiales précédemment définis. Le problème d'allocation de pilotes qui en résulte est formulé sous la forme d'un *max-cut problem* [81], ce qui permet d'utiliser un algorithme d'approximation de faible complexité pour le résoudre.

Dans le chapitre 4 (Enhancing performance by long term CSI estimation planning), nous abordons le problème d'estimation du CSI UL dans les systèmes Massive MIMO TDD, en utilisant une approche différente. En fait, en partant du constat que les systèmes sans fil actuels supposent la même durée de créneau horaire pour tous les appareils, sans tenir compte du fait que les utilisateurs sont soumis à des spreads Doppler hétérogènes, on remarque un DoF précédemment négligé, à savoir, la *périodicité d'estimation du CSI*. En fait, la durée de l'intervalle de cohérence dans les systèmes sans fil actuels est basée sur la propagation Doppler maximale prise en charge. La surcharge d'estimation CSI est définie en conséquence [1]. Cette approche est sous-optimale car elle implique que le réseau va consacrer des ressources temps-fréquence précieuses à l'estimation d'informations réutilisables. C'est particulièrement le cas pour les utilisateurs à faible mobilité car leurs canaux ne changent pas au même rythme que les utilisateurs plus mobiles. Par conséquent, aborder le goulot d'étranglement de la formation UL en permettant une estimation adaptative CSI basée sur la propagation Doppler semble tout à fait logique. Nous proposons une approche dans laquelle les ressources de formation nécessaires sont définies dynamiquement, à chaque tranche de temps. Cette idée est en accord avec le concept de **Dynamic TDD** [2], où les ressources de la trame TDD sont définies dynamiquement sans configuration préfixée.

Le concept de base est que, dans un créneau donné, si la corrélation entre l'estimation de CSI et le canal réel n'était pas considérablement dégradée en raison du vieillissement, le réseau n'est pas obligé de le réestimer. Cela permet d'utiliser une partie des ressources de formation pour la transmission de données ou pour programmer davantage d'utilisateurs. Étant donné que le vieillissement des canaux résulte principalement de la mobilité, la vitesse étant un paramètre important, il est nécessaire d'optimiser la politique de formation sur de longues périodes tout en intégrant la sensibilisation à la mobilité. Le développement de politiques de formation à long terme nécessite des estimations précises de l'emplacement des utilisateurs, ce qui peut être plutôt compliqué à obtenir, en pratique [1], [85].

Par conséquent, ce problème est abordé en supposant que le réseau a une connaissance partielle des positions des utilisateurs. En effet, nous supposons que le réseau est capable d'estimer l'emplacement d'un ensemble limité d'utilisateurs. L'adaptation à la modification des coefficients d'évanouissement à grande échelle et l'optimisation des décisions d'apprentissage UL, basées sur l'autocorrélation du canal, devraient se faire sur deux échelles de temps différentes [84]. En fait, les deux optimisations sont basées sur des informations qui changent sur des échelles de temps hétérogènes. Afin d'atteindre une SE cumulative maximale sur des intervalles de temps plus grands que le bloc de cohérence de l'évanouissement à grande échelle, un problème de contrôle à deux échelles temporelles est considéré.

Dans l'échelle de temps rapide (niveau inférieur), le réseau dérive une stratégie de décision d'entraînement optimale tout en supposant des statistiques de second ordre de canal constantes (coefficients d'évanouissement à grande échelle). En prenant en compte l'évolution dans le temps de la corrélation entre le CSI estimé et le canal actuel, le réseau est capable d'optimiser ses décisions d'ordonencement des utilisateurs pour l'entraînement UL pour un horizon fini. La prise en compte de la dimension temporelle permet au réseau d'être plus efficace puisqu'il devient capable de prédire l'impact de ses décisions sur les performances présentes et futures. Pour ce faire, nous proposons un système intelligent où le réseau n'est pas obligé d'optimiser sa décision d'entraînement au début de chaque créneau mais apprend la meilleure politique de formation pour des périodes importantes mais limitées. En raison de la lente évolution des coefficients d'évanouissement et d'autocorrélation à grande échelle, le réseau est capable de dériver une séquence de décisions d'entraînement optimales basées sur les mêmes informations. Dans ce cas, le problème d'optimisation de la formation UL peut naturellement être formulé comme un problème de planification discrète sur un horizon de temps fini [86]. En fait, l'optimisation des décisions d'ordonnancement du réseau équivaut à dériver une séquence d'actions qui maximise la SE cumulative au fil du temps. Puisque dériver la stratégie d'entraînement optimale peut être prohibitif en calcul pour de grands horizons d'optimisation, nous fournissons un cadre d'optimisation combinatoire qui permet de dériver une politique d'entraînement approximative avec un temps de calcul réduit.

Dans l'échelle de temps lente (niveau supérieur), le réseau s'adapte à la mobilité des utilisateurs en décidant quels utilisateurs doivent évaluer leur emplacement. La sélection efficace de ces utilisateurs est d'une importance primordiale en raison de la consommation d'énergie et des frais de signalisation qui en résultent. Le problème d'apprentissage à deux échelles est modélisé comme un Partially Observable Markov Decision Process (POMDP) [87]. Dans ce chapitre, nous fournissons des algorithmes efficaces pour le résoudre et nous présentons les gains obtenus en term de SE.

Dans le chapitre 5 (User-centric 5G networks: EE under popularity based user Clustering), nous mettons l'accent sur la mise en cache proactive. Nous explorons l'impact du placement de contenu et de la modélisation de la popularité sur un SCN avec mise en cache proactive.

Alors que la plupart des travaux précédents considèrent une popularité moyenne du contenu sur tous les utilisateurs, ce travail utilise un cadre de mise en cache alternative. En effet, les utilisateurs sont regroupés en fonction de leur profil de popularité. Ce choix est motivé par l'existence de schémas de trafic très divers chez les utilisateurs. En effet, le contenu demandé dépend du réseau social de l'utilisateur et des intérêts qui peuvent être très différents d'une personne à l'autre. Supposer une popularité de contenu homogène parmi les utilisateurs ne peut que conduire à perdre des informations précieuses. Afin de montrer la pertinence de cette approche, un critère de sélection de modèle statistique, à savoir *critère d'information Akaike* est utilisé. AIC permet de mesurer la véracité d'un modèle statistique donné. Il aborde également le compromis entre l'adéquation d'un modèle statistique, basé sur l'estimation du maximum de vraisemblance, et sa complexité, donnée par le nombre de paramètres à estimer. AIC permet d'adapter la mise en cluster des utilisateurs aux changements de configuration de trafic car il peut détecter les modifications dans le nombre optimal de clusters. Dans le cas d'une faible mobilité, il est judicieux d'adapter le placement des fichiers en fonction des informations de localisation. C'est le cas des utilisateurs en zone confinée (bureau, campus universitaire ...). La connaissance géographique est exploitée dans le cadre de mise en cache, plus précisément la corrélation spatiale dans les modèles de trafic. Ce problème est abordé juste après la détection du comportement principal grâce à la classification basée sur la popularité du contenu. Ce choix est motivé par les différentes échelles de temps selon lesquelles la popularité du contenu et la localisation de l'utilisateur évoluent. En effet, si la corrélation dans la popularité du contenu entre les utilisateurs d'un même groupe social est constante pendant de longues périodes, leurs positions peuvent changer en raison de la mobilité. Cela motive la nécessité d'adapter le placement de contenu mis en cache plus souvent que les fichiers sélectionnés afin de simplifier la gestion du réseau. Un cadre d'optimisation de l'emplacement du cache est donné en fonction du regroupement d'utilisateurs précédemment effectué. Le problème d'optimisation combinatoire qui en résulte vise à exploiter la corrélation spatiale dans le modèle de trafic utilisateur afin de réduire la puissance consommée. Il présente également un autre avantage intéressant du clustering basé sur le contenu. En fait, le regroupement effectué sur les utilisateurs permet également de regrouper les fichiers en conséquence. Par conséquent, la complexité du problème de placement de fichiers résultant est plus faible puisque l'espace de recherche est réduit de l'ensemble du catalogue à des groupes de fichiers de taille totale approximativement égale. Cela simplifie considérablement la gestion du système de mise en cache par rapport au travaux existants qui se basent uniquement sur la localisation, où la complexité des problèmes formulés est proportionnelle au nombre de fichiers total.

Enfin, le chapitre 6 inclut nos conclusions et un aperçu sur les futurs travaux potentiels liés à cette thèse. Nous notons que chaque chapitre ci-dessus contient sa propre notation mathématique.

1.6 Publications

La liste des travaux publiés au cours de cette thèse est donnée ci-dessous.

Papiers Journals

[106] Salah Eddine HAJRI and Mohamad Assaad, *A spatial basis coverage approach for uplink training and scheduling in Massive MIMO systems*, submitted to IEEE Transactions on Wireless Communications 2018. **(Chapter 3)**

[50] Salah Eddine HAJRI and Mohamad Assaad, *Energy Efficiency in Cache Enabled Small Cell Networks With Adaptive User Clustering*, IEEE Transactions on Wireless Communications (Volume: PP, Issue: 99), 17 November 2017.**(Chapter 5)**

[107] Salah Eddine HAJRI, Maialen Larranaga and Mohamad Assaad, *Heterogeneous Doppler Spread-based CSI Estimation Planning for TDD Massive MIMO* , submitted to IEEE Transactions on Wireless Communications 2017. **(Chapter 4)**

[108] Ali Maatouk, Salah Eddine Hajri, Mohamad Assaad, Hikmet Sari, *Optimal Scheduling for Joint Spatial Division and Multiplexing: Complexity Results and Approximation Algorithm*, submitted to IEEE Transactions on Wireless Communications 2018.

Papiers conférences

[109] Salah Eddine HAJRI and Mohamad Assaad, *An Exclusion zone for Massive MIMO With Underlay D2D Communication*, International Symposium on Wireless Communication Systems (ISWCS), Aug. 2015 .

[110] Salah Eddine HAJRI, Mohamad Assaad and Giuseppe Caire, *Scheduling in Massive MIMO: User clustering and pilot assignment*, 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2016. **(Chapter 3)**

[27] Ali Maatouk, Salah Eddine Hajri, Mohamad Assaad, Hikmet Sari, Serdar Sezginer, *Graph Theory Based Approach to Users Grouping and Downlink Scheduling in FDD Massive MIMO*, IEEE ICC, 2018.

[111] Salah Eddine Hajri, Mohamad Assaad, Maialen Larranaga, *Enhancing massive MIMO: A new approach for Uplink training based on heterogeneous coherence time*, accepted in IEEE ICT'18, 2018. **(Chapter 4)**

[49] Salah Eddine HAJRI and Mohamad Assaad, *Caching improvement using adaptive user clustering*, 17th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Edinburgh, UK, 3-6 July 2016. **(Chapter 5)**

[113] Mohamad Assaad, Salah Eddine Hajri, Thomas Bonald, Anthony Ephremides, *Power Control in Massive MIMO with Dynamic User Population*, submitted to WiOpt'18.

[112] Salah Eddine Hajri, Juwendo Denis and Mohamad Assaad, *Enhancing Favorable Propagation in Cell-Free Massive MIMO Through Spatial User Grouping*, accepted in IEEE SPAWC 2018.

Brevets

[114] Salah Eddine HAJRI and Mohamad Assaad, *Relay operations in a cellular network*, gb GB1710498.5, UK, TCL Communication Limited, 2017.

[115] Salah Eddine HAJRI and Mohamad Assaad, *IMPROVEMENTS IN OR RELATING TO DYNAMIC CHANNEL AUTOCORRELATION BASED ON USER SCHEDULING*, gb GB1710487.8, UK, TCL Communication Limited, 2017. **(Chapter 4)**

[116] Salah Eddine HAJRI and Mohamad Assaad, *Sparse uplink CSI estimation for MTC devices in massive MIMO systems*, TCL Communication Limited, 2018.

Chapter 2

Introduction

2.1 Background and Motivation

Mobile communications have been fundamental in producing our contemporary information societies. From legacy analog mobile systems to the more sophisticated LTE networks [1], the advances in wireless systems have radically changed the ways with which humans access and exchange information.

Presently, wireless communications are at crossroads. As a matter of fact, the ever growing demand for high data capacity and the proliferation of smart devices with applications that require high rates calls upon the definition of a more efficient next generation standards so that the substantial increase of data traffic can be handled.

The fifth-generation (5G) mobile communications systems are rapidly emerging to address a wide range of challenges brought by the thirst of our present and future societies for wireless communications. 5G is envisioned to tackle, in addition to a mountainous increase in traffic volume, the challenge of connecting billions of devices with heterogeneous service requirements. 5G networks are expected to provide improvements such as [2]:

- 10 folds increase in experienced throughput: Ushering the era of more uniform, multi-Gbps peak rates.

- 10 folds decrease in latency: Latency levels are expected to be as low as 1 ms.

- 10 folds connection density: Enabling IoT connectivity with low complexity and signaling overhead.

- 3 folds spectrum efficiency: More efficient utilization of the available bandwidth using advanced antenna techniques.

- 100 folds traffic capacity: Highly densified networks with more access point (AP)s everywhere.

- 100 folds network efficiency: More EE networks with effectual processing and hardware.

These high-level performance targets for 5G were developed as part of IMT-2020, the ITU initiative to define the basis for 5G. These requirements are mapped to three different main use cases, namely eMBB, massive MTC and URLLC [3],[4],[5]:

- Ultra-reliable and low latency communications: This use case focuses on latency-sensitive and reliability demanding services. It tackles the expectations of wireless controlled industrial manufacturing, remote medical surgery, smart grids and automated driving, etc.

- Enhanced mobile broadband: The enhanced mobile broadband use case comes with new applications and exigencies in addition to the legacy mobile broadband applications. It aims to meet the demands of a more and more digitalized human lifestyle. It focuses on services with broadband requirements such as virtual reality (VR), augmented reality (AR) and video streaming.

- Massive machine type communications: massive MTC focuses on meeting the demands of a high number of connected devices with low data capacity and latency-sensitivity requirements. This includes applications such as smart cities, IoT, etc.

The different aforementioned use cases have varying and, sometimes, conflicting characteristics. Indeed, as an example, 5G must scale from supporting low-data rate sensors at 10s of kbps to new immersive mobile experiences at several Gbps. This means that 5G networks need to be able to scale across diverse services with equally diverse mobile devices. Doing so comes with its toll of difficulties.

2.2 5G: A concentration of new paradigms and innovative technologies

Meeting the aforementioned requirements necessitate drastic changes in the network paradigm in addition to a large array of disruptive innovations. In this context, 5G networks can call upon a wide range of physical and higher layers new technologies. This enables a leap in wireless networks performance that dwarfs its predecessors. These innovations will touch on transmission and design of the networks physical layer in addition to introducing upheaval in the networking and application layer techniques. As a matter of fact, 5G NR will employ many key technologies in order to attain new levels of performance and efficiency. The combinations of the latter will expand the importance of mobile communications and enables it play a much central role in a world of changing use cases. Among the potential main innovation in 5G physical layer, one can cite [3]:

- mm-wave Mobile Communications.
- Massive MIMO Communications.
- Non-Orthogonal Multiple Access (NOMA).
- Full-DuplexWireless Communications.
- Carrier aggregation and Multicarrier Modulations.
- Larger spectrum.
- Sidelink communication.
- New waveform and heterogeneous OFDM numerology.

In addition to these physical improvement, the innovations in 5G will change how networking is performed. As a matter of fact, 5G concentrates a number of new paradigms which aim at enabling a more agile, automatic and intelligent network in each of its operations [5]. Among the main networking innovations in 5G, one can state [3]:

- cloud radio access network (CRAN).
- Energy Harvesting.
- Green Heterogeneous Wireless Access.
- Self Organizing Network (SON) networks.
- Fog Computing.
- User-Centric Wireless Network (proactive caching, etc).

In this thesis, we mainly focused on two major technologies that will help usher in the age of 5G, namely *Massive MIMO* and *proactive caching*. The main goal during the present thesis was to improve the performance of both technologies using intelligent optimization based on services and user awareness. This interest was based on a fundamental observation. Both technologies can be improved by optimizing network operations based on slow changing and easily obtainable side information on mobile users. Leveraging underexploited knowledge at the RAN side, such as spatial statistics, Doppler spread and efficient modeling of content popularity, the network performance can be substantially improved. This is achieved thanks to the incorporation of efficient algorithms in the networks procedures. This observation is the cornerstone of this thesis and special care was taken in order to develop schemes that enable the improvement of the network performance with the least possible signaling and training overhead counterpart. In the following sections, we first give an overview of past and recent advancements in both technologies. We then present the outline of the thesis and its contributions accordingly.

2.3 Massive MIMO for 5G: A brief history and related works

Any evolution in network generation, perforce, have to make substantial improvement in SE and area throughput. Massive MIMO is a technology that makes this possible for 5G since it brings a minimum of ten-fold improvement in SE [3]. This impressive gain is achieved using arrays of some hundred cheap antenna elements at the BSs. Thus enabling the spatial multiplexing of a considerable number of mobile devices. The basic idea is based on large scale statistical effects that results from drastically increasing the number of BS antennas (on the order of hundreds) [6]. This results in practically reducing the impacts of fast fading, interference and additive noise. More importantly, it enables to focus the radiated energy on intended targets. Consequently, by coherent processing of the signals over the BS antenna array, transmit precoding can be used in order to concentrate each signal at its intended terminal and receive combining can be used in order to discriminate between the signals of different users. The more antennas at the BS, the finer the spatial focusing can be. This allows to schedule many more users than is possible today, hence immensely increasing overall SE, connection density and area capacity. The excess of BS antennas results in increasing the number of data streams which can be exploited to serve more terminals, reducing the radiated power, while boosting the data rate. Massive MIMO can also improve link reliability through spatial diversity and provide more DoF in the spatial domain, resulting in enhancing performance irrespective of the noisiness of the measurements. Owing to the resulting aggressive spatial multiplexing, Massive MIMO can deliver an impressive gain in the network performance by simply steering the radiated waves in the right directions. Since the radiated energy is highly concentrated on user centric zones, Massive MIMO delivers considerable gains in EE [7], [8]. All of this is achieved with cost efficient antenna elements at the BSs and simplified, usually single antenna, user equipment. The main benefits of Massive MIMO systems can be summarized as follows:

- **High Spectral efficiency gain** Massive MIMO inherits the conventional MU-MIMO gains but on a massive scale, as its name specifies. As a matter of fact, with M BS antennas serving K single antenna users, a diversity of order M together with a multiplexing gain of $min(M, K)$ are achieved. These parameters can be tuned in order to further improve SE and achieve higher communication resilience.

- **High Energy efficiency gain** Massive MIMO achieves its performance thanks to the excess of BS antennas combined with coherent processing. This allows to considerably reduce transmit power. Consequently, through coherent combining at the reception and beamforming at the transmission, EE can be considerably improved.

- **Simple processing**

  Massive MIMO employs simple yet efficient signal processing schemes (linear precoding and deconding in the DL and UL, respectively). Indeed, the large number of BS antennas reduces the complexity of mitigating the impact of small scale fading and additive noise. In addition, when the number of antennas is sufficiently large, the resulting channel hardening further simplifies signal processing.

- **Increased robustness and reliability**

  The large number of BS antennas procures more diversity. This results in better link reliability and higher rate. In addition, as the number of antennas increases, additive noise, small scale fading, and intra- cell interference are bound to vanish.

- **Cost reduction in radio frequency (RF) power components**

  Massive MIMO employs coherent processing which allows to reduce the radiated power. This enables to use low cost radio frequency (RF) amplifiers in the milli-Watt range.

However, there always exists a compromise between achievable performance and complexity. The interesting gains of Massive MIMO come with their share of challenges:

- **Multiuser interference management** Massive MIMO deliver considerable gains in network performance. However, some users can have their channel suffer from an uneven impact of multiuser interference. Consequently, it may be necessary to implement interference cancellation schemes. Interference alignment [9], maximum likelihood multiuser detection [10] and dirty paper coding [11] can be used. These schemes suffer from a major shortcoming, namely high computational complexity.

- **CSI acquisistion**

  Coherent processing is the cornerstone of Massive MIMO systems. Consequently, timely and accurate CSI is required. This can be very challenging in both FDD and TDD modes, given the system scale (number of users, number of BS antennas). User mobility have also an important impact since it defines the correlation between estimated CSI and actual channel realization.

- **User scheduling** Massive MIMO is envisaged to handle a large number of connected devices. With the 5G requirements of high density connection, user scheduling is of paramount importance. In addition, when the same time-frequency resources are used, selecting the users that can be active simultaneously can considerably alter the system performance.

Due to the advantages and popularity of Massive MIMO, it have been the subject of ever increasing focus from the research community.This resulted in a rich literature that treats different aspects of this concept. In the following, we summarize some of the works on Massive MIMO based on their similarities and directions.

2.3.1 CSI estimation Methods:

The cornerstone of Massive MIMO is coherent signal processing which relays on accurate and timely CSI knowledge. In what follows, we provide an overview of CSI estimation in Massive MIMO systems in addition to the characterization of the most important challenges that relate to CSI acquisition.

TDD systems and pilot contamination

In TDD Massive MIMO systems, CSI estimates are obtained using channel reciprocity and UL training. In these systems, only BSs are required to have CSI in order to coherently precod and decode multiuser signals. The amount of time-frequency training resources depends on the number of users antennas. TDD is considered as more suitable for Massive MIMO operations since it implies that channel estimation needs to be performed in only one direction, and then can be used in both directions. This advantage can not be overlooked since it means that training overhead scales only with the number of users. Nevertheless, owing to the limited coherence interval, the training dimension is restricted and the same pilot sequences need to be reused which results in pilot contamination [6],[13]. This phenomenon has been identified as a major limiting factor of Massive MIMO performance and has attracted substantial attention in previous works [93]. Several methods have been proposed in order to reduce or, better yet, eliminate the impact of pilot contamination in TDD Massive MIMO systems. These methods are either pilot-based or subspace-based [93]. Time shifting for pilot transmission was proposed in [14], [15], as a mean to reduce pilot contamination. The main idea was to shifting pilot transmission in time so that users in different cells transmit at non-overlapping times. Results show that this protocol result is an efficient elimination of pilot contamination. In [16], Ashikhmin and Marzetta proposed a pilot contamination precoding (PCP) method based on the slow fading coefficients. The proposed method requires a certain level of cooperation between BS in order to construct the PCP matrices. Results show that this method can deliver non negligible SE gains. This work has been extended in [17] by proposing large-scale fading pre-coding (LSFP) and large scale fading decoding (LSFD) in the regime of a finite number of BS antennas. Results showed an interesting gain in the 5% outage rate. Subspace-based approaches, on the other hand, improve CSI estimation accuracy by leveraging the signal's higher order statistics. In [18], the authors showed that, leveraging spatial channel correlation and adequate precoding/combining, the capacity of Massive MIMO system can increase without limit as a function of the number of BS antennas. This is made possible by exploiting linear independence between copilot users channel covariance matrices. In [19], covariance matrix EVD is implemented in order to obtain CSI estimates. In order to decrease errors, EVD is combined with iterative least-square. The authors showed that the EVD method enable to mitigate the impact of pilot contamination and outperforms conventional pilot-based CSI estimation techniques. In [20, 21], blind detection was proposed. The main idea was to exploit singular value decomposition in order to discriminate user signals. Results showed that the knowledge of the subspace spanned by each channel vector is sufficient to obtain accurate CSI estimates through simple projection. Nevertheless, this approach assumes that all desired channels are stronger than all interfering channels which do not always hold in practice. Neumann et al. [22] proposed a MAP criterion for subspace channel estimation in order to overcome this issue. Although performance improvement was noticed, the utilization of MAP comes at the price of increased complexity. In [23], an iterative least-square projection with Diagonal jacket-based estimation was proposed in order to address the issue of pilot contamination.

FDD systems and feedback overhead

In FDD systems, since UL and DL utilize different frequency bands, CSI of both links need to be estimated. UL CSI is obtained by enabling users to send different pilot sequences. In the DL, CSI estimates are obtained using DL training followed by explicit or implicit CSI feedback. As the number of BS antennas increases, FDD channel estimation becomes very problematic since CSI feedback overhead scales linearly with the number of system antennas [24]. Enabling FDD Massive MIMO systems can only be done if the feedback bottleneck is addressed.

This problem was addressed in [24, 25], where JSDM for MU-MIMO DL was investigated. JSDM is a scheme that aims to serve users by clustering them into groups such that users within a group have approximately similar channel covariances, while users across groups have near orthogonal covariance eigenspaces. JSDM has been designed, originally, for FDD Massive MIMO systems without considering inter-cell interference. It enables to reduce CSI feedback overhead in FDD systems while incurring no loss of optimality with respect to the full CSIT case. Exploiting the covariance subspace is definitively suitable for Massive MIMO since, practically, the rank of the channel covariance matrix is likely smaller than the number of BS antennas. The performance of JSDM is based on grouping users according to their covariance eigenspaces. Consequently, the implemented grouping method is of paramount importance. In [25], a chordal distance based K-mean clustering was proposed for FDD systems with JSDM. In [26], the authors investigated a wide range of similarity measures such as weighted likelihood, subspace projection and Fubini-Study based similarity measures. In [26], two clustering methods namely, hierarchical and K-medoids clustering for user grouping with the aforementioned proximity measures. A comparison of the proposed grouping methods was performed and the combination that achieves the largest capacity was derived. In [27], we proposed a new similarity measure coupled with a novel clustering method in order to achieve appropriate user grouping based on the channels second order statistics. Using the same principle of two stage beamforming as in JSDM, we developed a graph theory based user grouping approach that mitigate the shortcomings of previously proposed user clustering methods. In [28], the spatial properties of the channel were leveraged in order to obtain CSI estimates in the mm-wave range. The authors proposed to implicate time correlation between two sequential block frames in the procedure owing to the special set of challenges that mm-waves impose. Reducing the feedback overhead can also be achieved using compressed sensing and channel sparsity [29, 30]. In [31], sparse channel modeling was used in order to show that CS can efficiently save DL time-frequency training resources. The same principle was also used in [30, 32].

Channel aging

Another reason for CSI inaccuracy is *channel aging*. This phenomenon results from the variation of the channel between the instant when it is learned and the instant when it is used for signal decoding in the UL and beamforming in the DL. This time variation is due to users mobility and processing delays at the BS. Performance degradation due to channel aging was studied in a MIMO system with CoMP in [33]. The authors showed that the impact of channel aging is mitigated when utilizing channel prediction filters in the low mobility regime. The authors in Truong et al. [34] provide an analysis of the achievable rate performance on UL and DL in the presence of channel aging and channel prediction. They showed that, although channel aging leads to degradation in the performance of Massive MIMO systems, channel prediction provides the means to overcome this issue. In Papazafeiropoulos et al. [35, 36], the effect of channel aging combined with channel prediction has been investigated in scenarios with regularized ZF precoders (DL) and MMSE receivers, respectively. In Kong et al. [37], lower bounds of the sum-rate for both MRC and ZF receivers with/without channel prediction have been derived with an arbitrary number of BS antennas and users. The impact of channel aging and prediction on the power scaling law has been studied. The authors demonstrated that, in the single-cell and the multi-cell scenario, the transmit power scaling is not affected neither by aged CSI nor channel prediction.

2.3.2 Detection of Encoded Signals

Signal detection implies accurate estimation of the transmitted data knowing the received signal. Signal detection can also exploit knowledge of the CSI when available. A wide range of detection algorithms is available for Massive MIMO systems. These algorithms can be classified as linear or nonlinear. Linear Detection has the advantage of low complexity which comes with the price of lower performance. In deed the output of linear detectors deteriorates rapidly as the number of transmitting users increases [38]. When the system becomes interference limited, nonlinear detection algorithms can be employed in order to improve performance. Such schemes implement interference cancellation at multiple stages. Such popular schemes include successive and parallel interference cancellation receivers.

2.3.3 Precoding and Decoding

Massive MIMO exploits CSI knowledge in order to spatially discriminate user signals. Precoding (or multiplexing) refers to the techniques that enable to focus the transmitted signal on a given receiver, thus minimizing energy loss in side lobes. Decoding (or demultiplexing) refers to coherent combining at the reception so that the received signal is detected in a given direction. Both are realized by adjusting the phases and amplitudes of the signals at the different BS antennas. In addition, precoding has the advantage of reducing PAPR which is very problematic for OFDM systems. Precoding techniques can be nonlinear or linear. Nonlinear methods, such as DPC and lattice-aided methods, have higher performance that come with more complex implementation. Linear precoders, on the other hand, have the advantage of simple implementation. Such precoders include MRC, MMSE, and ZF [12]. MRC maximizes the SNR by adding the signal components coherently over the antenna elements. This achieved using the wireless channel estimate. MRC is particularly suited for Massive MIMO which typically employs lower radiated power from the BS. ZF precoding is more suited to interference limited scenarios and performs quite well with high SNR. ZF aims at nulling the multiuser interference. While efficient in reducing interference, ZF comes with the penalty of higher computational cost and reduced array gain [39].

MMSE precoding is the optimal linear precoding in a Massive MIMO DL system. It strikes a compromise between amplifying the useful signal power and suppressing multiuser interference. Consequently, MMSE precoding performs well in both high and low SNR conditions. Since mm-wave frequencies are considered in 5G, more specialized precoding techniques have been proposed for these frequency range. In [40], a hybrid precoding scheme combining both analog and digital precoding has been proposed in order to deal with the high signal attenuation that happens at mm-wave frequencies using a non-complex sub array.

2.3.4 Energy efficiency

The huge potential of Massive MIMO in reducing energy consumption and consequently increasing EE is now very well established [7], [8]. In fact, since the transmit power can be significantly reduced, Massive MIMO can produce a non negligible EE gain. Moreover, the use of a large number of antennas make it possible to employ inexpensive component without considerable performance loss [117]. In [7], the power scaling law for UL Massive MIMO was derived and the large potential in improving EE was investigated. In [118], EE of Massive MIMO with non ideal hardware was analyzed. In [8], the authors showed, taking into consideration hardware impairments, that EE is maximized for a finite number of deployed antennas.

5G can not be enabled by, simply, increasing the physical layer performance. Indeed, a networking paradigm change is needed. Enabling the network to transit from a reactive, service and user agnostic paradigm to a more proactive and intelligent one, can produce substantial increase in its performance. Moreover, 5G is expected to handle diverse deployment scenarios with different requirements. This can become quite complicated with a fixed network topology.

In order to address these issues, user-centric 5G is envisaged [3]. This means that future 5G networks will be characterized by a flatter architecture with part of the intelligence shifted down to the RAN. One of the major enablers of User-centric 5G is, definitely, proactive caching. It refers to the ability of personalized local content provisioning. This can be achieved by enabling the RAN to obtain users' context information and predict traffic using analysis and recommendation algorithms. Having a proactive more intelligent RAN enable to locally store popular content which offloads the back-haul and reduces E2E latency while enhancing user experience.

2.4 Proactive Caching for User-centric 5G networks: A brief history and related works

The idea of proactive caching in wireless networks finds its roots in a rather old principle which was first considered in the field of operating systems [41]. The principle spread then to the web where it was found that caching contents in the proxy servers and other nodes of the network enables to enhance the scalability of the world wide web and offload the networks infrastructure [42]. Proactive caching at the edge of wireless networks is a rather recent concept. Based on the observation that human behavior is correlated and rather predictable [43], endowing the RAN with the ability to analyze user traffic and predict most likely content to be requested can considerably offload the back-haul while enhancing QoE and E2E latency [3]. Proactive caching enables to revolutionize the RAN. Indeed, instead of the conventional paradigm of reactive and user-agnostic data pipe, RAN is endowed with analysis and prediction capabilities that enable it to play an important role in content provisioning and management. The main benefits of proactive caching at the RAN side can be summarized as follows:

- **Reduce E2E latency**
- **Improve energy efficiency**
- **Manage traffic load**
- **Increase the network throughput and enhance QoE**

Owing to its huge potential in addressing 5G requirements, proactive caching has attracted considerable attention in both academic and industrial research communities. This resulted in a rich literature that treats different aspects of this concept. In the following, we summarize some of the works on proactive caching based on their similarities and directions.

Proactive Caching and Content Popularity Estimation

Proactive caching relays on user and traffic awareness, specifically, contents that the user may be interested in by considering traffic labels, user attributes, terminal types, etc. Consequently, content popularity modeling and estimation are of paramount importance.

In [44], supervised machine learning, specifically, collaborative filtering is used in order to estimate content popularity. Leveraging context-awareness, social networks and correlations of human behavior, content popularity can be efficiently estimated which results in considerable offloading gains.

In [46], transfer learning for caching in SCN is studied. Therein results showed that transfer of knowledge from a rich contextual information source domain to a target domain can considerably improve the offloading gain.

In [47], the centrality measures for content placement are exploited. The authors proposed a centrality-based content dissemination process where the complete information of content dissemination in social networks is not perfectly known. Therein, results showed that reasonable offloading gains can be obtained. In [48], an alternative game theoretical formulation of the proactive caching problem was proposed. After modeling the problem as a many-to-many matching game, the authors proposed a matching algorithm that reaches a pairwise stable outcome with considerable gain in satisfied requests ratio.

Alternatively, instead of generalizing the popularity distribution over all user profiles, we proposed a content popularity based user clustering approach in [49]. Therein, we investigated the impact of a more detailed analysis of user behavior and justified their approach using a statistical model selection tool, namely AIC. Results showed that clustering users according to their preference and caching content accordingly can significantly increase the system performance. An extension of this work with emphasis on the impact of such approach on EE is given in [50]. Clustering users according to their request pattern was also investigated in [51] with the goal of reducing service delay. The authors showed that the clustering scheme outperforms the unclustered and random caching approach.

Coded Caching Gains

An information-theoretic formulation of the caching problem is quite natural. Indeed, proactive caching is a paradigm that is completely defined by the information that can be collected on the users behavior. In [52], an information-theoretic approach of proactive caching is given. The authors derived local and global caching gains based on a coded scheme that exploits both gains. Therein, results showed that the coded approach leads to a multiplicative improvement in the peak rate compared to previously known schemes. These results are then extended to non-uniform content popularities in [53, 54], non-uniform cache access in [55], heterogeneous cache sizes in [56], online caching systems in [57], hierarchical caching networks in [58] and multi-server case in [59]. In [60, 61], random coded caching is studied in D2D wireless networks. Therein, it was found that the D2D scheme with spatial reuse and simple decentralized random caching achieves the same near-optimal throughput scaling law as coded multicasting. In the same line of thinking, the performance of decentralized random caching placement with a coded delivery scheme is given in [62, 63]. A D2D storage community is studied in [64]. Using regenerating codes and redundancy in the context of distributed coded caching, the authors showed that simple redundancy can lead to considerable gains in energy consumption. In [65], the effects of network coding on increasing the amount of available data to the users through the cache nodes is studied. Therein, a network coding-based content placement method was proposed, resulting in increasing fairness and offloading gain. Additionally, in [66], coded caching was studied in wireless networks with fading channel. Therein, it was shown that with power or bandwidth allocation, the throughput performance of coded caching under the frequency-division transport mode is significantly better than that under the time-division mode.

Deployment Aspects

Deployment aspects are of paramount importance in cache enabled networks. Optimal content assignment was studied in [45] where SBS (helpers) are in charge of delivering the contents to the users via short-range transmissions. Therein, both coded and uncoded cases were investigated. The authors provided content assignment algorithms in order to improve the total expected downloading delay. Extensions of this work, including D2D case, is given in [67, 68].

In [69], optimal content placement in a SBS with limited backhaul capacity is studied. Therein, conditioned on the knowledge of popularity distribution, content placement was formulated as a knapsack problem. When profiles of the available contents are not known in advance, the problem is formulated as a multi-armed bandit problem which enables to learn the popularity distribution and proceed to content placement. The authors provided three algorithms to pinpoint the important trade-offs between exploitation and exploration. This work was extended with more extensive analysis in [70].

The relation between collaboration distance and interference was studied in [71] for D2D networks. The authors showed that with enough content reuse, non-vanishing throughput per user can be attained, even with limited storage and delay.

In [72], the different trade-offs, in cache-enabled SBS with limited backhaul, where studied. Therein the authors investigated the impact of different system parameters on the achievable outage probability and average content delivery rate.

In [73], the problem of an optimal geographic placement of content is considered. Therein, it was shown that storing the most popular contents is beneficial only in some particular deployment scenarios. Particularly, when multi-coverage areas are significant, it is more beneficial to introduce more diversity in the cached content. Related conclusions were given in [50].

In [74], the Pareto front of the expected deployment cost of the caches in the plane and the expected cost for a client to retrieve a large data file from the cache is studied.

Proactive Caching for Green communication

Energy efficiency is considered as a major KPI for 5G networks. Since proactive caching trades off storage capacity with network throughput, it can bring a considerable gain in the networks energy efficiency. This impact of proactive caching was studied in [75]. Therein, the authors optimized the achievable EE with respect to transmit power and showed that enabling caching at the BSs can significantly improve the EE. The impact of proactive caching on EE was also investigated in [76]. The key factors that impact the EE of cache enabled networks were studied. In [77], the authors provided a GreenDelivery framework, with the joint design of Energy-Harvesting, push, and caching. It results in the reduction of BS activities and thus a reduced energy consumption. In [78], a joint caching and BS activation framework for green cellular networks is proposed. Therein, results demonstrated up to 45% energy savings. In [50], the impact of content popularity modeling on EE was studied. The authors optimized EE with respect to the optimal density of active SBSs and to content placement, under popularity based user clustering.

2.5 Thesis Outline and Contributions

In Chapter 3 (TDD Massive MIMO systems: Enhancing CSI estimation through Spatial Division based training), we address the bottleneck of UL training in TDD Massive MIMO systems. We adopt spatial division as a mean to schedule more user while keeping in check training overhead.

The main idea is based on the fact that scheduling more users per cell can be achieved without having to increase the training resources. This is performed by allowing pilot reuse within cells while separating copilot users signals based their spatial information. We propose an alternative approach for spatial user clustering in order to address the shortcoming of previously used grouping methods [25, 26]. Since a multi-cell system is considered, both intra-cell and inter-cell interference is addressed. In this chapter, these two problems are decoupled and dealt with successively.

In order to deal with intra-cell copilot interference, instead of grouping users based on the similarity of in their spatial signatures [25, 26], we adopt a different approach that aims at constructing copilot user groups. In each cell, any given copilot group is formed such that it contains users with minimum overlapping in their signals spatial signatures and that provide a maximum coverage of the systems' DoFs. The proposed approach is referred to as *Spatial basis coverage based copilot UE selection.* The idea is to associate each user with a set of beams that concentrate a large amount of its channel power. After obtaining the users specific decoding matrices, the BSs derive copilot user groups. Each group provides a maximum coverage of all available independent streams with minimum overlapping between users specific beam matrices. In opposition to prior spatial clustering schemes, where each user signal is processed with a group specific matrix [24–26, 82], the proposed approach enables to take into consideration the actual spatial information of each user. This approach enables also to couple the problems of user grouping and scheduling which reduces the complexity of the network management.

We provide two formulation of the copilot grouping problem and we propose two grouping algorithms accordingly. First, the problem of spatial basis coverage based copilot user selection is formulated as a *maximum coverage problem* [79]. In the second case, the problem of copilot group generation is formulated as a *Generalized maximum coverage problem* [80]. We provide two efficient algorithms to solve the two formulated problems and we assess their performance by deriving their respective approximation ratios.

Once copilot user groups are formed, we address the issue of inter-cell copilot interference through an efficient cross-cell training sequence allocation scheme. In order to do so, we formulate a combinatorial optimization problem that leverages the spatial information of interference links. The network is, consequently, able to allocate specific UL training sequences to copilot user groups in different cells, such that the resulting interference can be managed efficiently using the previously defined spatial signature based receivers. The resulting pilot allocation problem is formulated as a *max-cut problem* [81], which enables to use a low complexity approximation algorithm to solve it.

In Chapter 4 (Enhancing performance by long term CSI estimation planning), we address the bottleneck of UL training in TDD Massive MIMO systems, using a different approach. In fact, based on the observation that current wireless systems assume the same time slot duration for all devices regardless of the fact that users are subject to heterogeneous Doppler spreads, we notice a DoF that was previously neglected, namely, *CSI estimation periodicity*. As a matter of fact, the coherence slot duration in current wireless systems is based on the maximum supported Doppler spread and the CSI estimation overhead is defined accordingly [1]. This approach is suboptimal since it implies that the network is going to spend precious time-frequency resources on estimating information that may be reusable. This is particularly the case for users with low mobility since their channels do not change at the same rate as faster moving users. Consequently, addressing the UL training bottleneck by enabling a Doppler spread based adaptive CSI estimation seems to be quite logical. We propose an approach in which the needed training resources are defined dynamically, at each time slot. This idea is in according with the concept of Dynamic TDD [2], where the slot resources are defined dynamically without a prefixed configuration.

The basic concept is that, at a given slot, if the correlation between the estimated CSI and the actual channel was not considerably degraded, due to aging, the network is not required to reestimate it. Doing so enables to spear part of the training resources that can be used for data transmission or to schedule more users. Since channel aging results, primarily, from mobility, with speed being an important parameter, optimizing the training policy for long time periods while incorporating mobility awareness is required. Developing long term training policies requires accurate estimates of user locations, which can be rather complicated to obtain, in practice [1],[85]. Consequently, this problem is tackled while assuming that the network has a partial knowledge of the user positions. We suppose that the network is able to estimate the location of a limited set of users. Adapting to the change in the large-scale fading coefficients and optimizing UL training decisions based on the channel's autocorrelation should occur on two different time scales [84]. In fact the two optimizations are based on information that changes over heterogeneous time scales. In order to achieve the maximum cumulative average SE over time spans larger than the large-scale fading coherence block, a two time scale control problem is considered. In the fast time scale (lower level), the network derives an optimal training decision strategy while assuming constant channel second order statistics (large-scale fading coefficients). By taking into consideration the evolution over time of the correlation between the estimated CSI and the actual channel, the network is able to optimize its decisions to schedule users for UL training over a finite time horizon. Taking into consideration the time dimension allows the network to be more efficient since it becomes able to predict the impact of its decisions on present and future performance. In order to do so, we propose an intelligent system where the network is not required to optimize its training decision at the beginning of each time slot but learns the best training policy for large but finite time periods. Owing to the slow changing large scale fading and autocorrelation coefficients, the network is able to derive a sequence of optimal training decisions based on the same information. In this case, the UL training optimization problem can naturally be formulated as a discrete planning problem over a finite time horizon [86]. In fact, optimizing the network's scheduling decisions is equivalent to deriving a sequence of actions that will maximize the cumulative average SE over time. We formulate a finite horizon deterministic control problem. The optimal training decisions are derived for a predefined time duration, denoted here by H, for which the large-scale fading coefficients are supposed to be constant. This is quite advantageous since it allows the network to optimize its training over time without requiring the actual channel estimates. Since deriving the optimal training strategy can be computationally prohibitive for large optimization horizons, we provide a combinatorial optimization framework that enables to derive an approximate training policy with reduced running time. In the slow time scale (upper level), the network adapts to user mobility by deciding which users are required to feedback their locations. Efficiently selecting these users is of paramount importance owing to the resulting energy consumption and signaling overhead. The two time scale learning problem is modeled as a POMDP [87]. In this chapter, we provide efficient algorithms to solve it and we present the resulting gains in SE.

In Chapter 5 (User-centric 5G networks: EE under popularity based user Clustering), we focus on improving the performance of User-aware 5G networks, particularly, we focus on proactive caching. We explore the impact of content placement and popularity modeling on a cache enabled SCN.

While most previous works average content popularity over all users, this work uses an alternative caching framework. In fact, users are grouped according to their content popularity profiles. This choice is motivated by the existence of very diverse traffic patterns among users. Indeed, the requested content depends on the user social network and interests that can be very different from one person to the other. Assuming a homogeneous content popularity among users can only result in loosing valuable information. In order to showcase the pertinence of this approach, statistical model selection, namely, *Akaike information criterion* is used. AIC enables to measure the truthfulness of a given statistical model. It also addresses the trade-off between the fitness of a statistical model based on maximum likelihood estimation and its complexity which is given by the number of parameters to be estimated. AIC enables to adapt user clustering to any traffic pattern changes since it can detect modifications in the optimal number of clusters.

In the case of low mobility, where users do not change positions too often, it makes sense to adapt the files placement based on location information. This is the case for users in confined areas (office, university campus....). Geographical user awareness is then exploited in the caching framework, specifically, spatial correlation in traffic patterns. This problem is tackled right after main behavior detection through content popularity based clustering. This choice is motivated by the different time scales according to which content popularity and user location evolve. In fact, while the correlation in content popularity between users from the same social group is constant for long periods of time, their location can change due to mobility. This motivates the need to adapt the cached content placement more often than the selected files in order to simplify the management of the network. A cache placement optimization framework is given based on the previously performed user grouping. The resulting combinatorial optimization problem aims at exploiting the spatial correlation in user traffic pattern in order to reduce the average consumed power. It also showcases another interesting advantage of content based clustering. In fact, the clustering that is done on the users enables also to group the files accordingly. Consequently, the complexity of the resulting optimal file placement problem is lower since the search space is reduced from the whole file catalog to groups of file of approximately equal total size. This considerably simplifies the management of the caching system compared to existing work on location based optimization where, the complexity of the formulated problems is proportional to the number of files.

Finally, chapter 6 includes our conclusions and potential future works related to this thesis. We note that each chapter above contains its own mathematical notation.

2.6 Publications

The list of published works during the course of this PhD is given below.

Journal Articles

[106] Salah Eddine HAJRI and Mohamad Assaad, *A spatial basis coverage approach for uplink training and scheduling in Massive MIMO systems*, submitted to IEEE Transactions on Wireless Communications 2018. **(Chapter 3)**

[50] Salah Eddine HAJRI and Mohamad Assaad, *Energy Efficiency in Cache Enabled Small Cell Networks With Adaptive User Clustering*, IEEE Transactions on Wireless Communications (Volume: PP, Issue: 99), 17 November 2017.**(Chapter 5)**

[107] Salah Eddine HAJRI, Maialen Larranaga and Mohamad Assaad, *Heterogeneous Doppler Spread-based CSI Estimation Planning for TDD Massive MIMO* , submitted to IEEE Transactions on Wireless Communications 2017. **(Chapter 4)**

[108] Ali Maatouk, Salah Eddine Hajri, Mohamad Assaad, Hikmet Sari, *Optimal Scheduling for Joint Spatial Division and Multiplexing: Complexity Results and Approximation Algorithm*, submitted to IEEE Transactions on Wireless Communications 2018.

Conference Papers

[109] Salah Eddine HAJRI and Mohamad Assaad, *An Exclusion zone for Massive MIMO With Underlay D2D Communication*, International Symposium on Wireless Communication Systems (ISWCS), Aug. 2015 .

[110] Salah Eddine HAJRI, Mohamad Assaad and Giuseppe Caire, *Scheduling in Massive MIMO: User clustering and pilot assignment*, 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2016. **(Chapter 3)**

[27] Ali Maatouk, Salah Eddine Hajri, Mohamad Assaad, Hikmet Sari, Serdar Sezginer, *Graph Theory Based Approach to Users Grouping and Downlink Scheduling in FDD Massive MIMO*, IEEE ICC, 2018.

[111] Salah Eddine Hajri, Mohamad Assaad, Maialen Larranaga, *Enhancing massive MIMO: A new approach for Uplink training based on heterogeneous coherence time*, accepted in ICT'18, 2018. **(Chapter 4)**

[49] Salah Eddine HAJRI and Mohamad Assaad, *Caching improvement using adaptive user clustering*, 17th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Edinburgh, UK, 3-6 July 2016. **(Chapter 5)**

[113] Mohamad Assaad, Salah Eddine Hajri, Thomas Bonald, Anthony Ephremides, *Power Control in Massive MIMO with Dynamic User Population*, submitted to WiOpt'18.

[112] Salah Eddine Hajri, Juwendo Denis and Mohamad Assaad, *Enhancing Favorable Propagation in Cell-Free Massive MIMO Through Spatial User Grouping*, accepted in IEEE SPAWC 2018.

Patents

[114] Salah Eddine HAJRI and Mohamad Assaad, *Relay operations in a cellular network*, gb GB1710498.5, UK, TCL Communication Limited, 2017.

[115] Salah Eddine HAJRI and Mohamad Assaad, *IMPROVEMENTS IN OR RELATING TO DYNAMIC CHANNEL AUTOCORRELATION BASED ON USER SCHEDULING*, gb GB1710487.8, UK, TCL Communication Limited, 2017. **(Chapter 4)**

[116] Salah Eddine HAJRI and Mohamad Assaad, *Sparse uplink CSI estimation for MTC devices in massive MIMO systems*, TCL Communication Limited, 2018.

Chapter 3

TDD Massive MIMO systems: Enhancing CSI estimation through Spatial Division based training

3.1 Overview

Each evolution in wireless networks is required to provide considerable increase in SE and area throughput. These requirements especially vital in 5G networks owing to the considerable throughput demand that it should face. One of the most promising technologies to achieve the required performance improvement is, without doubt, Massive MIMO [6]. Indeed, by leveraging a large number of antennas at the BSs, Massive MIMO proved to be able to provide a considerable improvement in the network's spectral and energy efficiencies [88],[7], [8].

However, Massive MIMO gains depend heavily on acquiring accurate CSI estimates at the BSs. In TDD systems, CSI estimation is performed through UL training, leveraging channel reciprocity [13]. Unfortunately, owing to the limited coherence interval, the training dimension is restricted and the same pilot sequences need to be reused which results in the phenomenon of pilot contamination [13]. Addressing this issue lead to the development of numerous CSI estimation methods that exploit different channel statistics in order to mitigate copilot interference and enhance CSI accuracy. Several of these methods leverage spatial division multiplexing in order to efficiently discriminate user channels and thus substantially reduce copilot interference. The main idea resides in grouping users based on their spatial information and processing their signals accordingly. These methods proved to be able to provide considerable SE gains in both FDD and TDD Massive MIMO systems. Indeed, in TDD mode, users with non-overlapping spatial signatures can be allowed to reuse the same pilot sequence, even within a single cell, since a simple linear projection on each user dominant signal space can effectively mitigate copilot interference [92]. In FDD mode, spatial division multiplexing can be exploited in order to reduce the CSI feedback overhead with little or even negligible capacity loss. This enables to schedule more users for the same CSI feedback overhead which increases SE and connection density [25]. Previous works on spatial multiplexing proposed to group users based on their channel covariance eigenspace [25, 26], when UL or UL channel covariances are known, or simply based on their signals mean direction of arrivals [92]. Different grouping methods were implemented including, for example, K-mean [25], K-medoids and hierarchical clustering [26]. In [92], a greedy user scheduling algorithm was implemented in order to partition users into copilot groups based on their spatial signatures. In addition different proximity measures between users signal spaces were also considered[26]. These studies showed that the performance of spatial division multiplexing schemes depend heavily on the implemented user grouping scheme.

Although a considerable increase in SE was recorded, these methods come with a number of shortcomings. In fact, in order to achieve a good user clustering, the K-medoids and K-mean clustering methods require a prior estimation of the parameter K. In addition, these methods use an averaging in order to derive the group specific eigenspace matrix which can lead to a substantial overlapping between the clusters. As a matter of fact, practically, users might have similar but not necessarily identical second order channel statistics. This dictates the need to consider individual user spatial information. Another major shortcoming of these methods is the fact that they overlook the efficiency of leveraging the totality of the available DoFs of the system.

In this chapter, we propose an different approach for spatial user clustering. We consider a multi-cell TDD Massive MIMO system, in which, spatial diversity is exploited in order to allow for a more aggressive pilot reuse, within each cell, while mitigating copilot interference. This allows for an increase in the number of scheduled users for the same training overhead while improving SE. Since a multi-cell system is considered, both intra and inter-cell pilot contamination are addressed.

We choose to decouple these two problems and address them successively. In order to deal with intra-cell copilot interference, we propose a spatial grouping and scheduling scheme. Instead of grouping users based on the similarity of in their covariance eigenspaces [25,26], we adopt a different approach that aims at constructing copilot user groups based on the users spatial signatures. In each cell, any given copilot group is formed such that it contains users with minimum overlapping in their signals spatial signatures and that provide a maximum coverage of the systems' independent streams. The proposed approach is referred to as *spatial basis coverage based copilot users selection.* The idea is to associate each user with a set of beams that concentrate a large amount of its channel power. Since Uniform linear array (ULA)s are considered, the columns of a unitary discrete Fourier transform (DFT) matrix are used as spatial basis [25],[92]. After obtaining the users specific decoding matrices, the BSs derive copilot user groups. Each group provides a maximum coverage of all available independent streams with minimum overlapping between users specific beam matrices. This approach enables also to couple the problems of user grouping and scheduling which reduces the complexity of the network management. We provide two formulation of the copilot grouping problem and we propose two grouping algorithms accordingly. First, the problem of spatial basis coverage based copilot UE selection is formulated as a *maximum coverage problem* [79]. In the second case, the problem of copilot group generation is formulated as a *Generalized maximum coverage problem* [80]. The two formulations enable us to provide efficient algorithms that perform the desired grouping. We go one step further and derive the approximation ratio of each algorithm in order to assess its performance.

Once copilot user groups are formed, we address the issue of inter-cell copilot interference through an efficient cross-cell training sequence allocation. In order to do so, we formulate a combinatorial optimization problem in a graphical framework based on the copilot groups spatial signature. Using this information, the network is able to allocate specific UL training sequences to copilot user groups in different cells, such that the resulting interference can be managed efficiently using the previously defined spatial receivers.

The rest of the chapter is organized as follows. The systems model under consideration is provided in Section 3.2. In Section 3.3, we discuss the principal and performance of spatial division multiplexing in TDD Massive MIMO systems. In Section 3.4, we propose a spatial basis cover scheme for copilot user selection. Therein, we propose two algorithms to achieve the desired grouping and we assess their guaranteed performance. In Section 3.6, numerical results capturing the gains that the proposed scheme can provide are given. We finally conclude in Section 3.7.

3.2 System Model And Preliminaries

We consider a multi-cell, multi-user Massive MIMO network operating in TDD mode. The network is composed of N_c cells containing each, a BS that is equipped with a large M-element ULA. Each BS is serving K single omni-directional antenna users such that $K >> M$. Users are randomly distributed in each cell. Considering flat fading channels, the channel vector between user i in cell b and the BS of the r^{th} cell, $\mathbf{g}_{ib}^{[r]}$ is composed of an arbitrary number of i.i.d. P rays ($P >> 1$)[90]. Hence, the UL channel $\mathbf{g}_{ib}^{[r]}$ is given by the following multi-path model

$$\mathbf{g}_{ib}^{[r]} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \mathbf{a}(\theta_{ib}^{[r,p]}) \gamma_{ib}^{[r,p]}, \tag{3.1}$$

Here, $\gamma_{ib}^{[r,p]}$ represents the complex gain of the p^{th} ray from user i in cell b and the BS of the r^{th} cell and follows a $\mathcal{CN}\left(0, \mu_{ib}^{[r]^2}\right)$ distribution where $\mu_{ib}^{[r]}$ denotes the average attenuation of the channel. $\theta_{ib}^{[r,p]}$ denotes the direction of arrival of the p^{th} ray from user i in cell b and the BS of the r^{th} cell. Moreover, $\mathbf{a}(\theta_{ib}^{[r,p]}) \in \mathbb{C}^{M\times 1}$ is the array manifold vector which is given by:

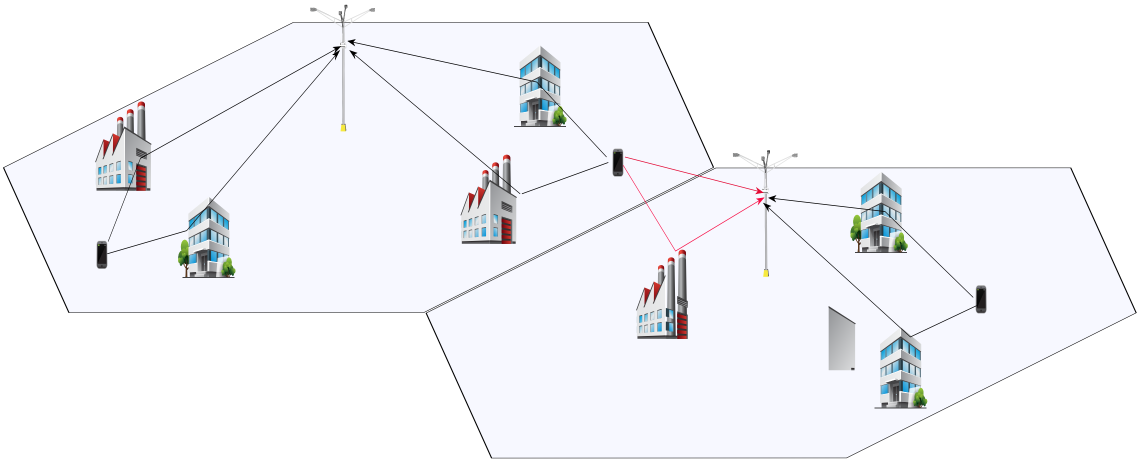

Figure 3.1: System Model

$$\mathbf{a}(\theta_{ib}^{[r,p]}) = \left[1, e^{\frac{j2\pi d}{\lambda} sin(\theta_{ib}^{[r,p]})}, \ldots, e^{\frac{j2\pi d}{\lambda} sin(\theta_{ib}^{[r,p]})(M-1)}\right], \tag{3.2}$$

where λ denotes the signal wavelength, d refers to the antenna spacing such that $d \leq \frac{\lambda}{2}$. As in [90] and [92], the incident angles of each user, with mean Direction of Arrival (DOA) $\theta_{ib}^{[r]}$, are considered to be restrained in a narrow angular range $\left[\theta_{ib}^{[r]} - \Delta\theta_{ib}^{[r]}, \theta_{ib}^{[r]} + \Delta\theta_{ib}^{[r]}\right]$. Within this range $\mathbf{a}(\theta_{ib}^{[r,p]}), p = 1, \ldots, P, \forall i, b, r$ are mutually correlated. Consequently, the covariance matrix of each channel $\mathbf{g}_{ib}^{[r]}$, which is given by $\mathbf{R}_{ib}^{[r]} = \mathbb{E}\left[\mathbf{g}_{ib}^{[r]}\mathbf{g}_{ib}^{[r]\dagger}\right]$, possesses a low-rank property.

As introduced above, in this chapter we focus on a TDD Massive MIMO system, where the entire frequency band is used for DL and UL transmission by all BSs and users. The BSs acquire CSI estimates using orthonormal training sequences (i.e., pilot sequences) in the UL. At each coherence interval, a maximum of τ users are scheduled for UL training in each cell with $\tau \leq K$. For that, we consider a set of orthonormal training sequences, that is, sequences $q_i \in \mathbb{C}^{\tau \times 1}$ such that $q_i^\dagger q_j = \delta_{ij}$ (with δ_{ij} the Kronecker delta). In this chapter, we consider an aggressive pilot reuse approach. In fact, in addition to reusing the same set of orthogonal pilot sequences in every cell, we consider that the same sequences are reused even within one cell. Consequently, the channel estimates are corrupted by both inter and intra-cell pilot contamination.

3.3 Spatial Division Multiplexing Based User Scheduling

3.3.1 Spatial Basis in Massive ULAs

Massive MIMO systems provide a substantial SE gain by spatially multiplexing a large number of mobile devices. This greater number of served devices requires higher signaling or feedback overhead in order to obtain CSI estimates. Consequently, new CSI acquisition schemes are required in order to address this issue, taking into consideration user clustering, grouping and beamforming. This issue promoted many research work which resulted in designing new transmission strategies for Massive MIMO leveraging low-rank approximation of the channel covariance matrix [25, 26, 90]. Indeed, based on the fact that the incident signals at the BSs are characterized by narrow angular spread, it was proven that the effective channel dimension can be reduced without capacity loss. This is achieved through eigen-decomposition of covariance matrices and eigenmode based user grouping[25, 26].

In this chapter, we build an alternative low-rank model, leveraging the characteristics of ULAs. Indeed, in this case, it was proven that a unitary DFT matrix constitute a good spatial basis of the signal [24, 92]. This means that, without accurate estimates of the channel covariance matrices, spatial division multiplexing can be implemented using a unitary DFT matrix. Indeed, for each user, it is sufficient to derive decoding matrices based on the DFT matrix vectors that concentrate the majority of its channel power [92]. The main idea is to group users according to their spatial signatures and allocate pilot sequence such that copilot users are spatially separated (i.e., their spatial signatures span independent subspaces). This principle is similar to previously proposed spatial division methods. However, we propose a novel grouping scheme that take into consideration the efficiency of the spatial space coverage. Indeed, copilot user grouping is performed based on two criterion. First, users are grouped such that they have minimum overlapping in their spatial signatures. Second, the users in each group provide a maximum coverage of the set of independent DFT streams. The latter criterion was completely sidestepped in previous work. Users signals are then processed using a per-user DFT-based decoding matrix.

We start by deriving the achievable average Signal-to-interference-plus-noise ratio (SINR) with a per-user DFT-based decoding matrix and we discuss the proposed user grouping in the next section. As in [92], we consider a DFT based spatial basis for ULA. In more details, the BSs proceed to DL training once each T_l, where T_l is the duration during which the channel spatial information remains unchanged. In this DL training phase, each user will feedback the indexes of the reference beams (i.e, columns of the DFT matrix $\mathbf{F}$) that concentrate a considerable percentage of its channel power. Particularly, each user i, b will feedback the indexes of the DFT matrix columns for which the following condition is verified:

$$\frac{\left\| \mathbf{g}_{ib}^{[b]\dagger} \mathbf{f}_s \right\|^2}{Tr(\mathbf{R}_{ib}^{[b]})} \geq \alpha, \tag{3.3}$$

where $0 < \alpha < 1$ is a design parameter that characterizes the projection on a reference beam and, consequently, the percentage of the total power received along this beam. The vectors for which the condition above is verified will form a detection matrix $\mathbf{F}_{ib}^{[b]}$ for each user ib:

$$\mathbf{F}_{ib}^{[b]} = \{\mathbf{f}_s \in \mathbf{F}, \frac{\left\|\mathbf{g}_{ib}^{[b]\dagger}\mathbf{f}_s\right\|^2}{Tr(\mathbf{R}_{ib}^{[b]})} \geq \alpha\}, \tag{3.4}$$

$\mathbf{F}_{ib}^{[b]}$ will be used as the bases in which the user's signal is detected and will henceforth be refereed to as spatial signature of user ib. The spatial signature of each user forms a subspace that concentrate a large percentage of the its channel power. Consequently, allocating the same copilot sequence to users with minimum overlapping in their spatial signature (3.4) enables to discriminate between their signals since the power of their channels is concentrated in different subspaces.

Note that this association can be done without requiring a covariance matrix estimate thanks to DL probing [110]. In the proposed scheme, we increase the reuse factor of training sequences. The same pilot sequence can be allocated to multiple users in a given cell, if their detection matrices are non-overlapping. This results in increasing the number of scheduled users while, at the same time, ensures that copilot interference, within the same cell, can be efficiently mitigated thanks to simple linear projections in the users spatial spaces.

3.3.2 Achievable Performance With Adaptive Spatial Division Based User Scheduling

For every user i, b, we consider the matrix $\mathbf{F}_{ib}$ formed by the vectors of the DFT matrix according to (3.4). During the UL training period, active users send their pilot sequence so that the BSs can estimate their CSI. For analytical simplicity, we consider that UL training sequences have the same reuse factor within all cells. We consider that the network schedules N_p users to use any given pilot sequences in each cell. During UL training, the received pilot signal $\mathbf{Y}_p^{[b]}$ at BS b is given by:

$$\mathbf{Y}_p^{[b]} = \sqrt{\rho_p}\sum_{r=1}^{N_c}\sum_{l=1}^{\tau}\sum_{i\in\Sigma(l,r)} \mathbf{g}_{ir}^{[b]}\,\mathbf{q}_l{}^{\dagger} + \mathbf{W}_p, \tag{3.5}$$

where $W_p \in \mathbb{C}^{M\times\tau}$ refers to an additive white Gaussian noise matrix with i.i.d. $\mathcal{CN}(0,1)$ entries and ρ_p denote the pilot transmit power. $\Sigma(l,r)$ denotes the set of user in cell r that are using pilot sequence l during UL training. The b^{th} BS then uses the orthogonality of the training sequences in order to obtain the MMSE estimate of the channel of user i, b [129]. In order to discriminate the copilot users signals, the BS exploits the previously defined spatial signature matrices (3.4). The BS estimates the channel of each user i, b after projecting the received signal on $\mathbf{F}_{ib}^{[b]}$ as

$$\hat{\mathbf{g}}_{ib}^{[b,ib]} = \mathbf{F}_{ib}^{[b]\dagger}\mathbf{R}_{ib}^{[b]}\left(\frac{1}{\rho_p}\mathbf{I}_M + \sum_{r=1}^{N_c}\sum_{u\in\Sigma(\chi(i,b),r)}\mathbf{R}_{ur}^{[b]}\right)^{-1}(\frac{\mathbf{Y}_p^{[b]}\mathbf{q}_{\chi(i,b)}}{\rho_p}), \tag{3.6}$$

where $\chi(i, b)$ denote the index of the training sequence used by user i, b. Note that indexes of the projection matrix $\mathbf{F}_{ib}^{[b]}$ are added to the channel estimate since its law depends on UE i, b spatial signature. Using the orthogonality characteristic of the MMSE estimate, the wireless channel of each user i, b can be decomposed as

$$\mathbf{g}_{ib}^{[b,ib]} = \hat{\mathbf{g}}_{ib}^{[b,ib]} + \tilde{\mathbf{g}}_{ib}^{[b,ib]}, \tag{3.7}$$

where $\tilde{\mathbf{g}}_{ib}^{[b,ib]} \sim \mathcal{CN}\left(0, F_{ib}^{\dagger}\mathbf{R}_{ib}^{[b]}F_{ib} - F_{ib}^{\dagger}\mathbf{R}_{ib}^{[b]}\left(\frac{1}{\rho_p}\mathbf{I}_M + \sum_{r=1}^{N_c}\sum_{u\in\Sigma(\chi(i,b),r)}\mathbf{R}_{ur}^{[b]}\right)^{-1}\mathbf{R}_{ib}^{[b]}F_{ib}\right)$ represents the uncorrelated estimation error. Next we study the achievable average UL SINR under the considered setting. During UL data transmission, BS b receives the following data signal

$$\mathbf{Y}_u^{[b]} = \sqrt{\rho_u}\sum_{r=1}^{N_c}\sum_{l=1}^{\tau}\sum_{i\in\Sigma(l,r)}\mathbf{g}_{ir}^{[b]}d_{ir} + \mathbf{w}_u, \tag{3.8}$$

where $\mathbf{w}_u \in \mathbb{C}^{M\times 1}$ refers to an additive white Gaussian noise vector with i.i.d. $\mathcal{CN}(0, 1)$ entries and ρ_u denotes the UL data transmission power. We consider linear detection where, the signal of each user i, b is estimated using a precoded matched filter receiver. In order to detect the signal of user i, b, BS b uses $\mathbf{F}_{ib}\hat{\mathbf{g}}_{ib}^{[b,ib]}$ as detection filter. The estimate of the signal of user i, b can be decomposed as follows:

$$\hat{\mathbf{g}}_{ib}^{[b,ib]\dagger}\mathbf{F}_{ib}^{\dagger}\frac{\mathbf{Y}_u^{[b]}}{\sqrt{\rho_u}} = \tag{3.9}$$

$$\hat{\mathbf{g}}_{ib}^{[b,ib]\dagger}(\hat{\mathbf{g}}_{ib}^{[b,ib]}d_{ib} + \tilde{\mathbf{g}}_{ib}^{[b,ib]}d_{ib} + \sum_{r=1}^{N_c}\sum_{l\neq\chi(i,b)}^{\tau}\sum_{u\in\Sigma(l,r)}\mathbf{g}_{ur}^{[b,ib]}\,d_{ur} + \sum_{r=1}^{N_c}\sum_{\substack{u\in\Sigma(\chi(i,b),r)\\ ur\neq ib}}\mathbf{g}_{ur}^{[b,ib]}\,d_{ur} + \frac{\mathbf{w}_u^{[ib]}}{\sqrt{\rho_u}}),$$

In what follows, we consider the large system limit, where M and K grow to infinity while keeping a finite ratio $\frac{K}{M}$. This assumption enables to derive a deterministic approximation of the achievable average SINR which is derived following the same approach as in [88].

Lemma 1. *In the case of a large antenna array with a spatially correlated channel, the average UL SINR of user i in cell b, with matched filter receiver and spatial signature based projection, can be written as :*

$$\gamma_{ib} = \frac{(\frac{1}{M}\operatorname{Tr}(\Psi_{ib}^{[b,ib]}))^2}{\sum\limits_{r=1}^{N_c}\sum\limits_{l\neq\chi(i,b)}^{\tau}\sum\limits_{u\in\Sigma(l,r)}\frac{\operatorname{Tr}(F_{ib}^{\dagger}\boldsymbol{R}_{ur}^{[b]}F_{ib}\Psi_{ib}^{[b,ib]}))}{M^2}+\sum_{r=1}^{N_c}\sum\limits_{\substack{u\in\Sigma(\chi(i,b),r)\\ ur\neq ib}}\frac{\operatorname{Tr}(\Psi_{ur}^{[b,ib]}))^2}{M^2}+\frac{\operatorname{Tr}(\Psi_{ib}^{[b,ib]}))}{\rho_u M^2}}, \tag{3.10}$$

where $\Psi_{ur}^{[b,ib]} = \boldsymbol{F}_{ib}^{\dagger}\boldsymbol{R}_{ur}^{[b]}\left(\frac{1}{\rho_p}\boldsymbol{I}_M+\sum_{r=1}^{N_c}\sum_{k\in\Sigma(\chi(i,b),r)}\boldsymbol{R}_{kr}^{[b]}\right)^{-1}\boldsymbol{R}_{ur}^{[b]}\boldsymbol{F}_{ib}$.

We can see in (3.10) that the achievable average SINR of any given user, in the asymptotic regime, depends on the power of copilot interference in the subspace spanned by the user's spatial signature. Although increasing the pilot reuse factor increases the number of active users, intra-cell copilot interference can be completely mitigated if the same pilot sequence is allocated to users with non-overlapping spatial signatures, i.e, independent detection matrices. Selecting copilot users in each cell is then of paramount importance. In the next section, we address the problem of intra-cell copilot interference through appropriate spatial signature-based user grouping.

3.4 An Alternative Approach to Spatial User Grouping: A Spatial Basis Coverage Problem

Exploiting the channel low-rank property in Massive MIMO transmission strategies proved to provide non-negligible gains in performance for both TDD and FDD systems. In FDD mode, spatial division multiplexing allows to reduce the CSI feedback overhead while incurring no capacity loss [25]. In TDD such methods enable to reduce training resources while mitigating the impact of pilot contamination [92]. These gains are mainly due to the capacity of spatial division methods to utilizes the independent spatial spaces of different users in order to discriminate between their signals. Spatial division based methods relay on an efficient spatial information-based user grouping. Several works proposed to perform this grouping using the classical K-mean algorithm with different proximity measures [25, 26]. K-medoids and hierarchical clustering have also been proposed [26]. A DFT-based greedy user grouping was also considered in SBEM [92].

Although the aforementioned grouping approaches get the work done and provide considerable performance increase for both FDD and TDD modes, they suffer, nevertheless, from a range of shortcomings that may limit the potential of spatial division multiplexing.

Indeed, previously proposed methods neglect a very important criterion, namely the coverage of all the available DoFs. Indeed, when applied to the spatial division problem, classical clustering approaches concentrate on the mutual distance between user channels subspaces with little regard to the final coverage of independent streams. This means that, although the condition of independent spatial information is met, the DoFs that the Massive MIMO system provides can be underexploited. Addressing these shortcomings can help boost the performance of spatial division multiplexing methods.

In this chapter, we propose a method that, in addition to the requirement of independent spatial subspaces, emphasizes on leveraging all independent streams that the Massive MIMO system can provide. Since a TDD system is considered in this work, the focus will be on addressing the ULs training bottleneck. To mitigate pilot contamination while realizing UL training for an excess of users with only τ pilot sequences, the same pilot sequence is allocated to users with minimum spatial signature overlapping which constitute a copilot group within each cell. In addition, the users within each copilot group achieve a maximum coverage of all interdependent streams. Consequently, in each cell, a total of τ copilot groups need to be constructed, each of which is associated with a distinct training sequence. In what follows, we formulate the spatial basis coverage based copilot user selection problem. We then provide efficient algorithms that enables to solve it. Two approaches are considered. The first approach is power agnostic. This means that it neglects the channel power along each beam and concentrate only on minimum spatial signature overlapping and maximum coverage without discriminating users based on their channel gains. The second approach is power aware. This means that users are prioritized based on their achievable channel gains in addition to the criterion considered in the power agnostic approach. The differences between the two approaches lie mainly in complexity and fairness. These differences will be discussed in more details further in this chapter.

3.4.1 Power Agnostic Spatial Basis Coverage

In this subsection, we focus on solving the spatial basis coverage copilot user selection problem in a power agnostic approach. In this case, the BSs know only the set of DFT beams that concentrate a large percentage of channel power (i.e, $\mathbf{F}_{ib}^{[b]}, \forall i, b$). The user-beam association is performed as specified in the previous section thanks to Massive MIMO training (3.4) [110] or using UL preamble [92]. As previously discussed, the BSs perform spatial basis coverage based copilot user selection in order to schedule users for UL training. This is done in order to deal with intra-cell copilot interference. The out of cell copilot interference will be addressed later in this chapter. In the power agnostic case, the BSs do not take into consideration the achievable gain along each beam. Consequently, in this case, the problem reduces to scheduling users with minimum spatial signature overlapping and maximum coverage of the DFT beams. This actually simplifies the problem at hand and enables to derive the desired grouping with low complexity. The power agnostic approach is also characterized by the upside of fairness since it does not discriminate scheduled users based on their channel power. However, this means that more flexibility should be allowed when constructing copilot users. In fact, since we cannot prioritize users based on their channel gain, it may be wise

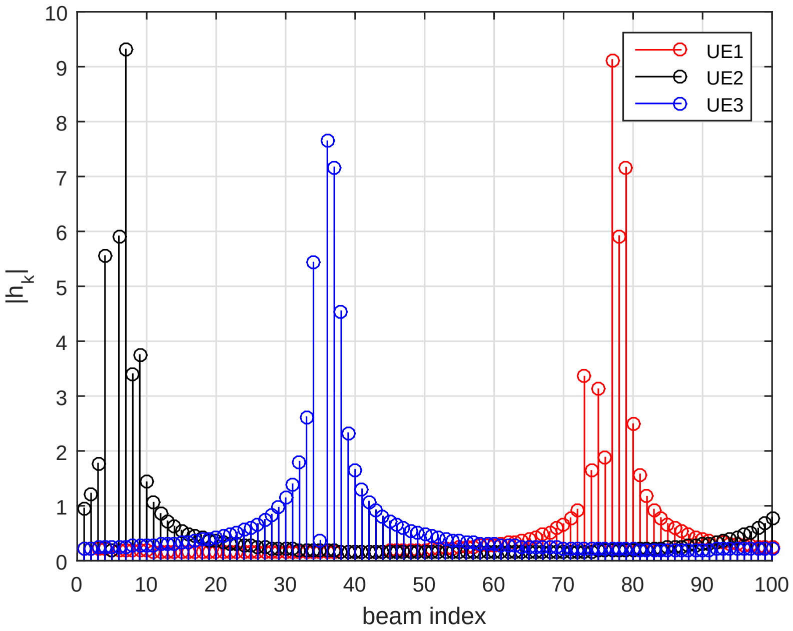

Figure 3.2: Example of spatial basis coverage for $M = 100$

to allow for some spatial overlapping. We also allow for another degree of flexibility in this problem, namely, pilot reuse in each cell. In fact, in this work, we consider that the reuse factor of each pilot sequence can vary from one cell to the other. This implemented in order to allow for a more flexible specific training sequences allocation to the copilot groups when dealing with inter-cell copilot interference. The main principle of the spatial basis coverage problem is depicted in Figure (3.2).

We consider τ copilot groups (covers) per cell $C_k^{[b]}, k = 1, \ldots, \tau, b = 1, \ldots, N_c$. Each copilot group in each cell will be associated with a distinct pilot sequence. We start by defining $x_{i,b}^{[k]}$, $\forall i, b, k$ and $y_{s,b}^{[k]}$, $\forall s = 1, ..., M, b = 1, ..., N_c, k = 1, ..., \tau$, which are given by

$$
\begin{aligned}
x_{\{i,b\}}^{[k]} &= \begin{cases} 1 & \text{if user } i, b \text{ is selected in copilot group } C_k^{[b]}. \\ 0 & \text{otherwise.} \end{cases} \\
y_{s,b}^{[k]} &= \begin{cases} 1 & \text{if beam } \mathbf{f}_s \text{ is covered in cell } b \text{ and copilot group } C_k^{[b]}. \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.11}
$$

Formally, under the power agnostic approach, the Beam coverage based copilot UE selection problem can be formulated as follows:

$$\max_{Y} \sum_{k=1}^{\tau} \sum_{b=1}^{C} \sum_{s=1}^{M} y_{s,b}^{[k]} \tag{3.12}$$

$$\text{subject to} \sum_{i} x_{\{i,b\}}^{[k]} \leq U_b^{[k]} \ \ \forall k = 1...\tau, \ \ \forall b = 1, ..., N_c \tag{2.12a}$$

$$\sum_{i, f_s \in F_{ib}} x_{\{i,b\}}^{[k]} \geq y_{s,b}^{[k]} \ \ \forall k = 1...\tau, \ \ \forall b = 1, ..., N_c, \tag{2.12b}$$

$(2.12a)$ guarantees that the number of users in a given copilot group $C_k^{[b]}, k = 1, \dots, \tau, b = 1, \dots, N_c$, is upper bounded by $U_b^{[k]}, k = 1, \dots, \tau, b = 1, \dots, N_c$. Note that $U_b^{[k]}$ is a design parameter that defines the reuse factor of a given pilot sequence in each cell. Depending on the considered setting, $U_b^{[k]}$ can be the same or differs from one cell to the other. $(2.12b)$ guarantees that, for any covered beam f_s in cell b, in copilot group $C_k^{[b]}$, at least one user i, b with $\mathbf{f}_s \in \mathbf{F}_{ib}^{[b]}$ is scheduled for UL training in copilot group $C_k^{[b]}$. We start by showing the computational intractability of problem (3.12).

Lemma 2. *The considered spatial basis coverage based copilot UE selection problem* (3.12) *is NP-hard.*

Proof. For $C = 1$ and $\tau = 1$, (3.12) is equivalent to a *maximum coverage problem* which is known to be NP-hard [80]. Consequently, (3.12) is also NP-hard. □

The proof of computational intractability provides us with insight on how to solve this problem in a low complexity manner.

In order to solve (3.12), we use two nested greedy phases. In the upper phase, the algorithm produces τ maximum coverages of the DFT matrix vectors (**F**), in each cell. The maximum covers $C_k^{[b]}, k = 1..\tau, b = 1, .., N_c$, are computed successively in a greedy manner. Each of the maximum covers is computed using another greedy method that goes as follows. For each $C_k^{[b]}, k = 1..\tau, b = 1, .., N_c$, the set of uncovered beams is initialized as the vectors of the DFT matrix. Then users are added to $C_k^{[b]}$ successively while selecting, at each iteration, the user with the spatial signature that cover a maximum of the uncovered DFT columns. This procedure is repeated until attaining the reuse constraint $U_b^{[k]}$ for each copilot group, in each cell. Different from previously proposed algorithms, the present approach enables to satisfy the spatial independence requirements within each copilot group while offering a maximum utilization of the excess of DoFs. The detailed algorithm is given in table 3.1. We denote by $\Gamma(b)$ the set of users in cell $b = 1, \dots, N_c$.

```
Initialize: Copilot groups sets $C_k^{[b]} = \emptyset, k = 1, \dots, \tau, b = 1, \dots, N_c$,
User specific beam matrices $\mathbf{F}_{ib}, \forall i \in \Gamma(b), b = 1, \dots, N_c$
For $b = 1 : N_c$ do:
    For $k = 1 : \tau$ do:
        Define the set $Un = \mathbf{F}$ as the set of uncovered beams.
        For $j = 1 : U_b^{[k]}$ do:
            $i^* \longleftarrow \underset{i \in \Gamma(b)}{\text{argmax}} |\mathbf{F}_{ib} \cap Un|$
            $Un \longleftarrow Un \setminus \{\mathbf{F}_{i^* b} \cap Un\}$
            $C_k^{[b]} \longleftarrow C_k^{[b]} \bigcup i^*$
        End for
    End for
End for
```

Table 3.1: **Power Agnostic Beam coverage based copilot UE selection**

The algorithm in table 3.1 produces τ copilot user groups in each cell. Each copilot group maximizes the coverage of the DFT beams while minimizing the overlapping between copilot users spatial signatures. Consequently, the CSI estimation of each user can be enhanced by a simple linear projection. Note that the proposed algorithm allows for some subspace overlapping. This is actually needed since users are not prioritized based on their channel gains. We now proceed by deriving the performance guarantee of the proposed algorithm.

Theorem 3. *The algorithm in table 3.1 provides an* $(1 - (\frac{\tau-1}{\tau})^{\tau})(1 - \frac{1}{e})$*-approximation of the optimal solution of problem* (3.12).

Proof. See A.1. □

3.4.2 Power Aware Spatial Basis Coverage

In the power aware approach, users are prioritized based on the power of their signals along each direction. The resulting problem provides a more efficient grouping since it takes into consideration the overlapping between copilot users spatial signatures, the coverage of the signal space and the power of each user channel. However, this efficiency comes at the price of augmented complexity and reduced fairness since users are discriminated based on their channel gains. In this case, we define a different value for each beam depending on which user is covering it. For each user i, b, the value associated with beam $\mathbf{f}_s, s = 1, \ldots, M$ is given by $\zeta_{ib}^{[s]}$, where $\zeta_{ib}^{[s]}$ is the power of user i, b channel along $\mathbf{f}_s, s = 1, \ldots, M$. This consideration changes the formulation of the spatial basis coverage based copilot UE selection problem (3.12). The main idea of providing maximum coverage of the DFT beams, in each cell and for each pilot sequence, still holds but the actual gain associated with each beam will also be taken into consideration. The resulting combinatorial optimization problem can be formulated as follows

$$\max_{Y} \sum_{k=1}^{\tau} \sum_{b=1}^{N_c} \sum_{i \in \Gamma(b)} \sum_{f_s \in \mathbf{F}} \zeta_{ib}^{[s]} y_{\{i,b\}}^{[s,k]} \tag{3.13}$$

$$\text{subject to} \quad \sum_{i \in \Gamma(b), f_s \in F_{ib}} y_{\{i,b\}}^{[s,k]} \leq 1 \;\; \forall k = 1...\tau, \;\; \forall b = 1...N_c \tag{2.13a}$$

$$\sum_{i \in \Gamma(b), f_s \in F_{ib}} x_{\{i,b\}}^{[k]} \geq y_{\{i,b\}}^{[s,k]} \;\; \forall k = 1...\tau, \;\; \forall b = 1...N_c, \tag{2.13b}$$

$$\sum_{i \in \Gamma(b)} x_{\{i,b\}}^{[k]} \leq U_b^{[k]} \;\; \forall k = 1...\tau, \;\; \forall b = 1...N_c, \tag{2.13c}$$

The constraints, in $(2.13a)$, guarantees that each beam is covered by at most one user. $(2.13b)$ guarantees that, for any covered beam f_s in cell b, in copilot group $C_k^{[b]}$, at least one user i, b with $\mathbf{f}_s \in \mathbf{F}_{ib}$ is scheduled for UL training in copilot group $C_k^{[b]}$. $(2.13c)$ guarantees that the total number of the users associated with a given pilot sequence in a given cell is bounded. The difference between (3.12) and (3.13) is mainly the fact that the actual gain along each stream is taken into consideration. This means that, the BS can optimize its pilot allocation accordingly with the final aim of maximizing the total weight of the covered streams. This means that, in addition to reducing copilot interference thanks to the non-overlapping spatial signatures, the users are selected such that the achievable gain along all the available independent streams is maximized. We start by showing the computational intractability of problem (3.13).

Lemma 4. *The considered spatial basis coverage based copilot UE selection problem* (3.13) *is NP-hard.*

Proof. For $C = 1$ and $\tau = 1$, the optimization problem (3.13) is equivalent to a *Generalized*

Maximum Coverage Problem (GMC) which is known to be NP-hard. Consequently, (3.13) is also NP-hard.

□

The proof of computational intractability provides us with insight on how to solve problem 3.13 efficiently. Indeed, to solve 3.13, we adopt a successive coverage approach as the previous algorithm. The difference here comes in the construction of each copilot group where the actual power along each beam needs to be considered. To this end, $\forall\ k = 1..\tau, b = 1,..,N_c$, a GMC problem is solved. Solving the GMC problem that produces each copilot group will be performed based on a modification of the coverage algorithm in [80].

The present approach enables to satisfy the spatial independence requirements within each copilot group while offering a maximum utilization of the excess of DoFs. Since it takes into consideration the actual power of users signals along each beam, the present approach enables also to discriminate between users based on their channel gain in each direction. This, ultimately, results in more efficient utilization of the system's DoFs by prioritizing users with high signal power in each direction. Before providing the detailed algorithm to solve (3.13), some definitions are now in order. We define a user allocation A as a triple $A = (\phi, \xi, h)$, where ϕ represents the set of selected users, ξ denotes the set of corresponding covered beams and h is an assignment from ξ to ϕ such that $\forall, f_s \in \xi, h(f_s)$ denotes the user covering beam f_s. For a given allocation A, we define $V(A) = \sum_{f_s \in \xi} \zeta^{[s]}_{h(f_s),b}$ as the value of A and $W(A) = |\phi|$ as its weight. We also define the residual value of $([i,b], f_s)$ with respect to A as follows

$$V_A([i,b], f_s) = \begin{cases} \zeta^{[s]}_{i,b} & \text{if } f_s \text{ is not covered by A.} \\ \zeta^{[s]}_{i,b} - \zeta^{[s]}_{h(f_s),b} & \text{otherwise.} \end{cases} \tag{3.14}$$

The main idea of the algorithm is to use two nested greedy phases. In the upper phase, the maximum coverages $C^{[b]}_k, k = 1..\tau, b = 1,..,N_c$ are computed successively in a greedy manner. In order to obtain each coverage, in the lower phase, the algorithm uses the residual value 3.14 in a greedy procedure so as to choose a subset of DFT beams that are part of a given user spatial signature with the highest density. Users are then added to the selection as long as their residual value is positive. When the greedy phase ends, its resulting selection is compared with the highest value of a single user. The selection with the highest value is then selected. This procedure is repeated in the each cell $b = 1,..,N_c$ for each copilot group $k = 1..\tau$. We now present the detailed algorithm in table 3.2.

```
For $b = 1 : N_c$ do:
    For $k = 1 : \tau$ do:
        $A_g \longleftarrow$ Greedy($S = \{i, i \in \Gamma(b)\}$ )
        Find the single user $i^*, b$ with most efficient coverage of the reference beams.
        $V(A_s) \longleftarrow \sum_{\mathbf{f}_s, \mathbf{f}_s \in \mathbf{F}_{i^*b}} \zeta_{i^*b}^{[s]}$
        If $V(A_g) \geq V(A_s)$:
            $C_k^{[b]} \longleftarrow (A_g)$
        Else
            $C_k^{[b]} \longleftarrow (A_s)$
    End for
End for
```

Table 3.2: **Power Aware Beam coverage based copilot UE selection**

```
$j \longleftarrow 0$
While new beams with positive residual value can be added to $A$ without violating the cardinality constraint $U_r^{[k]}$ do:
    Use the greedy algorithm for Knapsack problems in order to derive $([i^*, r], \mathbf{F}_{i^*r})$ such that $W_A([i^*, r], \mathbf{F}_{i^*r}) \leq U_r^{[k]} - W(A)$, which has the maximum density
    $A \longleftarrow A \oplus ([i^*, r], \mathbf{F}_{i^*r})$
    For $u \notin A$ do:
        If $W_{A \oplus ([u,r], \mathbf{F}_{ur})}([u, r], \mathbf{F}_{ur}) \leq 0$ and $\forall \mathbf{f} \in \mathbf{F}_{ur}, V_{A \oplus ([u,r], \mathbf{F}_{ur})}([u, r], \mathbf{f}) > 0$ do:
            $A \longleftarrow A \oplus ([u, r], \mathbf{F}_{ur})$
    End for
    $j \longleftarrow j + 1$
End While
Return($A$)
```

Table 3.3: **Greedy(S)**

The algorithm in table 3.2 consists of solving a generalized maximum coverage problem for each copilot group $C_k^{[b]}, k = 1, \ldots, \tau, b = 1, \ldots, N_c$, successively. We now derive the performance guarantee of the proposed algorithm.

Theorem 5. *The algorithm in table 3.2 provides an* $(1 - (\frac{\tau-1}{\tau})^{\tau})\frac{\frac{3}{2} - \frac{e^{-2}}{2}}{1-e^{-2}}$*-approximation of the optimal solution of problem* (3.13).

Proof. See A.2. □

The proposed algorithm in table 3.2 provide τ covers of the DFT matrix beams in each cell. This results in τ copilot user groups, in each cell, that fully exploit all available DoFs with minimum overlapping between the beam sets of each user. This leads to an efficient reduction of intra-cell copilot interference.

The main differences between the power agnostic and aware cases are performance guarantees and complexity. Indeed, the simplified power agnostic case enables to achieve the desired grouping with good performance guarantee (see Theorem 3) and low complexity. The power aware case provide a more efficient grouping since it takes into consideration the users channel gains. Nevertheless, this comes with a penalty in the approximation ratio (see Theorem 5) and a higher computational complexity. Constructing the copilot user groups enables to efficiently address the issue of intra-cell interference. Nevertheless, further spectral efficiency gain can be achieved by addressing the issue of inter-cell copilot interference. This will be the focus of the next section.

3.5 Cross Cell Pilot Allocation: A Graphical Approach

3.5.1 Cross Cell Pilot Allocation Problem

The proposed spatial basis coverage copilot UE selection approach enables to manage intra-cell copilot interference. Using predefined spatial signature-based detection filters provides the means to substantially reduce intra-cell copilot interference. Nevertheless, another performance limiting factor, namely inter-cell copilot interference, needs to be treated. This can be achieved by leveraging the spatial signatures of the inter-cell interference channels. Adopting the same approach as in the previous section will not be as fruitful in this case. Indeed, constructing copilot groups across cells, in order to address the problems of intra-cell and inter-cell interference simultaneously, proves to be quite complex. This is due to the complexity of defining a proper grouping metric that is based on the spatial signatures of both useful and interference links.

This fact motivated the present approach of dealing with interference through two consecutive subproblems. A major advantage of such division is the reduction of complexity. Indeed, since copilot user groups have already been constructed in each cell, addressing out-of-cell copilot interference reduces to allocating specific training sequences to copilot groups. If copilot interference is to be addressed from the beginning as a whole, it will result in a complex pilot allocation problem among all users in all cells which proves to be complicated.

Practically, complete removal of interference is not physically possible. In addition, copilot user grouping was performed with the clear goal of managing intra-cell interference. Consequently, when dealing with inter-cell copilot interference, previously formed copilot groups should be maintained. We propose a scheme in which pilot allocation is done such that high interference links are suppressed when spatial signature based receivers are used. In this section, we address this problem using an intelligent pilot assignment scheme. The basic idea is to infer inter-cell copilot interference from the spatial signatures of interference links. A training phase to obtain spatial information of the interference links is required. This can be implemented without a large signaling overhead owing to the slow changing spatial information. We now consider that, each user $i, b, i = 1...K,\ b = 1...N_c$ is associated with N_c matrices $\{\mathbf{F}_{ib}^{[r]}, r = 1...N_c\}$. Each matrix $\mathbf{F}_{ib}^{[r]}$ is constructed in a similar manner to (3.4) as follows

$$\mathbf{F}_{ib}^{[r]} = \{\mathbf{f}_s \in \mathbf{F}, \frac{\left\|\mathbf{g}_{ib}^{[r]\dagger}\mathbf{f}_s\right\|^2}{Tr(\mathbf{R}_{ib}^{[r]})} \geq \alpha\}, \tag{3.15}$$

3.5.2 Graphical Modeling and Proposed Solution

The first step to manage inter-cell copilot interference is to construct an interference graph that corresponds to the considered system setting. We construct an undirected interference graph $\mathcal{G}(\mathcal{C}, \mathcal{E})$. Each node $C_k^{[b]} \in \mathcal{C},\ k = 1...\tau, b = 1...N_c$ represents copilot user group of index k in a give cell b. Each edge in $e_{C_b^{[j]}, C_l^{[k]}} \in \mathcal{E}$ represents an interference link and is associated with a given weight $w_{C_b^{[j]}, C_l^{[k]}}$. We propose a method for determining the edge weight without accurate SINR measurements since it cannot be obtained before pilot allocation. In this work, we propose to infer interference levels form spatial information. This consideration is due to the practical low signaling overhead that is required.

Since the weight of each edge quantifies the level of interference between two copilot groups, an appropriate measure need to be considered. This task is not an easy one since the weight of the edge between two different copilot groups need to properly characterize the levels of resulting interference. The research papers that investigated spatial division multiplexing proposed different metrics to characterize subspace distances. The most used one is chordal distance. Such metric has the considerable downside of neglecting the actual signal power in each subspace. In addition, the present framework implies that the weight of each edge need to characterize the mutual interference between two groups of users. Consequently, defining a distance measure between copilot groups is a major issue in our case. In order to solve this issue, we call upon hierarchical clustering where measuring distances between groups is commonly encountered. We adopt a linkage method in hierarchical clustering [26], namely weighted average linkage. To quantify interference on each link, we use spatial signature overlapping between the users forming the two copilot groups which is obtained using the chordal distance between spatial signatures. The weight of each edge $w_{C_b^{[j]}, C_l^{[k]}}$ is then given by

$$w_{C_b^{[j]}, C_l^{[k]}} = \min_{y \in C_b^{[j]}, z \in C_l^{[k]}} \{\frac{1}{2}\|\mathbf{F}_{yb}^{[b]}\mathbf{F}_{yb}^{[b]\dagger} - \mathbf{F}_{zl}^{[b]}\mathbf{F}_{zl}^{[b]\dagger}\|_F^2 + \frac{1}{2}\|\mathbf{F}_{yb}^{[l]}\mathbf{F}_{yb}^{[l]\dagger} - \mathbf{F}_{zl}^{[l]}\mathbf{F}_{zl}^{[l]\dagger}\|_F^2\}, \tag{3.16}$$

Here $||\mathbf{F}^{[b]}_{yb}\mathbf{F}^{[b]^{\dagger}}_{yb} - \mathbf{F}^{[b]}_{zl}\mathbf{F}^{[b]^{\dagger}}_{zl}||^2_F$ represents the chordal distance between the spatial signatures of the useful signal of user $y \in C^{[j]}_b$ and the interference generated by $z \in C^{[k]}_l$. $||\mathbf{F}^{[l]}_{yb}\mathbf{F}^{[l]^{\dagger}}_{yb} - \mathbf{F}^{[l]}_{zl}\mathbf{F}^{[l]^{\dagger}}_{zl}||^2_F$ denotes the chordal distance between the spatial signatures of the useful signal of user $z \in C^{[k]}_l$ and the interference generated by $y \in C^{[j]}_b$.

The weight expression (3.16) capture the minimum chordal distance between the spatial signatures of the interference and useful signals for all users in the two copilot groups and is inspired by the single and weighted average linkage, commonly used in hierarchical clustering [26].

In each cell, users from the same copilot group are the only devices allowed to transmit the same UL training sequence. Consequently, during pilot allocation, we need to make sure that any given pilot sequence $\mathbf{q}_l, l = 1...\tau$ should be allocated to only one copilot group in each cell. In order to do so, the weight of the links between copilot groups from the same cell will be given a very large value w_∞, because intra-cell interference between copilot groups must be avoided. Using this metric we are able to construct the interference graph $\mathcal{G}(\mathcal{C}, \mathcal{E})$, which is a first step in the proposed pilot allocation scheme. An illustration of $\mathcal{G}$ is presented in figure 3 for the case of $N_c = 3$ and $\tau = 2$.

The considered pilot allocation problem is closely related to MAX-CUT problem [81]. Indeed, the task of interference management in our problem reduces to suppressing high pilot contamination between copilot groups. This can be performed by allocating the same training sequence to copilot groups with minimum mutual interference weights. In the considered graphical framework, this task is equivalent to partitioning the interference graph into τ subgraphs where the copilot groups in each subgraph will be allocated the same training sequence. In the graph theory, a cut is a partition of the vertices of the graph into multiple sets or clusters. The size of a cut is the total number of edges that cross the cut. In our weighted graphs, the size of the cut is the sum of weights of the edges that cross the cut. A cut is said to be maximal if the size of the cut is not smaller than the size of any other cut. By generalizing this notion to τ cuts, the MAX-τ-CUT problem is to find a set of τ cuts that are not smaller in size than any other τ cuts. Given τ UL training sequences and N_c cells, containing each τ copilot groups, our pilot allocation problem is a MAX-τ-CUT problem on the interference graph and can be stated as follows:

Pilot sequence allocation problem: Given the interference graph $\mathcal{G}(\mathcal{C}, \mathcal{E})$ with $\tau \times N_c$ nodes and edge weight $w_{C^{[j]}_b, C^{[k]}_l}$ for each edge $e_{C^{[j]}_b, C^{[k]}_l} \in \mathcal{E}$, partition the graph into τ disjoint sets $P_g, g = 1, ..., \tau$, such that
$\sum_{\substack{C^{[j]}_b \in P_g, C^{[k]}_l \in P_{g'} \\ g \neq g'}} w_{C^{[j]}_b, C^{[k]}_l}$ is maximized.

The training sequence length constraint is already taken into consideration by the defini-

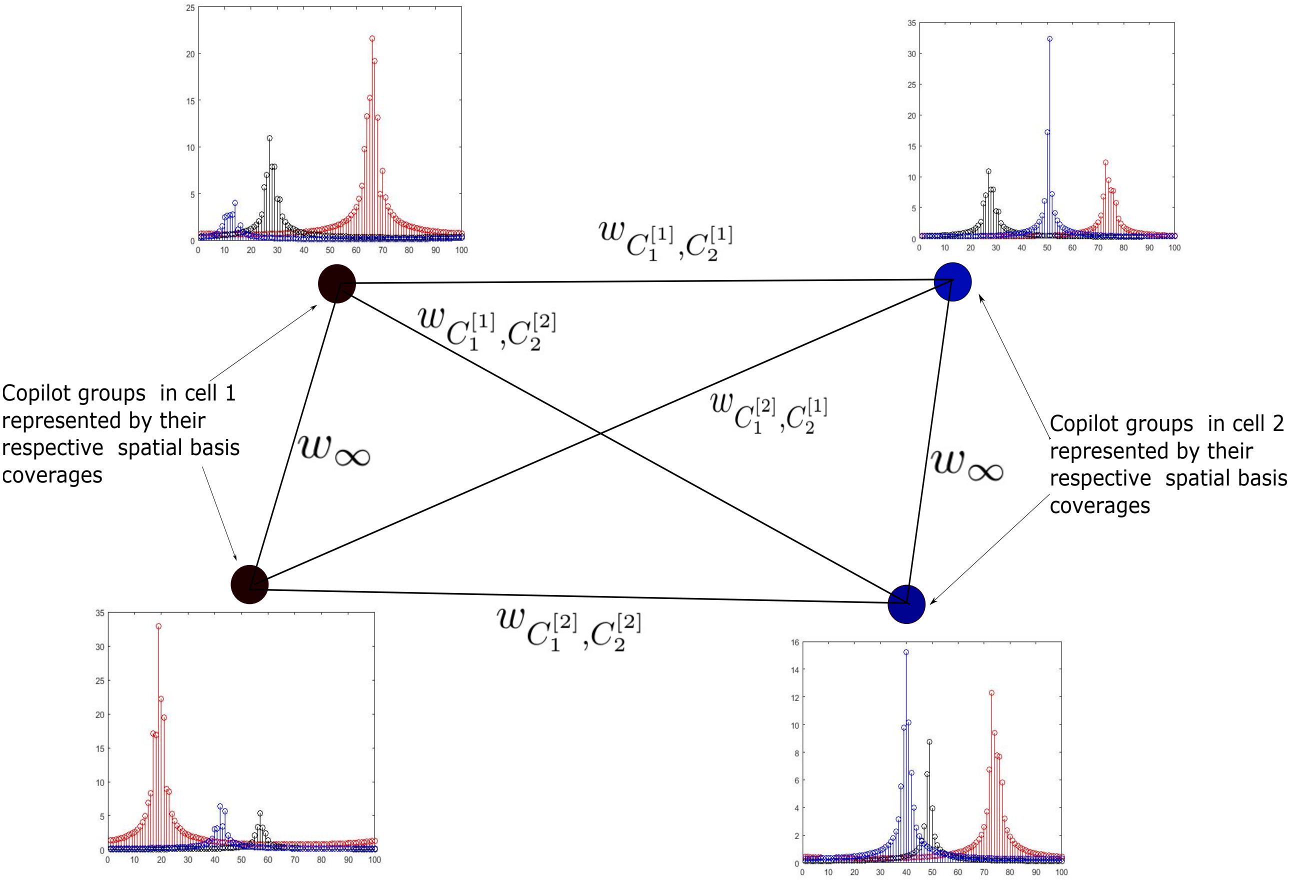
$w_{C_1^{[1]},C_2^{[1]}}$
$w_{C_1^{[1]},C_2^{[2]}}$
$w_{C_1^{[2]},C_2^{[1]}}$
$w_{C_1^{[2]},C_2^{[2]}}$
w_∞
w_∞
Copilot groups in cell 1 represented by their respective spatial basis coverages
Copilot groups in cell 2 represented by their respective spatial basis coverages

Figure 3.3: Interference graph example

tion of the number of resulting sets τ. Since interference links between copilot users in the same cell was assigned a large weight w_∞, we are sure that all copilot groups within a given cell will be allocated to different sets. The max-τ- cut algorithm assigns different training sequences to copilot groups with strong spatial signatures overlapping between the useful and interference signals. The complexity of the proposed pilot allocation algorithm depends on the number of copilot groups, edges and training sequences.

A remark on the complexity of this algorithm is now in order. Proceeding to sequence allocation, once copilot groups are formed, results in a substantial simplification of the problem. In fact, instead of processing each user individually, the proposed method exploits the previously formed copilot groups in order to reduce running time of the pilot allocation algorithm. This impact becomes very interesting in an IoT communication scenario. In fact with a high number of scheduled devices per copilot group, it makes sens to implement this grouping in order to deal with inter and intra cell copilot interference, separately, while reducing the running time of the algorithm.

The Pilot sequence allocation problem is NP-hard in a graph constituted of a large number of nodes [187]. Meaning that the optimal solution is computationally prohibitive to obtain. Consequently, we use the low complexity algorithm in [81]. The heuristic algorithm, in [81], provides an approximate solution that achieves at a ratio of $(1 - \frac{1}{\tau})$ of the optimal one for a general MAX-τ-CUT problem, given that all weights in the graph are nonnegative integers. Since all weights in the considered interference graph are positive, using the heuristic from [81] provides us with a $(1 - \frac{1}{\tau})$-approximation of the optimal solution for the considered problem. The detailed algorithm for cross cell Pilot assignment is given in table 3.4.

```
Initialize: intra-set weights $W_g = 0, \forall g = 1...\tau$ $P_g = \emptyset, g = 1, ..., \tau$

1. Assign the $\tau$ copilot groups in cell $1$ to different pilot sets

2. Randomly order the rest of copilot groups.

3. Select the next copilot group $v$ and assign it to set $g^*$ for which

$W_{g^*}^v$ is minimized where $W_g^v = \sum_{u \in P_g} w_{v,u}$

4. Update the Average weight of group $g^*$ such that $W_{g^*} = W_{g^*} + W_{g^*}^v$

5. Repeat steps $3 - 4$ until all copilot groups are assigned.
```

Table 3.4: **Cross cell Pilot assignment algorithm**

3.6 Numerical Results And Discussion

In this section, we provide numerical results demonstrating the performance of the proposed spatial basis coverage copilot user selection. We compare the proposed approach with a conventional TDD Massive MIMO system. We then extend the simulation results to include MAX-τ-CUT pilot allocation.

We consider a network constituted of $N_c = 4$ hexagonal cells. Each cell has a radius 0.5 Km from center to vertex. Each cell contain a Massive MIMO BS at its center, equipped with $M = 100$ equally spaced isotropic antennas. The minimum distance between antenna elements is equal to $\frac{\lambda}{2}$. Each cell contains $K = 20$ users with randomly generated mean direction of arrivals. The channel vectors of the different users are generated according to 3.1 where $P = 50$. Each coefficient $\mu_{ib}^{[r]^2}, \forall i, b, r$ denotes the path-loss between the user and the target BS. The path-loss coefficient is 2.8. For each user i, b, the angles of its rays

$\theta_{ib}^{[r,p]}, p = 1, \ldots, P$ are uniformly distributed in the interval $\left[\theta_{ib}^{[r]} - \Delta\theta_{ib}^{[r]}, \theta_{ib}^{[r]} + \Delta\theta_{ib}^{[r]}\right]$ where the Angular spread (AS) is supposed to be the same for all users with $\Delta\theta_{ib}^{[r]} = \Delta = 4°$. The coherence interval is set to $T_s = 200$, split between training and data transmission. We take $\alpha = 0.01$. In order to assess the accuracy of channel estimation, we take as metric the average individual mean square error.

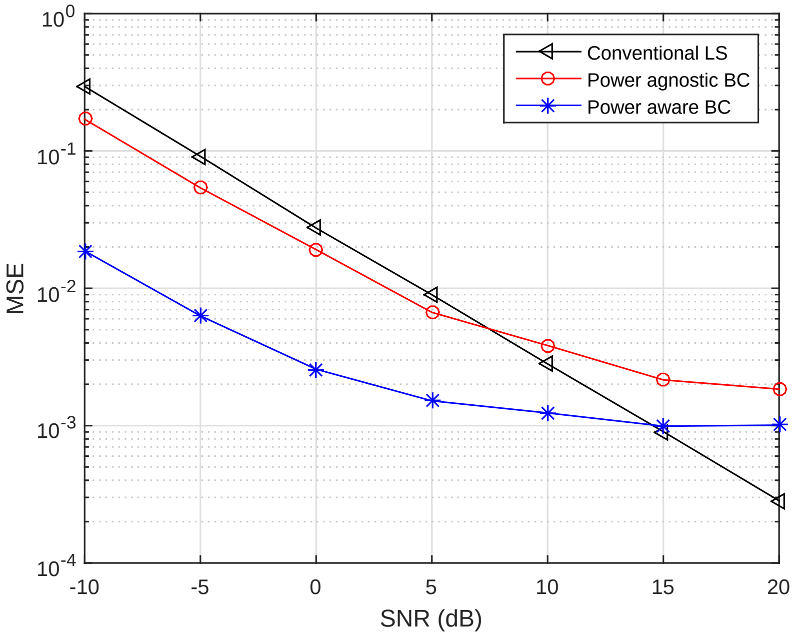

Figure 3.4: Comparison of uplink channel estimation MSE with $\tau = 5$ and $U_b^{[k]} = 4, \forall k, b$

Figure (3.4) illustrates a comparison of Mean square error (MSE) performances of UL channel estimation, as a function of SNR. Figure (3.4) shows that UL channel estimation is improved when using the two proposed spatial basis coverage algorithms in the low SNR range (up to approximately 7.5dB for the power agnostic approach and up to 15 dB for the power aware approach). It shows that power aware spatial basis coverage outperform the power agnostic approach. This is mainly due to the fact that the power aware approach do not allow any overlapping between users spatial signatures while some overlapping is permitted in the power agnostic approach. Figure (3.4) shows also that, as SNR increases, the performances of the two proposed algorithms reach two distinct error floors. This phenomenon is due to the truncation error that results from projection on the users specific signal subspaces. These error floors depend on the rank of the user's spatial signatures and can be reduced by considering the DFT beams that concentrate lower levels of the user channel power (This can be done by decreasing α in (3.4)) .

Figure (3.5) illustrates a comparison of Cumulative density function (CDF)s of the achievable SEs between the proposed spatial basis coverage algorithms and a conventional TDD Massive MIMOs system, for different SNR values. Figure (3.5) shows that, for an SNR of

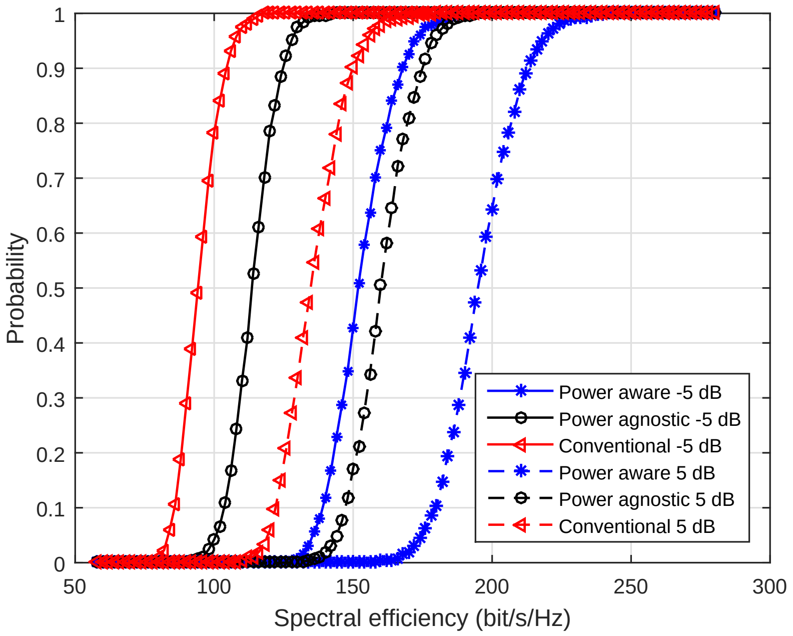

Figure 3.5: Comparison of CDFs of achievable SE for different SNR values with $\tau = 5$ and $U_b^{[k]} = 4, \forall k, b$

-5dB, the power aware spatial basis coverage and the power agnostic spatial basis coverage approaches achieves 5% outage rate around 136 bit/s/Hz and 102 bit/s/Hz, respectively. This represents gains of 52bit/s/Hz and 18bit/s/Hz, respectively, in comparison with a conventional TDD Massive MIMO system. For 5dB, these gains become 56bit/s/Hz and 24bit/s/Hz, respectively. This increase is mainly due to the reduced impact of additive noise since the system becomes interference limited which emphasizes the ability of the proposed schemes to mitigate intra-cell copilot interference.

Figure (3.6) illustrates a comparison of CDFs of the achievable SE between the proposed spatial basis coverage algorithms and a conventional TDD Massive MIMO system, for different for different τ and $U_b^{[k]}$ values. Figure (3.6) shows that, for $\tau = 10$ and $U_b^{[k]} = 2\forall k, b$, the power aware spatial basis coverage and the power agnostic spatial basis coverage approaches achieves 5% outage rate around 136 bit/s/Hz and 102 bit/s/Hz, respectively.

Figure (3.7) illustrates the impact of the proposed max-τ-cut pilot allocation algorithm. Figure (3.7) shows that addressing the issue of inter-cell copilot interference through efficient pilot sequence allocation results in an improvement in the system SE. Indeed, while the power aware spatial basis cover approach achieves 5% outage rate around 202 bit/s/Hz, the combination with the max-τ-cut pilot assignment algorithm result achieves 5% outage rate around 216 bit/s/Hz. Consequently, after constructing copilot user groups based on the spatial basis approach, the same diversity in spatial signatures can be leveraged in order to address the problem of inter-cell copilot interference. Although complete removal of interference is still not possible, especially since copilot user groups are constructed based on the

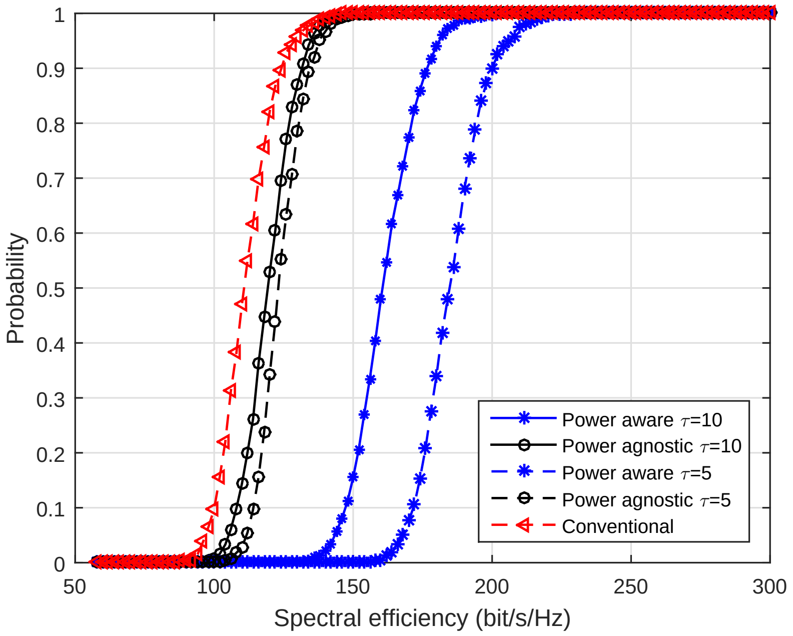

Figure 3.6: Comparison of CDFs of achievable SE for different τ and $U_b^{[k]}$ values with $SNR = 0$ dB

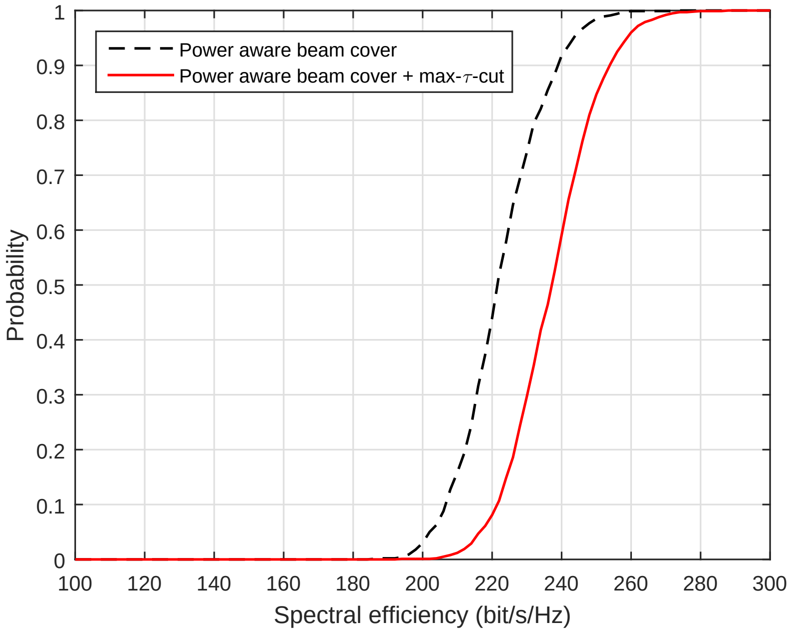

Figure 3.7: Comparison of CDFs of achievable SE for $\tau = 4$, $U_b^{[k]} = 5$ and with $SNR = 0$ dB

useful links spatial information, non-negligible performance improvement can be achieved by efficient pilot sequence allocation across cells.

3.7 Closing Remarks

In this chapter, we have studied user scheduling and pilot allocation based on spatial division multiplexing for TDD Massive MIMO systems. We proposed a copilot user grouping approach based on spatial basis cover coverage. After associating each user with the DFT matrix vectors that concentrate the majority of its channel power, users are assigned to copilot groups in order to achieve a maximum coverage of the DFT vectors per group. The proposed approach enables to increase the SE without requiring more training overhead. This is made possible by more aggressive pilot reuse within each cell taking into account spatial diversity. Various numerical results were provided to demonstrate the effectiveness of the proposed method. In order to efficiently manage inter-cell copilot interference, further optimization is performed. We have proposed a graphical approach for training sequence allocation across cells. The proposed approach exploits the interference links spatial information in order to reduce the impact of inter-cell copilot interference. The training sequence allocation optimization is formulated as a max-cut problem. Consequently, we are able to provide a low complexity algorithm that allocates training sequences to copilot groups while minimizing interference thanks to spatial diversity. Although, the proposed approach does not remove entirely the bottleneck of Massive MIMO systems, namely copilot interference, it provides a practical method to increase the achievable SE for the same training overhead. It also provides an efficient tool to increase the number of connected devises which is an essential requirement for 5G and beyond networks.

Chapter 4

Enhancing performance by long term CSI estimation planning

4.1 Overview

Mobile communication networks are shifting from being dominated by voice traffic with symmetric DL/UL capacity needs to more asymmetric data traffic [3]. Consequently, the conventional design of wireless networks including fixed FDD or TDD settings with little flexibility for varying the capacity split between resources needs to be phased out in future generation. In addition, the advent of 5G with high density connectivity and spectral efficiency requirements, imposes a more efficient and flexible use of time-frequency resources. In this context, Dynamic TDD is starting to attract more attention owing to its flexible utilization of the available radio resources by means of dynamic adaptation of the TDD UL / DL subframe configuration [119]. Thereby Dynamic TDD can greatly improve user experience especially in low to medium load.

Another bottleneck that a dynamic allocation of TDD resource can address is UL training. In fact, in TDD systems, the actual number of active users is restricted due to training overhead and to the limited coherence time. Consequently, increasing connection density can be enabled by optimizing the training procedure so that the network can acquire a maximum amount of CSI for as little as possible training. The importance of such optimization is taken to another dimension in massive MIMO systems [6]. In order to be able to achieve the needed energy and spectral efficiency gains, accurate CSI estimates is required at the BS end. In TDD systems, CSI estimates are acquired using UL training with orthogonal pilot sequences [13]. Consequently, addressing the UL training bottleneck through an adaptive TDD frame structure seems to be quite logical. Nevertheless, this optimization needs to rest on user-specific information with the final goal of increasing the networks performance. The defining parameter that was chosen to achieve this is *Doppler spread* or, equivalently, the wireless channel *coherence time*. In fact, based on the observation that current wireless systems assume the same time slot duration for all devices regardless of the fact that users are subject to heterogeneous Doppler spreads, we notice a degree of freedom that was previously neglected, namely, *CSI estimation periodicity*. As a matter of fact, the coherence slot duration in current wireless systems is based on the maximum supported Doppler spread and the CSI estimation overhead is defined accordingly [1]. This approach is suboptimal since it implies that the network is going to spend precious resources on estimating information that may be reusable. This is particularly the case for users with low mobility.

In this chapter, we address this issue by exploiting the heterogeneous channel aging effect among users. We propose a novel approach for UL training in massive MIMO systems that defines the needed training resources dynamically, at each time slot, depending on all available information about users. We notice that further improvement can be achieved by considering a training optimization over time, while taking into consideration user mobility. In fact, since Doppler spread results from mobility, with velocity being a defining parameter, it makes sense to consider changes in user locations. In order to enable the network to cope with mobility, an UL training strategy that takes into consideration the evolution of user positions, and, consequently, large-scale fading coefficients, should be developed. Doing so requires exact information about user locations, which can be quite complicated to obtain, in practice. Consequently, we tackle this problem while allowing the network to have a partial knowledge of the user positions. We suppose that the network is able to estimate the location of a limited set of users. Adapting to the change in the large-scale fading coefficients and optimizing UL training decisions based on the channel's autocorrelation should occur on two different time scales [84]. Consequently, we develop a two time scale control problem where the network learns the best position estimation and UL training decisions for long time periods.

In the fast time scale (lower level), the network derives an optimal training policy while assuming constant channel second order statistics. By taking into consideration the evolution over time of the correlation between the estimated CSI and the actual channel, the network is able to optimize its decisions to schedule users for UL training over time. Taking into consideration the time dimension allows the network to be more efficient since it becomes able to predict the impact of its training decisions on present and future performance. In the slow time scale (upper level), the network adapts to user mobility by deciding which users are required to feedback their locations. In fact estimating the exact location of all users requires a non negligible signaling overhead. Consequently, efficiently selecting the users that are required to feedback their location is of paramount importance. The combined optimization on the two time scales provides the network with an optimal training strategy that considerably improves the achievable cumulative average SE.

The rest of the chapter is organized as follows. The network model under consideration is provided in Section 4.2. In Section 4.3, we propose an adaptive Doppler based UL training framework that exploits outdated CSI estimates. We also assess its performance with different linear receivers. In Section 4.4, a two time scale UL training learning framework is presented and optimization algorithms are provided. In Section 4.5, numerical results capturing the gains that the proposed scheme can provide are given. We finally conclude in Section 4.6.

4.2 System Model And Preliminaries

We consider the UL of a multi-cell multiuser massive MIMO system constituted of C macro BSs operating in TDD mode. Each macro BS is equipped with M omnidirectional antennas and serves K mobile devices equipped, each, with a single omnidirectional antenna. We will refer to the latter as users.

All users in the network move according to different speeds and directions. Consequently, their signals are subject to heterogeneous Doppler spreads which results in different wireless channel autocorrelations in time. We consider a system where time is slotted $t \in \{0, 1, \dots\}$ and the duration of each time slot t is given by T_s OFDM symbols. We note that T_s is a system parameter that depends on the maximum Doppler spread supported by the network, see for instance Toufik et al. [1].

The wireless channel of each user can be decomposed as a product of small and large scale fading coefficients. The wireless channel from user k (in cell c) to BS j, at time slot t, i.e., $g_{kc}^{[j]}(t)$, is given by

$$g_{kc}^{[j]}(t) = \sqrt{\beta_{kc}^{[j]}} h_{kc}^{[j]}(t), \text{ for all } k = 1, \dots, K, \text{ and } j, c = 1, \dots, C, \tag{4.1}$$

where $h_{kc}^{[j]}(t) \in \mathbb{C}^{M \times 1}$ is the fast fading vector which is described by means of the uncorrelated circular-symmetric complex Gaussian channel vector having zero mean and unit variance, i.e. $h_{kc}^{[j]}(t) \sim \mathcal{CN}(0, I_M)$. $\beta_{kc}^{[j]} \in \mathbb{R}^+$ models the large-scale effect including shadowing and pathloss, which are assumed to remain constant during large-scale coherence blocks of T_β OFDM symbols. Typically ($T_\beta >> T_s$). For different large-scale coherence blocks coefficients $\beta_{kc}^{[j]}$ are assumed to be independent.

Remark 1. *In Sections IV and V we will consider a system where, after T_β OFDM symbols, $\beta_{kc}^{[j]}$ evolves according to a Markovian model.*

4.2.1 Channel Estimation

As introduced above, in this work we focus on a TDD system, where the entire frequency band is used for DL and UL transmission by all BSs and users. The BSs acquire CSI estimates using orthonormal training sequences (i.e., pilot sequences) in the UL. We consider a pilot reuse factor of 1, i.e. the same sets of pilot sequences are used in all cells.

We also consider that, during each coherence interval, a maximum of τ users are scheduled for UL training in each cell with $\tau \leq K$. For that, we consider a set of orthonormal training sequences, that is, sequences $q_i \in \mathbb{C}^{\tau \times 1}$ such that $q_i^\dagger q_j = \delta_{ij}$ (with δ_{ij} the Kronecker delta).

During the UL training phase of the coherence interval t, the l^{th} BS receives the pilot signal $Y_p^{[l]}(t) \in \mathbb{C}^{M \times \tau}$

$$Y_p^{[l]}(t) = \sum_{c=1}^{C} \sum_{k=1}^{\tau} \sqrt{P_p} g_{kc}^{[l]}(t) q_k^\dagger + W_p(t), \tag{4.2}$$

where $W_p(t) \in \mathbb{C}^{M \times \tau}$ refers to an additive white Gaussian noise matrix with i.i.d. $\mathcal{CN}(0,1)$ entries. P_p refers to the training signal power. The l^{th} BS then uses the orthogonality of training sequences in order to obtain the MMSE estimate of the channel of user k, l [129] as

$$\hat{g}_{kl}^{[l]}(t) = \frac{\beta_{kl}^{[l]}}{\frac{1}{P_p} + \sum_{b,b\neq l}^{C} \beta_{kb}^{[l]}} \frac{Y_p^{[l]}(t)}{\sqrt{P_p}} q_k. \tag{4.3}$$

Note that the MMSE channel estimate $\hat{g}_{kl}^{[l]}(t)$ follows a $\mathcal{CN}\left(0, \frac{{\beta_{kl}^{[l]}}^2}{\frac{1}{P_p} + \sum_{b,b\neq l}^{C} \beta_{kb}^{[l]}} I_M\right)$ distribution. Taking into consideration the MMSE estimation result, the wireless channel between user k (in cell l) and BS l can then be decomposed as follows

$$g_{kl}^{[l]}(t) = \hat{g}_{kl}^{[l]}(t) + \tilde{g}_{kl}^{[l]}(t), \tag{4.4}$$

where $\tilde{g}_{kl}^{[l]}(t)$ represents the estimation error and follows a $\mathcal{CN}\left(0, \left(\beta_{kl}^{[l]} - \frac{{\beta_{kl}^{[l]}}^2}{\frac{1}{P_p} + \sum_{b,b\neq l}^{C} \beta_{kb}^{[l]}}\right) I_M\right)$ distribution. Moreover, $\hat{g}_{kl}^{[l]}(t)$ and $\tilde{g}_{kl}^{[l]}(t)$ are independent due to the orthogonality property of linear MMSE estimators.

4.2.2 Channel aging

In practice, the wireless channel varies between the time when it is learned and used for precoding in DL and decoding in UL. This variation is due mainly to user movement and processing delays. Such phenomenon is referred to as *channel aging*. Its impact can be captured by a time varying wireless channel model. To this end we consider a stationary ergodic Gauss-Markov block fading regular process (or auto-regressive model of order 1) [124]. The evolution of the channel vector of user k, l between the two slots t and $t-1$ is expressed as

$$g_{kl}^{[l]}(t) = \rho_{kl}^{[l]} g_{kl}^{[l]}(t-1) + \sqrt{\beta_{kl}^{[l]}} \varepsilon_{kl}^{[l]}(t), \tag{4.5}$$

where $\varepsilon_{kl}^{[l]}(t)$ denotes a temporally uncorrelated complex white Gaussian noise process with zero mean and variance $(1 - {\rho_{kl}^{[l]}}^2) I_M$. $\rho_{kl}^{[l]}$ represents a temporal correlation parameter of the channel of user k,l. This parameter is given by Jakes et al. [124] and reads as follows

$$\rho_{kl}^{[j]} = J_0(2\pi f_{kl}^{[j]} T), \tag{4.6}$$

where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind and $f_{kl}^{[j]}$ represents the maximum Doppler shift of user k in cell l with respect to the antennas of BS j. In our work, we adopt a realistic setting in which, mobile users have different frequency shifts since we consider heterogeneous movement velocities and directions. For every user k in cell l, the maximum Doppler shift with respect to the antennas of BS j is given by

$$f_{kl}^{[j]} = \frac{\nu_{kl} f_c}{c} \cos(\theta_{kl}^{[j]}), \tag{4.7}$$

where ν_{kl} is the velocity of user k in cell l in meters per seconds, $c = 3 \times 10^8$ mps is the speed of light, f_c is the carrier frequency and $\theta_{kl}^{[j]}$ represents the angular difference between the directions of the mobile device movement and the incident wave. From the properties of the Bessel function, we deduce that the channel autocorrelation is bounded as, $0 \leq |\rho_{kl}^{[j]}| \leq 1$. Taking into consideration the combined effects of estimation error and impairments due to channel aging, we can express the wireless channel of user k,l at time t as

$$g_{kl}^{[l]}(t) = \rho_{kl}^{[l]} \hat{g}_{kl}^{[l]}(t-1) + \rho_{kl}^{[l]} \tilde{g}_{kl}^{[l]}(t-1) + \sqrt{\beta_{kl}^{[l]}} \varepsilon_{kl}^{[l]}(t), \tag{4.8}$$

4.3 An adaptive uplink training approach for Massive MIMO TDD systems

In this section, we present a novel approach for UL training in TDD Massive MIMO systems. We argue that using an adaptive training scheme that leverages the users heterogeneous channel coherence times can improve the achievable spectral efficiency. We present a detailed analysis of the impact of the proposed scheme on the achievable spectral efficiency with MRC and ZF receivers respectively.

In current Massive MIMO models, the same coherence interval T_s is considered for all users in the network. T_s is defined as a system parameter that is based on the maximum Doppler spread supported by the network [1]. This consideration results in a suboptimal use of the time-frequency resources and a loss of flexibility that can be leveraged otherwise. As a matter of fact, in practice, users have heterogeneous Doppler spreads. This is due to different users velocities and movement directions. Consequently, their channels do not age at the same rate. Considering that all users need to perform UL training with the same periodicity while sidestepping the important of heterogeneous coherence times, causes vain redundancy and a loss of resources. A more efficient approach should adapt the periodicity of each user CSI estimation according to its actual coherence time [111] [123]. This means that, at a given slot, if the correlation between the estimated CSI and the actual channel was not considerably degraded, due to aging, the network is not required to reestimate it. Doing so enables to spear part of the training resources that can be used for data transmission or to schedule more users. In all cases, the latter results in an increase in the achievable spectral efficiency.

In this section, we investigate the achievable spectral efficiency when such training scheme is used. We also derive an important condition which ensures that a coherence time-based training scheme is able to provide a substantial improvement of the network performance.

4.3.1 An adaptive coherence time-based uplink training scheme

We consider a massive MIMO system in which CSI estimation is adapted according to the actual users' coherence times. We consider that the network groups users according to their channel autocorrelation coefficients into N_g copilot user groups $\lambda_g,\ g = 1, ..., N_g$. The users in each group are either scheduled for UL training synchronously, using the same pilot sequence, or not scheduled at all. This requirement guarantees that copilot users always have the same CSI delay and a similar channel aging effect. For each copilot group $\lambda_g,\ g = 1, ..., N_g$, the CSI delays are denoted by $d_g, g = 1, ..., N_g$. At each slot, all N_g copilot user groups are scheduled for data transmission and a maximum of τ $(\tau < N_g)$ copilot groups are selected for UL training. The rest will have their signals processed using the last estimated version of their CSI. The proposed TDD protocol consists of the following seven steps.

1. In the beginning of each large-scale coherence block, the BSs estimate the large scale fading and channel autocorrelation coefficients, i.e., $\beta_{kc}^{[j]}$ and $\rho_{kc}^{[j]}$ for all $k = 1, \dots, K$, and $c, j = 1, \dots, C$. All coefficients are then fed back to a central processing unit (CPU).

2. Next, the CP uses the K-mean algorithm in order to cluster users according to their autocorrelation coefficients, see Young et al [189]. The resulting clusters will be characterized by an average autocorrelation coefficient or, equivalently, an average Doppler spread and a variance of the corresponding users autocorrelation coefficients. The considered number of clusters is N_c. Defining N_c is of paramount importance. In this work, we choose to define N_c according to

$$N_c = \lceil \frac{T_{max}}{T_s} \rceil, \tag{4.9}$$

where T_{max} represents the maximum coherence slot of the users' channels. (3.9) guarantees that the average coherence slot per cluster is approximately equivalent to a multiple of T_s. This is needed in order to appropriately define CSI estimation periodicity as a function of the parameter T_s.

3. Next, the CP allocates all users in the network (K per cell) to N_g copilot user groups. Each group contains at maximum C users from the same channel autocorrelation based cluster and from different cells. These N_g copilot groups are formed so that the variance of the autocorrelation coefficients in each group is minimized. This is done in order to guarantee that copilot users has similar channel aging effects. An example of the proposed adaptive coherence time-based user grouping procedure is given in Figure 4.1.

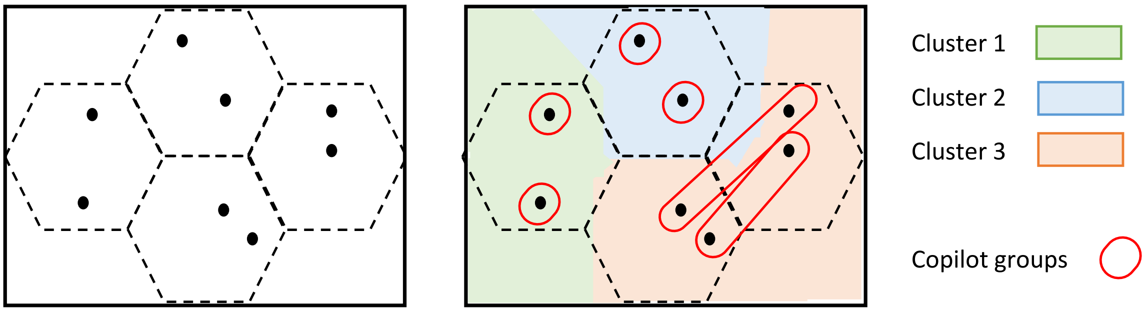

Figure 4.1: Illustration of Steps 2) and 3) of the adaptive coherence-time based user scheduling procedure with $C = 4, K = 2, N_g = 6, N_c = 3$.

4. At each coherence slot, the network schedules at maximum τ copilot user groups for UL training synchronously. Depending on the main KPI to optimize, different scheduling algorithms can be used to select these copilot groups [1].
5. All N_g copilot groups transmit their UL signal in a synchronous manner.
6. The BSs process the received pilot signal and estimates the channels of the active users during UL training using MMSE estimators. The BSs decode and precode the UL and DL data signals, respectively, using the last estimated version of each user CSI.
7. All BSs synchronously transmit DL data signals to the N_g copilot groups.

4.3.2 Spectral efficiency with outdated CSI

In what follows, we analyze the impact of the aforementioned training procedure on the achievable spectral efficiency of the network with linear receivers. In particular, two linear receivers are considered, namely, MRC and ZF receivers. We derive closed-form lower bounds of the achievable spectral efficiency with outdated CSI. Moreover, we provide a condition in order to ensure that the spectral efficiency of all users is improved when outdated CSI is used. For the sake of analytical traceability, we consider that the N_g copilot groups contain exactly C users. We, henceforth, refer to each user by its copilot group and serving BS indexes.

During UL data transmission, at time slot t, BS l receives the data signal $Y_u^{[l]}(t)$ which is given by

$$Y_u^{[l]}(t) = \sum_{c=1}^{C}\sum_{k=1}^{N_g} \sqrt{P_u} g_{kc}^{[l]}(t) S_{kc} + W_u(t), \tag{4.10}$$

where $W_u(t) \sim CN(0, I_M)$ is the additive noise, S_{kc} denotes the UL signal of the user from copilot group $k, k = 1, \ldots, N_g$ in cell $c, c = 1, \ldots, C$ and P_u denotes the reverse link transmit power.

MRC Receivers

In what follows, we derive a lower bound on the achievable spectral efficiency of aforementioned procedure with a MRC receiver. At the reception, each BS applies a MRC receiver based on the latest available version of CSI estimates. BS $l, l = 1, \ldots, C$ detects the signal of user $g, g = 1, \ldots, N_g$, within the same cell, by applying the following filter

$$u_{gl}(t) = \frac{\hat{g}_{gl}^{[l]}(t - d_g)}{\|\hat{g}_{gl}^{[l]}(t - d_g)\|}, t \geq d_g, \tag{4.11}$$

where $\hat{g}_{gl}^{[l]}(t - d_g)$ denotes the latest available channel estimate of user g in cell l. The resulting average achievable spectral efficiency in the system with MRC receivers is given in Theorem 6.

Theorem 6. *For N_g active copilot groups during UL transmission, τ of which are scheduled for UL training and using a MRC receiver $u_{gl}(t)$ that is based on the latest available CSI estimates of each user g, l, the average achievable spectral efficiency in the UL $\bar{R}_u^{MRC}$ is lower bounded by:*

$$\bar{R}_u^{MRC} \geq \sum_{l=1}^{C}\sum_{g=1}^{N_g} \left(1 - \frac{\tau}{T_s}\right) log \left(1 + \frac{(M-1){\beta_{gl}^{[l]}}^2 {\rho_{gl}^{[l]}}^{2d_g}}{(M-1) \times I_{gl}^p + I_{gl}^n}\right), \tag{4.12}$$

where $d_g, g = 1...N_g$ represents the CSI delays of users in copilot groups $g, g = 1, \ldots, N_g$. I^p_{gl} and I^n_{gl} are given by:

$$I^p_{gl} = \sum_{c \neq l}^{C} \rho^{[l]^{2d_g}}_{gc} \beta^{[l]^2}_{gc}, \tag{4.13}$$

$$I^n_{gl} = (\sum_{c=1}^{C} \sum_{k \neq g}^{N_g} \beta^{[l]}_{kc} + \sum_{c=1}^{C} (\beta^{[l]}_{gc} - \rho^{[l]^{2d_g}}_{gc} \frac{\beta^{[l]^2}_{gc}}{\frac{1}{P_p} + \sum_{b=1}^{C} \beta^{[l]}_{gb}}) + \frac{1}{P_u}) \times (\frac{1}{P_p} + \sum_{b=1}^{C} \beta^{[l]}_{gb}).$$

Proof. See B.1. □

Equation (4.12) provides further insights into the impact of channel aging on the achievable average spectral efficiency as a function of the CSI time offset. We can clearly see that the spectral efficiency decreases as a function of its CSI time offset. This is an intuitive result since the correlation between the estimated CSI and the actual channel fades over time. Equation (4.12) shows also that for a same CSI time offset, the degradation due to channel aging is higher for users with lower autocorrelation coefficients. Although outdated CSI causes an SINR degradation, the speared resources from UL training can lead to an increase in spectral efficiency.

ZF Receivers

We now consider that the BSs use ZF receivers that are based on the latest available version of CSI estimates. BS $l, l = 1, \ldots, C$ detects the signal of the user $g, g = 1, \ldots, N_g$, within the same cell, by applying the following filter:

$$U^{zf}_l(t) = (\hat{G}^{o\dagger}_l(t)\hat{G}^o_l(t))^{-1}\hat{G}^o_l(t),\ t \geq d_g, \tag{4.14}$$

where $\hat{G}^o_l(t) \in \mathbb{C}^{M \times N_g}$ is the outdated CSI matrix with $[\hat{G}^o_l(t)]_g = \hat{g}^{[l]}_{gl}(t - d_g)$ for each user g in cell l. The resulting average achievable spectral efficiency in the system with ZF receivers is given in Theorem 7.

Theorem 7. *For N_g active copilot groups during UL transmission, τ of which are scheduled for UL training and using a ZF receiver $U^{zf}_l(t)$ that is based on the latest available CSI estimates of each user, the average achievable spectral efficiency in the UL $\bar{R}^{ZF}_u$ is lower bounded by:*

$$\bar{R}^{ZF}_u \geq \sum_{l=1}^{C} \sum_{g=1}^{N_g} \left(1 - \frac{\tau}{T_s}\right) log \left(1 + \frac{(M - N_g)\beta^{[l]^2}_{gl} \rho^{[l]^{2d_g}}_{gl}}{(M - N_g) \times I^p_{gl} + I^n_{gl}}\right), \tag{4.15}$$

where $d_g, g = 1...N_g$ represents the CSI delays of users in copilot groups $g, g = 1, \ldots, N_g$. I^p_{gl} and I^n_{gl} are given by:

$$I^p_{gl} = \sum_{c\neq l}^{C} \rho_{gc}^{[l]^{2d_g}} \beta_{gc}^{[l]^2}, \tag{4.16}$$

$$I^n_{gl} = \left(\frac{1}{P_u} + \sum_{c}^{C} \sum_{k}^{N_g} \beta_{kc}^{[l]} - \rho_{kc}^{[l]^{2d_k}} \frac{\beta_{kc}^{[l]^2}}{\frac{1}{P_p} + \sum_{b,b\neq c}^{C} \beta_{kb}^{[l]}}\right) (\frac{1}{P_p} + \sum_{b=1}^{C} \beta_{gb}^{[l]}).$$

Proof. See. B.2. □

4.3.3 ASYMPTOTIC Performance

We now analyze the potential gain that the proposed training approach can provide. To do so, we compare it with a reference model that follows a classical TDD protocol in which all of N_G copilot groups are scheduled for UL training at each time slot. We consider a worst case scenario with random delays and random copilot groups allocation. In this scenario, each user experiences the lowest channel autocorrelation coefficient in comparison with its copilot users. This means that each user suffers from the heaviest channel aging impact in its copilot group.

Theorem 8. *In the asymptotic regime (M grows large), with $\bar{\rho}_g^{[min]}$ and $\bar{\rho}_g^{[max]}$ denoting, respectively, the minimum and maximum autocorrelation coefficients in copilot group g, $g = 1, ..., N_G$, the proposed training framework enables to improve the SE of each user when* (4.17) *is satisfied*

$$\left(\frac{\bar{\rho}_g^{[min]^2}}{\bar{\rho}_g^{[max]^2}}\right)^{d_g} \geq \frac{\left(1 + \mathit{SINR}_{g,l}^{[\infty]}\right)^{\frac{T_s - N_G}{T_s - \tau}} - 1}{\mathit{SINR}_{g,l}^{[\infty]}}, \tag{4.17}$$

with

$$\mathit{SINR}_{g,l}^{[\infty]} = \frac{\beta_{gl}^{[l]^2}}{\sum_{c\neq l} \beta_{gc}^{[l]^2}}, \tag{4.18}$$

Proof. See B.3. □

Condition (4.17) ensures that the SE of each user increases when outdated CSI is used. Equation (4.17) shows that the speared resources due to the reduced training overhead is a defining parameter. In fact, SE is improved as long as the SINR degradation is compensated for by the spared resources from training. It also shows the importance of the ratio between the minimum and maximum autocorrelation coefficients in a copilot group. A high ratio is required in order to achieve the needed SE gain. This requirement become tighter as the CSI time offset increases. (4.17) shows that the use of the proposed procedure can improves the achievable SE even with random delays and random pilot sequence allocation.

Remark 2. *In order to satisfy condition* (4.17)*, copilot users need to have similar autocorrelation coefficients. This explains Steps* 2) *and* 3) *in the protocol in Section III.A. In fact, clustering users based on their autocorrelation coefficients and grouping them accordingly results in copilot user groups with homogeneous channel aging within each group. This allows to tolerate higher CSI time offset.* (4.17) *also shows that the use of the aforementioned training procedure can improves the achievable SE of the network, even with random pilot allocation. Consequently, one can do better if a coherence time adaptive scheduling for UL training is implemented. More importantly, the proposed scheme shows the impact of the time dimension. This fact justifies the need for a time-aware training optimization which will be the focus of the next section.*

4.4 Optimal training strategy with outdated CSI and user mobility: a two-time scale decision process

We proved that adapting UL training periodicity to the actual channel coherence time can provide a considerable increase in network performance, even with random pilot sequence allocation. Nevertheless, higher performance gain can be obtained if more sophisticated and adapted scheduling policy is used. Developing such policy is the focus of this section.

As a matter of fact, knowing that CSI estimation periodicity should depend on the rate of channel aging, it makes sense to develop an UL training policy that takes into consideration the evolution in the difference between the estimated CSI and the actual wireless channels. In opposition to a per slot UL training optimizing, such policy enables to take into consideration the impact of past scheduling decisions on the long term performance. User mobility should also be included. In fact, channel aging results, primarily, from mobility, with velocity being a defining parameter. Consequently, developing an UL training policy that takes into consideration the evolution of large-scale fading coefficients, in addition to channel aging, is of paramount importance. Developing such strategy requires accurate estimates of user locations, which can be rather complicated to obtain, in practice. As a matter of fact, localizing all covered users requires a non negligible signaling overhead, when the localization capabilities of the network are used (observed time difference of arrival (OTDOA) [1] for example), on the one hand and large energy consumption with global positioning system (GPS) on the other hand [85]. Consequently, this problem should be addressed while assuming a partial knowledge of the user positions. Adapting to the change in user locations and optimizing UL training decisions based on the channels' autocorrelation coefficients, should occur on two different time scales [84]. In fact, the two optimizations are based on

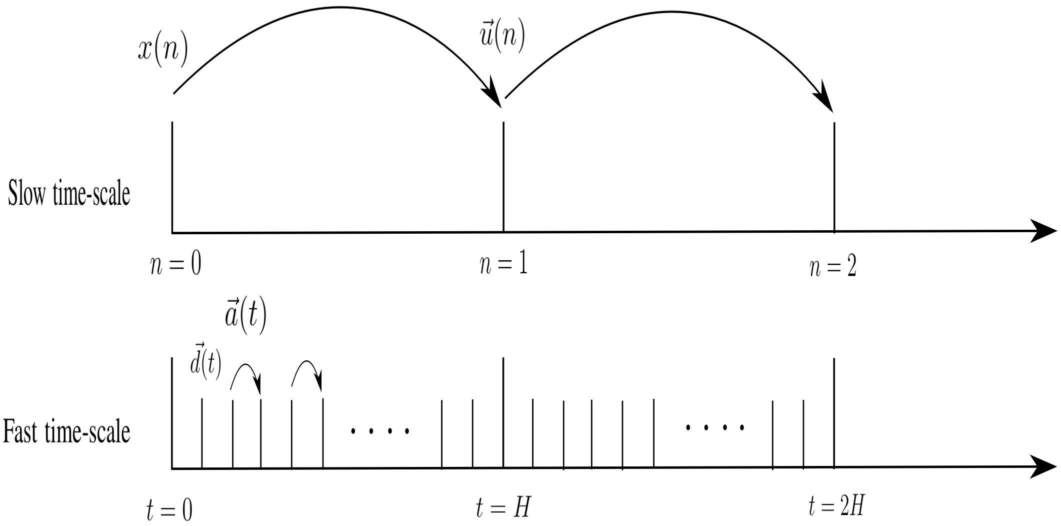

Figure 4.2: A two time-scale planning problem

information that change over heterogeneous time scales (The wireless channel changes faster than user position). Consequently, a two time scale control problem should be formulated. This will be the focus of the present section.

4.4.1 Optimizing uplink training: A two-time scale control problem

We now model the two-time scale system introduced above as a Partially Observable Markov Decision Process [87]. We assume that in both time scales (i.e., the slow and the fast-time scales, see Figure 4.2) the action and state spaces are finite.

We consider that, in the slow-time scale (*upper level*), the position of the users evolves according to a Markovian Mobility model [125] within their serving cells. These position variations occur at decision times $n = 0, 1, \ldots$. Let $\ell_g(n)$ be the combination of the positions of users from copilot group g at time n. Considering the combination of copilot users positions instead of each individual one enables to reduce the complexity of the present model. We assume, for the sake of simplicity, that all copilot groups have L possible position combinations, hence $\ell_g(n) \in \{1, \ldots, L\}$. Building this model requires a portioning of the coverage area of each cell into a number of disjoint regions. The area of each region is chosen such that the variation of the large scale fading coefficients can be considered as negligible within the region. For copilot group g, each position $\ell_g(n) \in \{1, \ldots, L\}$ corresponds to a combination of regions in each cell. The transition probabilities are characterized by the matrix

$$P_g = (p_g(i, j))_{i,j \in \{1,\ldots,L\}}, \text{ for copilot group } g. \tag{4.19}$$

The large scale fading coefficients for user g in cell l, i.e., $\beta_{gl}^{[j]}, j \in \{1, \ldots, C\}$ depend on the users' position. In previous sections, we assumed that this values were constant. In this

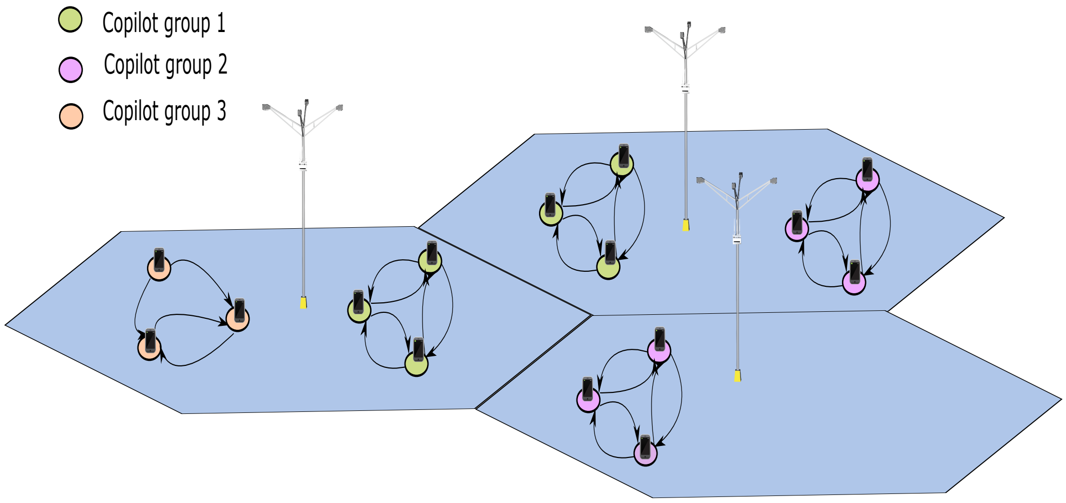

Figure 4.3: Markovian Mobility Model

section, we add a time dependency to it, namely, $\beta_{gl}^{[j]}(n) = \beta_{gl}^{[j],\ell_g(n)} \in \{\beta_{gl}^{[j],1}, \dots, \beta_{gl}^{[j],L}\}$.

Acquiring the information on the position of all users can be really expensive. In fact, in 4G networks, mobile devices were required to monitor and process positioning reference signals (PRS) from all neighboring cells which requires a non-negligible signaling and processing overhead [1]. Acquiring, user position using GPS results also in a considerable energy consumption for the mobile devices [85]. Consequently, we consider that a limited number of users can feedback its positions to the network. In particular, we assume that, in every decision epoch, the users from U_{max} copilot groups can feedback their positions (with $U_{max} < N_g$). The CP therefore can only acquire the positions of the users from U_{max} copilot groups, at each time n. The positions of the rest of the user will be inferred from previous estimations. This estimation is characterized by the belief state vector. The belief state vector of copilot group g, at decision-time n, will be denoted by $\vec{b}_g(n)$, where the i^{th} entry in $\vec{b}_g(n)$ refers to the probability that the users of copilot group g are in positions of combination i. We define by $\mathcal{X}_g$ the set of all belief states for copilot group g and we let $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_{N_g}$ be the state space in the upper level. A remark on the notation is now in order.

Remark 3. *The state in the upper level $x \in \mathcal{X}$ is an $L \times N_g$ matrix, whose columns represent the belief state vectors of all copilot groups g, for $g = 1, \dots, N_g$. That is, $x = (\vec{b}_1, \dots, \vec{b}_{N_g})$.*

In the upper level, at every decision epoch $n = 0, 1, \ldots$, the decision is to select which U_{max} copilot groups out of the N_g will transmit their positions to the BSs. That is, we consider the action vector $\vec{u}(n) = (u_1(n), \dots, u_{N_g}(n)) \in \mathcal{A} = \{0, 1\}^{N_g}$, such that

$$u_g(n) = \begin{cases} 1 & \text{users in copilot group } g \text{ feedback their positions at decision epoch } n, \\ 0 & \text{otherwise.} \end{cases} \tag{4.20}$$

At decision epoch n, the transition probability from belief state matrix $x(n) \in \mathcal{X}$ to belief state matrix $x(n+1) \in \mathcal{X}$ is defined by

$$\begin{aligned}\mathbb{P}^{up}(x(n+1) = x'|x(n) = x, \vec{u}(n)) =& \mathbb{P}(\vec{b}_1(n+1) = b'_1|\vec{b}_1(n) = b_1, \vec{u}(n)) \cdot \ldots \\ & \cdot \mathbb{P}(\vec{b}_{N_g}(n+1) = b'_{N_g}|\vec{b}_{N_g}(n) = b_{N_g}, \vec{u}(n)),\end{aligned} \tag{4.21}$$

where, $x' = (\vec{b}'_1, \ldots, \vec{b}'_{N_g})$, $x = (\vec{b}_1, \ldots, \vec{b}_{N_g})$ with $b'_g, b_g \in \mathcal{X}_g$ for all $g = 1, \ldots, N_g$ and $\vec{u}(n) \in \mathcal{A}$. The latter is satisfied because all users have independent movements. Recall that each position combination of users in copilot group g is characterized by a set of large scale fading coefficients $\beta_{gl}^{[j]}, j \in \{1, \ldots, C\},\ l \in \{1, \ldots, C\}$.

In the fast-time scale, we define the state-space by $X = \{0, \ldots, H-1\}^{N_g}$, that is, the set of all possible delay vectors. Namely, $\vec{d} = (d_1, \ldots, d_{N_g}) \in X$ is such that d_g is the CSI delay of all users in copilot group g, i.e., λ_g. The action space is given by $A = \{0, 1\}^{N_g}$. For $\vec{a} = (a_1, \ldots, a_{N_g}) \in A$, $a_g, 1, \ldots, N_g$ is given by

$$a_g = \begin{cases} 1 & \text{copilot group } g \text{ is scheduled for uplink training,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.22}$$

When a given copilot group is not scheduled for training, its signals are processed using the last available CSI estimates. The decision times at the fast-time scale (*lower level*) will be denoted by $t = \{t_0, t_1, \ldots\}$, with $t_{nH} = n$ for all $n = 0, 1, \ldots$ and H the finite-time horizon in the *lower level*. Moreover, we make the assumption that the decision $\vec{u}(n)$ in the slow-time scale (at decision epoch n) is made right after the decision at time t_{nH}. We denote by $\vec{d}(0) = \vec{d}^0 \in X$ the initial state in the fast-time scale at $n = 0$ and $x_0 \in \mathcal{X}$ the initial state in the slow-time scale. In this particular model, the fast time scale transitions from time t_{nH} until time $t_{(n+1)H-1}$ for all $n \geq 0$ are deterministic. Namely,

$$d_g(t_{nH+j}) = (1 + d_g(t_{nH+j-1}))(1 - a_g(t_{nH+j})), \text{ for all } n \geq 0, \text{ and } 1 \leq j \leq H. \tag{4.23}$$

At the fast time scale, we therefore encounter a finite-state finite-horizon deterministic sequential-decision making problem [86]. In what follows we consider MRC receivers at the BSs. The same analysis can be done in the case of ZF receivers by applying simple changes which will be indicated when needed. The reward in this lower level, at time t with MRC receivers, is the following

$$R^{low}(\vec{d}(t), \vec{a}(t), x, \vec{u}) = \sum_{g=1}^{N_g} \sum_{l=1}^{C} \left(1 - \frac{1}{T_s} \sum_{i=1}^{N_g} a_i(t)\right) \log\left(1 + \text{SINR}_{gl}^{MRC}(\vec{d}(t), x, \vec{u})\right), \tag{4.24}$$

where $x \in \mathcal{X}$ and $\vec{u} \in \mathcal{A}$ are fixed and

$$\text{SINR}_{gl}^{MRC}(\vec{d}(t), x, \vec{u}) = \frac{(M-1)(\beta_{gl}^{[l]})^2 (\rho_{gl}^{[l]})^{2d_g(t)}}{(M-1) \times I_{gl}^p + I_{gl}^n}, \tag{4.25}$$

the SINR of user g in cell l with MRC receiver. I^p_{gl} and I^n_{gl} are given in Theorem 6. Note that SINR^{MRC}_{gl} can be replaced with SINR^{ZF}_{gl} (Theorem 7), if ZF receivers are used. We also need to state that the reward function at the lower level, i.e., R^{low}, depends on the belief state and the decision in the upper level.

We now define the sequence $\pi^{low} = \{\vec{\phi}^{low}_n\}^{\infty}_{n=0}$, where for each n,

$$\vec{\phi}^{low}_n = (\phi^{low}_{t_{nH}}, \phi^{low}_{t_{nH+1}}, \ldots, \phi^{low}_{t_{(n+1)H-1}}). \tag{4.26}$$

Each function $\phi^{low}_{t_{nH+j}} : X \times \mathcal{X} \times \mathcal{A} \to A$ prescribes the action to be taken at decision time t_{nH+j} (in the lower level), for all $n \geq 0$ and all $0 \leq j \leq H-1$. For this model we only look at the set of stationary decision rules, π^{low} with respect to the upper level, such that $\vec{\phi}^{low}_n(\vec{d}, x, \vec{u}) = \vec{\phi}^{low}_{n'}(\vec{d}, x, \vec{u})$ for all n and n' given $\vec{d} \in X$, $x \in \mathcal{X}$ and $\vec{u} \in \mathcal{A}$. The set of all possible lower level decision rules will be denoted by Π^{low}, i.e., $\pi^{low} \in \Pi^{low}$. Moreover, we drop the dependency on n, since we only consider policies that are n-independent, and we denote by Φ^{low} the set of all H-horizon policies $\vec{\phi}^{low}$, i.e., $\vec{\phi}^{low} \in \Phi^{low}$. We now define $\Phi^{low}_{x,\vec{u}} \subset \Phi^{low}$ as follows

$$\Phi^{low}_{x,\vec{u}} = \{\vec{\phi}^{low}_{x,\vec{u}} : \vec{\phi}^{low}_{x,\vec{u}} = (\phi^{low}_{x,\vec{u},t_0}, \ldots, \phi^{low}_{x,\vec{u},t_{H-1}}), \phi^{low}_{x,\vec{u},t_j} : X \times \{x\} \times \{\vec{u}\} \to A \text{ and } j = 0, \ldots, H-1\}. \tag{4.27}$$

The latter is the set of all H-horizon policies given initial belief state matrix x and action in the upper level $\vec{u}$. Note that, in the definition of $\Phi^{low}_{x,\vec{u}}$ to introduce the policy $\vec{\phi}^{low}_{x,\vec{u}}$, we use the decision times $t_0, \ldots, t_{H-1}$. This is without loss of generality, since we recall that these policies are independent from n.

Next we define the reward in the upper level. Namely,

$$R^{up}(\vec{d}, \vec{\phi}^{low}, x(n), \vec{u}(n)) = \sum_{t=t_{nH}}^{t_{(n+1)H-1}} R^{low}(\vec{d}(t), \phi^{low}_t(\vec{d}(t), x(n), \vec{u}(n)), x(n), \vec{u}(n)), \tag{4.28}$$

where $\vec{d}$ is the delay state vector at time t_{nH}. We remark that none of the upper level decisions incur in an immediate cost. Let us denote by Φ^{up} the set of all possible stationary decision rules in the upper level, such that $\pi^{up} \in \Phi^{up}$, $\pi^{up} : X \times \mathcal{X} \to \mathcal{A}$. Consequently, the objective is to find $\pi^{up} \in \Phi^{up}$ and $\pi^{low} \in \Phi^{low}$ such that

$$\max_{\pi^{up} \in \Phi^{up}} \max_{\pi^{low} \in \Phi^{low}} \lim_{Z \to \infty} \frac{1}{Z} \sum_{n=0}^{Z-1} \mathbb{E}\left(R^{up}(\vec{d}(t_{nH}), \pi^{low}, x(n), \pi^{up}(\vec{d}(t_{nH}), x(n)))\right). \tag{4.29}$$

The latter problem is a POMDP [87]. To see this, it suffices to note that the slow time scale sequential decision making problem is just a POMDP with a reward that depends on the fast time scale deterministic decision making problem. Therefore the standard theory on

Bellman's optimality equations follows. The optimal decision-rule for this POMDP can be obtained as a solution of the optimality equation for $0 < \alpha < 1$

$$V(\vec{d}, x) = \max_{\vec{u} \in \mathcal{A}} \left(\max_{\vec{\phi}^{low}_{x,\vec{u}} \in \Phi^{low}} \{ R^{up}(\vec{d}, \vec{\phi}^{low}_{x,\vec{u}}, x, \vec{u}) + \alpha \sum_{y \in \mathcal{X}} \mathbb{P}^{up}(y|x, \vec{u}) V(\vec{d}^{\vec{\phi}^{low}_{x,\vec{u}}}, y) \} \right). \tag{4.30}$$

We will now make an assumption that will simplify the model significantly. Let us define $\overline{\Phi}^{low} \subset \Phi^{low}$ where

$$\begin{aligned} \overline{\Phi}^{low} &= \{ \vec{\phi}^{low} : \vec{\phi}^{low} = (\phi^{low}_{t_0}, \ldots, \phi_{t^{low}_{H-1}}), \qquad (4.31) \\ &\quad \phi^{low}_{t_j} : X \times \{x\} \times \{\vec{u}\} \to A \text{ for } j = 0, \ldots, H-1, \text{ and } \phi^{low}_{t_0} = (1, \ldots, 1) \}. \end{aligned}$$

For all $\vec{\phi}^{low} \in \overline{\Phi}^{low}$, $\vec{\phi}^{low}$ is such that, in the first stage of the H-horizon problem, all copilot groups are scheduled for UL training. This allows us to start every slow-time scale with the same delay state $\vec{d}(nH) = (0, \ldots, 0)$ for all $n = 0, 1, \ldots$. Optimality equation in Eq. (4.30) then reduces to

$$V(x) = \max_{\vec{u} \in \mathcal{A}} \left(\max_{\vec{\phi}^{low}_{x,\vec{u}} \in \overline{\Phi}^{low}} \{ R^{up}(\vec{\phi}^{low}_{x,\vec{u}}, x, \vec{u}) \} + \alpha \sum_{y \in \mathcal{X}} \mathbb{P}^{up}(y|x, \vec{u}) V(y) \right), \tag{4.32}$$

where $R^{up}(\vec{\phi}^{low}_{x,\vec{u}}, x, \vec{u}) = R^{up}((0, \ldots, 0), \vec{\phi}^{low}_{x,\vec{u}}, x, \vec{u})$. If we further denote

$$R^{max}(x, \vec{u}) = \max_{\vec{\phi}^{low}_{x,\vec{u}} \in \overline{\Phi}^{low}} \{ R^{up}(\vec{\phi}^{low}_{x,\vec{u}}, x, \vec{u}) \}, \tag{4.33}$$

we then obtain a standard one-time scale POMDP, and its optimality equation reduces to

$$V(x) = \max_{\vec{u} \in \mathcal{A}} \left(R^{max}(x, \vec{u}) + \alpha \sum_{y \in \mathcal{X}} \mathbb{P}^{up}(y|x, \vec{u}) V(y) \right). \tag{4.34}$$

Although Markov Decision Process (MDP)s have been long studied in the literature, little can be said about the optimal solution of complex problems like Problem (4.34). In the next subsections, we provide methods to solve Problem (4.34).

4.4.2 Fast time scale: learning an optimal training strategy for finite horizon

We start by solving the fast time scale problem in order to derive $R^{max}(x, \vec{u})$, see equation (4.33). In the current literature, the number of scheduled users, on each slice of the spectrum, is limited by the length of the UL training signal, which is already fixed. A more appropriate approach would be to define the needed training resources dynamically at each time slot depending on all available information about users. This is in accordance with the concept of dynamic TDD that is already considered in the current development of the 5G standard [2].

In order to do so, we consider a dynamic system in which the base stations selects, at the beginning of each slot, the users that are scheduled for training, if there is any, and the users that will be using an outdated version of their CSI. In order to enjoy full flexibility, no maximum delay constraint is fixed a priori and the network is free to allow any delay it deems acceptable. All copilot groups are scheduled for data transmission. All available CSI estimates are used in order to decode the users signal. We choose to base the proposed UL training optimization on the lower bound in Theorem 6 and on the estimated autocorrelation of the users channels. This is quite advantageous since it allows the network to optimize its training over time without requiring instantaneousness channel estimates. UL training optimization can naturally be formulated as a discrete planning problem over a finite time horizon [86]. In fact, optimizing the network's scheduling decisions is equivalent to deriving a sequence of actions that will maximize the projected cumulative average spectral efficiency over time.

We formulate a finite horizon deterministic control problem. The optimal training decisions are derived for a predefined time duration H. The actions of the network on the fast time scale are optimized while assuming a given belief state x, an initial state $\vec{d}(0) = (0, \dots, 0)$ for all $n = 0, 1, \dots$ and a given action in the upper level $\vec{u}$.

The control horizon H is selected to be equal to the large-scale fading block. The main goal, in this section, is to derive the optimal training strategy over a finite optimization horizon H. To the authors knowledge, this is the first work that addresses the issue of UL training using an optimal control approach over a finite time horizon. Without loss of generality, we consider $n = 0$. The problem of optimal users scheduling for UL training can be formulated as follows:

$$\max_{\vec{\phi}^{low}_{x,\vec{u}} \in \overline{\Phi}^{low}} \{\sum_{t=t_0}^{t_{H-1}} \sum_{g=1}^{N_g} \sum_{l=1}^{C} \left(1 - \frac{1}{T_s} \sum_{i=1}^{N_g} a_i(t)\right) \log\left(1 + \text{SINR}_{gl}^{MRC}(\vec{d}(t), x, \vec{u})\right)\}, \tag{4.35}$$

with

$$\sum_{g=1}^{N_g} a_g(t) \leq \tau, \forall t = t_1, \dots, t_{H-1},$$
$$\vec{d}(0) = (0, \dots, 0).$$

A naive approach to solve problem (4.35) is to generate all H-length sequences of actions and then select the sequence that results in the higher cumulative average SE after H slots (brute force). Clearly, this approach can be quite computationally prohibitive when the action space and the optimization horizon are large. A more appropriate approach is to use the Dynamic Programming (DP) algorithm, see [192] (based on the Bellman Equation). The DP approach can be used for sequential decision making problems like the one proposed in Eq. (4.35). Next we present the DP algorithm for Problem (4.35). Let $X \cup \{\gamma\}$ be the set of all delay states, with γ an artificially introduced final state. We assume that the reward (or the spectral efficiency) from state $\vec{d_i}$ to $\vec{d_j}$ with $\vec{d_i}, \vec{d_j} \in X$ is given by

$$r_{ij} = \sum_{g=1}^{N_g} \sum_{l=1}^{C} \left(1 - \frac{1}{T_s} \sum_{n=1}^{N_g} \vec{a}_n^{ij}\right) R_{gl}(\vec{d_j}), \tag{4.36}$$

if there exists an action vector $\vec{a}^{ij} = (a_1^{ij}, \ldots, a_{N_g}^{ij})$ that allows the transition from state $\vec{d_i}$ to $\vec{d_j}$ in one stage. Otherwise, $r_{ij} = -\infty$. We further define $r_{i\gamma} = 0$ the cost to go from state $\vec{d_i}$ to the final state γ. Let us now define $V_h(\vec{d})$ as the optimal reward to get from state $\vec{d}$ to γ in $H - h$ stages, then the optimal reward from initial state $\vec{d_0} = (0, \ldots, 0)$ to final state γ, i.e., $V_0(\vec{d_0})$, is obtained as follows. Define

$$V_h(\vec{d_i}) = \max_{\vec{d_j} \in X}\{r_{ij} + V_{h+1}(\vec{d_j})\}, \text{ for all } h = t_1, \ldots, t_{H-1}, \text{ and } V_H(\vec{d_i}) = r_{i\gamma} = 0, \text{ for all } \vec{d_i} \in X.$$

In the next table we provide the DP algorithm (proposed above) in details.

```
Input: Optimization horizon $H$, Maximum length of training sequences $\tau$, Channel
autocorrelation coefficients estimates for all users, $N_g$ copilot groups.

Initialize: $V_{H-1} = \{0\}^{N_g \times H}$ and $a^*_{H-1n} = \{0\}^{N_g \times H}$ for all $n \leq N_g$,

1. for $t_1 < h < t_{H-1}$:

2. for $\vec{d_i} \in X$ at stage $h$:

      define $r_{ij} = -\infty$ if the transition from $\vec{d_i}$ to $\vec{d_j}$ is not allowed

      define $r_{ij} = \sum_{g=1}^{N_g} \sum_{l=1}^{C} \left(1 - \frac{1}{T_s} \sum_{n=1}^{N_g} \vec{a}_n^{ij}\right) \log\left(1 + \mathrm{SINR}_{gl}^{MRC}(\vec{d_j}, x, \vec{u})\right)$

      if transition from $\vec{d_i}$ to $\vec{d_j}$ is allowed

3. $V_h(\vec{d_i}) = \max_{\vec{d_j} \in X}(r_{ij} + V_{h+1}(\vec{d_j}))$

4. $(a^*_{h1}(\vec{d_i}), \ldots, a^*_{hN_g}(\vec{d_i})) = \mathrm{argmax}_{\vec{d_j} \in X}(V_{h+1}(\vec{d_j}) + r_{ij})$

5. The optimal training strategy is retrieved as follows $\vec{a}^*_1(\vec{d^0} + 1)$ is the optimal action
vector at stage 1. $\vec{a}^*_2(\vec{d_j})$ with $d_{jn} = (2 + d^0_n)(1 - a^*_{1n})$ we retrieve the optimal
action vector at stage 2, and so on.
```

Table 4.1: **Finite horizon optimal uplink training strategy**

Due to user grouping, deriving the optimal training strategy is simplified. In fact, instead of deciding which user is scheduled for UL training with which pilot sequence at, each time slot, the network optimizes its decisions for predefined groups. Consequently, the search space is reduced and the optimal strategy can be derived faster. Enabling the network to optimize its training decisions over time has another interesting impact since it enables to reduce the needed signaling between the BS and the users. In fact, the scheduling decisions are communicated once every HT_s slots. Since users are grouped and are aware of their grouping, the BS is not required to notify each user, individually, about its training strategy. Feeding back the decisions to the users can be done on a group bases which significantly reduces the amount of required signaling in the network.

Remark 4. *We note that the algorithm in Table 4.1 can be computationally expensive for* *large optimization horizons H with a running time $\mathcal{O}(H\,|X|\,|A|)$. In the next subsection we provide an algorithm with a lower complexity that provides an approximate policy that reaches a guaranteed fraction of the optimal solution.*

4.4.3 Fast time scale: A faster approximate learning solution

Deriving a training strategy, in the lower level (fast time scale), using the aforementioned value iteration algorithm provides an optimal solution to the considered control problem (4.35). Nevertheless, in low mobility cases, the optimization horizon can be quite large due to slow changing user locations. Consequently, the search space at each iteration of the algorithm in table I becomes large resulting in a long running time that can hinder the UL training procedure. This motivates us to find an alternative approach to solve problem (4.35).

Instead of adopting a dynamic programming approach, we trait this problem by combinatorial optimization. In order to do so, expressing the CSI delays $\vec{d}(t_j) = (d_1(t_j), \ldots, d_{N_g}(t_j))$ as a function of the action vectors $\vec{a}(t) = (a_1(t), \ldots, a_{N_g}(t)),\ \forall t = t_0, \ldots, t_{j-1}$ and $\vec{d}(t) = (d_1(t), \ldots, d_{N_g}(t)),\ \forall t = t_0, \ldots, t_j$, is now in order. The delay $d_g(t_j), \forall g = 1, \ldots, N_g$, can be written as follows

$$d_g(t_j) = t_j \prod_{t=t_1}^{t_j}(1 - a_g(t)) + \sum_{t=t_1}^{t_j} t\ a_g(t_j - t) \prod_{h=t_j-t+1}^{t_j} (1 - a_g(h)). \tag{4.37}$$

Consequently, the objective function in problem (4.35) can be transformed into the following

$$\max_{\vec{a}(t_0),\ldots,\vec{a}(t_{H-1})} \{\sum_{t=t_0}^{t_{H-1}} \sum_{g=1}^{N_g} \sum_{l=1}^{C} \left(1 - \frac{1}{T_s} \sum_{i=1}^{N_g} a_i(t)\right) \log\left(1 + \text{SINR}_{gl}^{MF}(\vec{d}(t), x, \vec{u})\right)\}, \tag{4.38}$$

with

$$\sum_{g=1}^{N_g} a_g(t) \le \tau, \forall t = t_1, \ldots, t_{H-1}, \tag{4.39}$$

$$\text{SINR}_{gl}^{MRC}(\vec{d}(t_j), x, \vec{u}) = \frac{(M-1)(\beta_{gl}^{[l]})^2 (\rho_{gl}^{[l]})^{2(t_j \prod_{t=t_1}^{t_j}(1-a_g(t)) + \sum_{t=t_1}^{t_j} t\, a_g(t_j-t) \prod_{h=t_j-t+1}^{t_j}(1-a_g(h)))}}{(M-1) \times I_{gl}^{p} + I_{gl}^{n}},$$

and

$$I_{gl}^{p} = \sum_{c\neq l}^{C} \rho_{gc}^{[l]^{2(t_j \prod\limits_{t=t_1}^{t_j}(1-a_g(t)) + \sum\limits_{t=t_1}^{t_j} t\, a_g(t_j-t) \prod\limits_{h=t_j-t+1}^{t_j}(1-a_g(h)))}} \beta_{gc}^{[l]^2}, \tag{4.40}$$

$$I_{gl}^{n} = (\sum_{c=1}^{C}\sum_{k\neq g}\beta_{kc}^{[l]} + \sum_{c=1}^{C}(\beta_{gc}^{[l]} - \frac{\beta_{gc}^{[l]^2}\rho_{gc}^{[l]^{2(t_j \prod\limits_{t=t_1}^{t_j}(1-a_g(t)) + \sum\limits_{t=t_1}^{t_j} t\, a_g(t_j-t) \prod\limits_{h=t_j-t+1}^{t_j}(1-a_g(h)))}}}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]}}) + \frac{1}{P_u}) \tag{4.41}$$

$$\times (\frac{1}{P_p} + \sum_{b=1}^{C}\beta_{gb}^{[l]}).$$

In the following theorem, we show that problem (4.38) is equivalent to a maximization of a submodular function subject to matroid constraints.

Theorem 9. *Problem* (4.38) *is equivalent to maximizing a submodular set function subject to matroid constraints.*

Proof. See appendix C. □

The structure of problem (4.38) is quite convenient. In fact, even-though the objective function is not monotone, efficient approximation algorithms exist for the non-monotone submodular set function case.

In this work, we make use of the approximation algorithm proposed in [190] which provides a $\left(\frac{1}{k+2+\frac{1}{k}+\epsilon}\right)$-approximation of the optimal solution under k matroid constraints. In our case, we consider $H-1$ matroid constraints. Each one is associated with a given optimization stage $t, t = t_1, \ldots, t_{H-1}$. In fact, the action at each t_{nH} is already fixed $\vec{d}(t_{nH}) = (0, \ldots, 0)$. Consequently, the proposed algorithm in this subsection provides a $\left(\frac{1}{H+1+\frac{1}{H-1}+\epsilon}\right)$-approximation of the optimal cumulative average SE. The detailed algorithm is given in table 4.2. We define the ground set $G = \{v_{1t_1}, \ldots, v_{N_g t_1}, \ldots, v_{1t_{H-1}}, \ldots, v_{N_g t_{H-1}}\}$, where each element v_{gt} represents the scheduling of copilot group g for training at slot t. We also define the sets $\mathcal{I}_t,\ t = t_1, \ldots, t_{H-1}$. Each $\mathcal{I}_t$ contains the selected elements at stage t with $|\mathcal{I}_t| \leq \tau$.

```
1. Set $G_0 = G$:
2. for $t_1 < h < t_{H-1}$:
3. Apply Approximate local search Procedure (table III) on the ground set $G_h$ to obtain
   a solution $S_h \subset G_h$ corresponding to the problem: $\max_S (R^{up}(S, x, \vec{u}) : S \subset G_h)$
4. set $G_{h+1} = G_h \setminus S_h$
5. Return the best solution ($R^{max}(x, \vec{u}) = \max_{S_1,\ldots,S_{H-1}}(R^{up}(S_h, x, \vec{u}))$).
```

Table 4.2: **Algorithm for Approximate Finite horizon training strategy**

```
Input: Ground set $X$ of elements
1. Set $v \longleftarrow \text{argmax}_{u \in X}(f(u))$ and $S \longleftarrow \{v\}$
2. While one of the following local operations applies, update $S$ accordingly
•Delete Operation on $S$:
If $e \in S$ such that $f(S \setminus \{e\}) > (1 + \frac{\varsigma}{N_g^4}) f(S)$ then $S \longleftarrow S \setminus \{e\}$
•Exchange Operation on $S$:
If $d \in X \setminus S$ and $e_h \in S \cup \{\emptyset\}$ (for $t_1 < h < t_{H-1}$) are such that $(S \setminus \{e_h\}) \cup \{d\} \in \mathcal{I}_h$
for all $h$ and $f((S \setminus \{e_1, \ldots, e_{H-1}\}) \cup \{d\}) > (1 + \frac{\varsigma}{N_g^4}) f(S)$,
then $S \longleftarrow (S \setminus \{e_1, \ldots, e_{H-1}\}) \cup \{d\}$.
```

Table 4.3: **Approximate Local search Procedure**

4.4.4 Slow time scale: adapting to user mobility

Once the fast time scale planning problem is solved, we tackle the infinite horizon positioning problem of the slow time scale. Since we have chosen to decompose (4.29) into two levels, the combination of the policies, in the two time scales, will provide an infinite horizon policy that solves (4.29). The mobility of each copilot group g is modeled by an L-state Markov chain. The positions of users in a each copilot group g remain the same for a given period which is equal to the large scale fading coefficients coherence block and evolves according to the probability transition matrix P_g.

Solving the slow time scale control problem, directly, becomes intractable for a large number of users and possible positions, owing to the resulting complexity of belief-state monitoring [193]. Nevertheless, practical methods exist if policy optimality is abandoned for the sake of convergence speed. We adopt the approximate approach in Nourbakhsh et al. [194], which solves a POMDP by exploiting its underlying MDP. This is done by ignoring the agent's confusion (uncertainty about users locations) and assuming that it is in its most likely state (MLS). Replacing a complicated POMDP Problem by its underlying MDP enables to considerably reduce complexity since the belief space is replaced by a more practical and smaller state space.

We now discuss in more details how the upper level policy is derived. Particularly, in our case, the state of the underlying MDP, at a given decision epoch, $s \in \mathcal{S}$ is an $N_G \times 1$ vector whose elements represent the location of all copilot groups. That is, $s = (\ell_1, \ldots, \ell_{N_G})$. The most likely positions of users, for each decision epoch $n = 0, 1, \ldots$, are obtained as

$$\{\ell_1^*(n), \ldots, \ell_{N_G}^*(n)\} = \text{argmax}_{\{\ell_1(n),\ldots,\ell_{N_G}(n)\}\in\{1,\ldots,L\}^{N_G}}\Big(\prod_{g=1}^{N_G} \vec{b}_{g\ell_g}(n)\Big), \tag{4.42}$$

Recall that the belief position at decision epoch n depends on the belief state transition given in (4.21). Using (4.42), the agent's uncertainty about user locations is removed and the upper level planning problem is transformed to a more practical MDP. The resulting MDP is solved using value iteration [86]. At each iteration, the CP updates its belief-state (according to (4.21)) and assumes that the users are in their most likely positions (according to (4.42)). Then, a training policy is derived in the fast time scale, based on the assumed positions. Deriving the latter can be done using the algorithm in table I. This provides the upper level reward which is equivalent to the H-horizon lower level reward $R^{max}(x, \vec{u}) = \max_{\vec{\phi}^{low}_{x,\vec{u}} \in \overline{\Phi}^{low}} \{R^{up}(\vec{\phi}^{low}_{x,\vec{u}}, x, \vec{u})\}$. The same procedure is repeated until deriving the best position estimation decision for each most likely state. Although the derived policy provides only an approximate location estimation strategy, it enables, nevertheless, to solve a problem otherwise intractable in realistic scenarios.

4.5 Numerical Results

In this section we provide some numerical results to validate the derived analytical expressions and to demonstrate the performance of the proposed training/copilot group scheduling scheme. We also showcase the performance of the proposed UL training learning procedures. We compare the obtained results for the proposed schemes with a massive MIMO system operating according to the classical TDD protocol, considered as a reference model. We consider $C = 7$ hexagonal cells, each of which has a radius of 1.5Km. The possible positions of the mobile users are generated randomly in each cell with minimum distance of 10m to their serving BSs. The movement velocities and directions are generated randomly for all users. User speeds are drawn randomly from $[4Km/h; 80Km/h]$. This interval includes, pedestrian and public transportation speeds. The angle separating the movement direction of the mobile devices and the directions of their incident waves are drawn from $[0, 2\pi]$. The path-loss exponent is considered to be equal to 3.5. A coherence block of $T_s = 200$ is assumed with a coherence time of 1 ms. The system operates over a bandwidth of 200Mhz [94]. Once the copilot groups formed, we consider $L = 5$ possible position combinations for each group. The transition probabilities matrices $P_g, g = 1, \ldots, N_g$ are also generated randomly with $\sum_{j=1}^{L} p_g(i, j) = 1, \forall g = 1, \ldots, N_g, \forall i = 1, \ldots, L$.

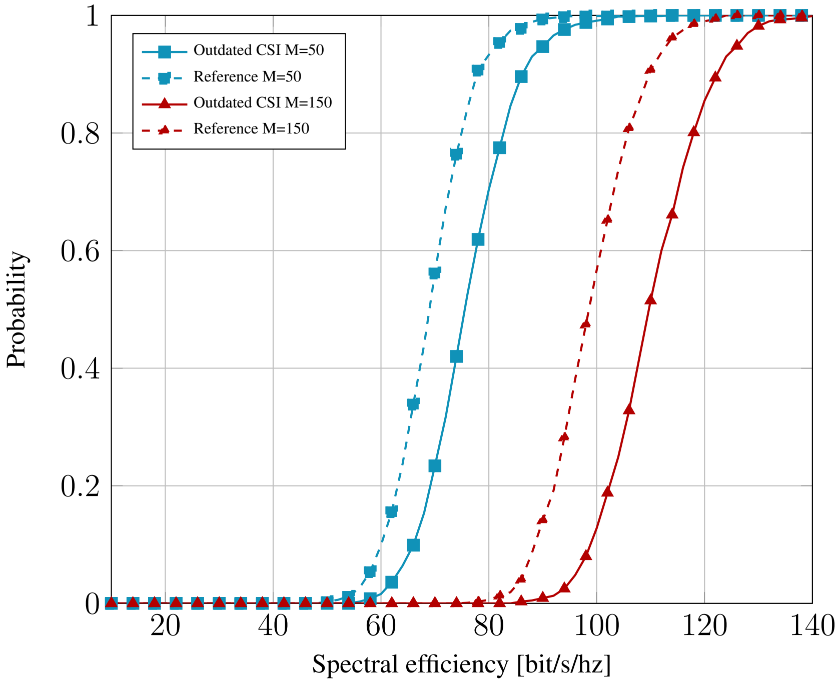

Figure 4.4: Comparison of the CDFs of spectral efficiency ($N_g = 30,\ M = 50$)

Figure 4.4 presents a comparison of CDFs of the achievable spectral efficiency between the reference model and the proposed training scheme for different numbers of antennas at the BS. For 50 BS antennas, the proposed training scheme achieves a gain in the 5% outage rate of 6 bit/s/hz. For 150 antennas, the gain in the 5%-outage rate grows to 8 bit/s/hz This increase in the performance is mainly due to the reduced training resources which can be used to transmit more data.

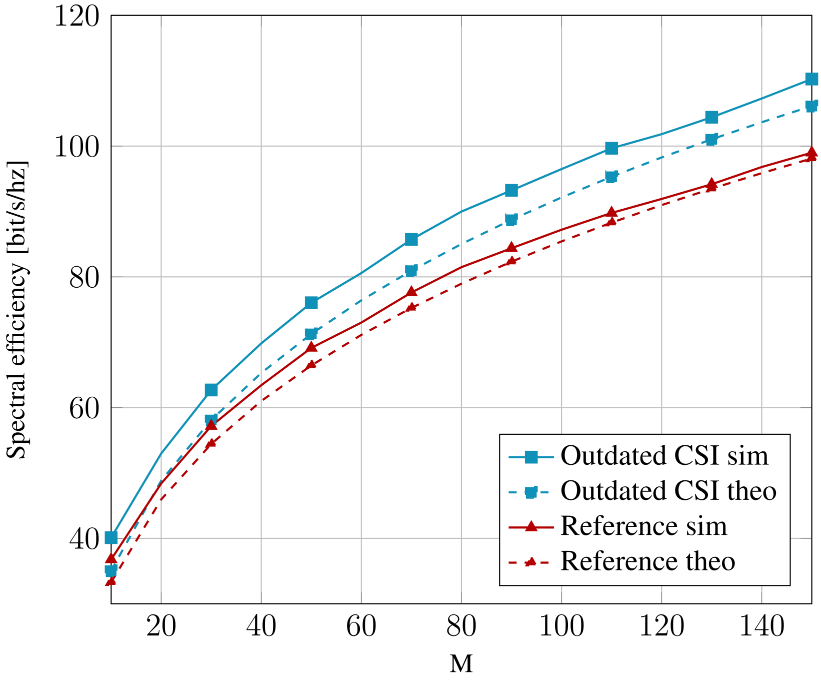

Figure 4.5: Spectral efficiency for varying values of M

Figure 4.5 examines the tightness of the proposed analytical lower bound given in Theorem 6. As can be observed, the proposed lower bound almost overlap with the simulation curve. In addition, we readily see that using outdated CSI with the implicated decrease of training resources increases the spectral efficiency by 6.91 bit/s/hz for $M = 50$. This gain attains 11.2 bit/s/hz for $M = 150$.

Figure 4.6 shows a comparison of the achievable SE as a function of the number of BS antennas with ZF and MRC receivers, respectively. As we can see in Figure 4.6, the proposed training scheme achieves a considerable gain in SE with both ZF and MRC receivers. For $M = 105$, the proposed training scheme achieves SE gains of 12.2 bit/s/hz and 3.6 bit/s/hz with ZF and MRC receivers, respectively. We readily see that the speared resources from UL training enable to mitigate any degradation in SINR that the use of outdated CSI may generate.

We now investigate the performance of the proposed two time scale learning algorithms. The performance is evaluated as the difference between the achievable cumulative average SE of the considered methods and a classical Massive MIMO TDD protocol.

In Figure 4.7 , we illustrate the performance of the UL training learning algorithms for the fast time scale. The performance of optimal policy (Algorithm table 4.1) and the approximate

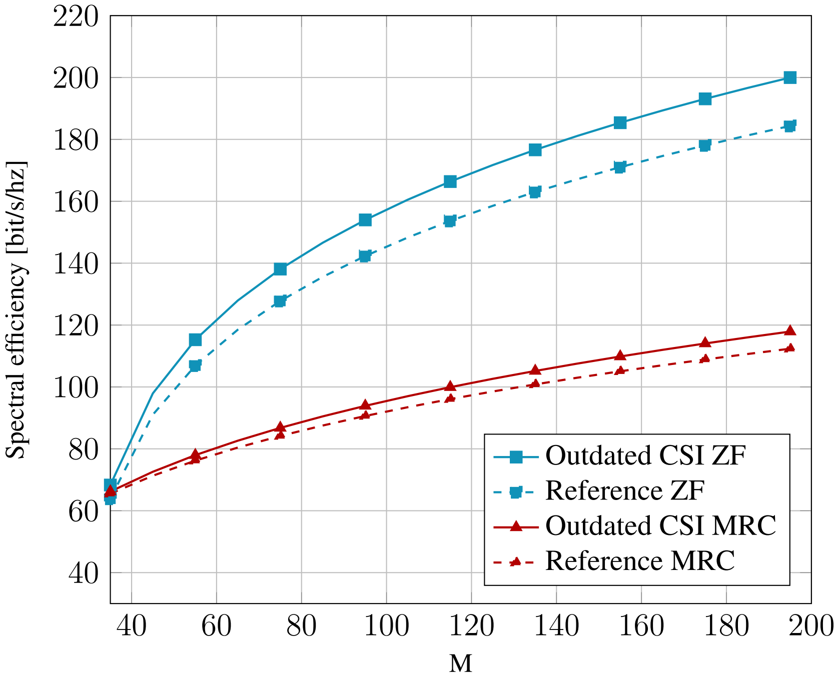

Figure 4.6: Comparison of Spectral efficiency for varying values of M with ZF and MRC receivers, respectively

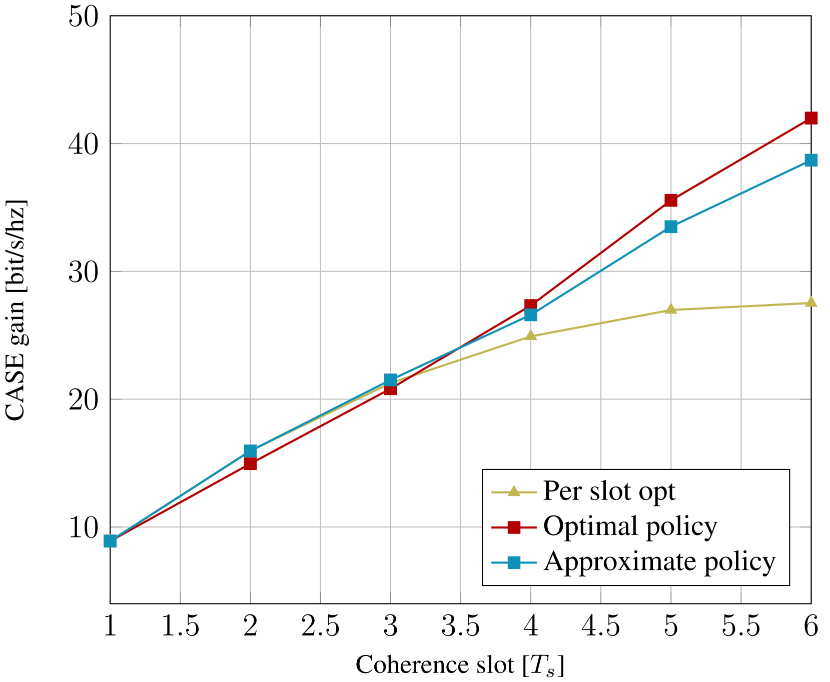

Figure 4.7: CASE gain for different lower level algorithms

one (Algorithm table 4.2) are compared with the case where outdated CSI is used with a per slot optimization. The latter means that the evolution of the correlation between the estimated CSI and the actual channel according to the time dimension is not taken into consideration and the scheduling of copilot groups for UL training is optimized in order to maximize the ASE at each slot. Figure 4.7 shows that using the algorithms in table 4.1 & 4.2, the gain in cumulative average SE is maintained and attains 41.99 bit/s/hz and 38.7 bit/s/hz respectively, at the final stage of the optimization horizon H. However, although per slot optimization achieves also a gain in cumulative average SE, we can see that this method performs poorly in comparison with the proposed policies which shows the paramount importance of taking the time dimension into consideration when optimizing UL training decisions. Finally, due to its good performance, we can deduce that the approximate method (Algorithm table 4.2) represents an efficient low complexity substitute to the more computationally prohibitive DP approach (Algorithm table 4.1).

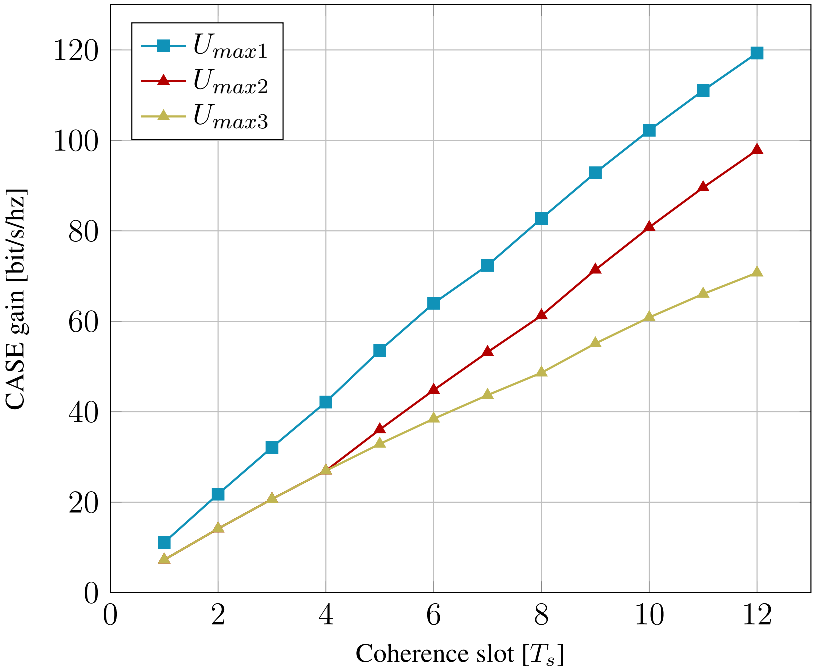

Figure 4.8: CASE gain for different U_{max} values ($H = 4$, $n = 0, \dots, 2$, $U_{max1} = N_g$, $U_{max2} = N_g - 4$ and $U_{max3} = N_g - 8$

In Figure 4.8, we illustrate the achievable cumulative average SE gain after 3 upper level decision epochs with a lower level of horizon $H = 4$. In this example, 3 values for U_{max} were considered. As can be readily observed, decreasing U_{max} results in lower cumulative average SE gain. This is quite intuitive since a lower U_{max} results in more confusion about the users locations. In fact, the CPU commits more errors when inferring users positions from its belief states for lower U_{max} values. Nevertheless, despite the positioning errors the proposed two time scale learning approach is able to provide a considerable cumulative average SE gain of 110.26 bit/s/hz, 97.863 bit/s/hz and 70.73 bit/s/hz with U_{max1}, U_{max2} and U_{max3}

respectively, after $12\,T_s$ slots. These results did not showcase the energy and signaling gains that result from reducing positioning estimation but are sufficient to prove the advantages of allowing the network to proactively plan its UL training decisions for long time periods.

4.6 Closing Remarks

In this chapter, we have proposed a novel UL training paradigm for TDD massive MIMO systems. The main idea is to adapt the periodicity of each users' CSI estimation based on its actual coherence time. The proposed scheme highlighted the importance of the time dimension in the training process and proved that this dimension can be leveraged in order to substantially improve the achievable cumulative spectral efficiency. As a matter of fact, we proposed a two time scale control problem in order to allow the network to learn the best UL training strategy taking into consideration user mobility, channel coherence time and practical signaling overhead limitations. In the fast time scale, the network learns an optimal training strategy by choosing which users are requested to send their pilot signals for a predefined optimization horizon that is equivalent to the large scale fading coefficients coherence block. In the slow time scale, owing to practical signaling and processing overhead limitation, the network needs to choose which users are required to feedback their positions, based on their belief states. The proposed approach enables to leverage the time evolution of the correlation between the wireless channel and the estimated CSI and provides an impressive increase in the achievable cumulative average SE that cannot be obtained otherwise.

Chapter 5

User-centric 5G networks: Energy Efficiency under popularity based Clustering in cache enabled SCN

5.1 Overview

5G systems are expected to fulfill multiple user experience requirements such as low latency, high reliability, low energy consumption and dense connectivity. In order to meet these requirements, network designers can call upon a number of innovative technologies, namely, dense small-cell deployment, millimeter-wave communications and massive MIMO among others. Nevertheless, the performance gap between 5G requirements and previous wireless networks, such as LTE, is considerable. Addressing this difference cannot be fulfilled by only concentrating on improving the physical layer in a user and service agnostic fashion. As a matter of fact, under the current reactive networking paradigm, physical layer technologies falls short of solving peak traffic demands. This situation is expected to worsen with the surge in the number of connected devices and the advent of high density connectivity which put a considerable strain on back-haul links. Consequently, to meet the low latency and diverse 5G requirements, some data plane functionalities should be offloaded on the edge of the network in order to facilitate data provision and reduce the strain on the back-haul. This is exactly what user-centric networks are envisioned to achieve. User-centric 5G network trades the conventional service agnostic for a more intelligent and proactive one. By monitoring and analyzing the sources of traffic, user behavior can be predicted and local copies of popular content can be provided from local storage devices at the edge of the network. This can reduce E2E latency and enhance user experience while reducing the needed resources to achieve it. User-centric networks are emerging as a promising technology to address the required performance increase for future wireless access networks. Predicting users behavior and proactively caching popular content in the edge of the network has shown a considerable potential in terms of back-haul savings and user experience improvement. In addition, a worldwide challenge for the design of future cellular systems is to meet the increasing traffic demand, while, on the other hand, to lower the emission of greenhouse gases for achieving the environment sustainability. User-centric 5G networks have the potential to address one of the key requirements of 5G networks, namely, EE. As a matter of fact, thanks to the improvement of memory devices, proactive caching offers a very practical and energy efficient alternative to network densification since it replaces back-haul link by caching capacity at the BSs. The impact of user-centric networks on EE is a fundamental subject that is attracting increasing attention [170, 171]. Previous works already showed the considerable potential of proactive caching in reducing energy consumption [76]. When coupled with energy harvesting communications, the green impact of proactive caching is taken to another level[170].

A key component of information-centric networks is users behavior modeling (i.e. content popularity profiles). Most previous works assume similar popularity profiles for all users. We argue that this approach is suboptimal. As a matter of fact, assuming homogeneous content popularity among users can only result in loosing valuable information that can be leveraged otherwise. In realistic scenarios we observe the existence of very diverse traffic patterns among users. In fact, the requested content depends on the user social network and interests that can be very different from one person to the other. In this chapter, we address this issue by providing a more adaptive framework for proactive caching that takes into consideration the diversity of user traffic patterns. We also EE in the context of heterogeneous traffic patterns.

The main contribution of this chapter is to provide a content popularity based clustering framework for caching. In particular, given heterogeneous user profiles, we propose a content popularity based clustering scheme. In order to achieve an efficient user grouping, we use the Akaike Information Criterion. This allows to effectively estimate the number of clusters and the associated average popularity profiles, thus providing an estimate of the main users request patterns. Based on the derived average profiles, we develop two optimization frameworks in order to improve the achievable EE. First, the optimal active SBS density vector is obtained through quasi-concave optimization. This allows to switch SBSs on and off and cache content accordingly so that the EE is optimized. Second, we leverage any spatial correlation in user request patterns in order to improve EE by optimizing content placement in the SCN. We develop a combinatorial optimization framework that enables to place popular content in SBSs caches so that the total transmit power is minimized.

The rest of chapter is organized as follows. Our network model for proactive caching is detailed in Section 5.2. The proposed content popularity based clustering is presented in Section 5.3. An analysis of EE in addition to its optimization with respect to active SBSs densities is carried out in Section 5.4. In Section 5.5, the problem of cache placement is addressed. Discussions of numerical results are carried out in Section5.6. Finally, Section 5.7 draws some conclusions.

5.2 System Model And Preliminaries

5.2.1 Network Model

We consider a small cell network deployed over a disc with radius R_n. The SBSs are spatially distributed according to a homogeneous Poisson point process ϕ_s with density $\lambda_{s_{max}}$. The available SBSs can be in idle or active modes. The density of active SBSs is given by λ_s such that $\lambda_s \leq \lambda_{s_{max}}$.

We consider an orthogonal frequency-division multiple access (OFDMA) system where, users served by the same SBS, are scheduled on orthogonal resources. Consequently, each user will be subject to interference coming from users served by other SBSs. The users are also distributed in $\mathbb{R}^2$ according to an independent homogeneous Poisson point process (PPP) ϕ with density λ such that, $\lambda >> \lambda_{s_{max}}$. The average number of users in the network is then given by $U = \lambda \pi R_n^2$. Each user is equipped with a single antenna and is allowed to communicate with any SBS within a radius R. This restriction enables to control the level of interference. We consider that R is defined so that each user is covered with high probability by more than one SBS. We consider that a packet can be successfully transmitted and decoded if and only if SINR $> \theta$. This means that, if the SINR is lower than the threshold θ, the link undergoes an outage and the transmission fails.

A general power law pathloss model is used where, the power decay is given by $r_{us}^{-\alpha}$. r_{us} represents the distance between user u and its serving SBS s and $\alpha > 2$ denotes the pathloss exponent.The wireless channel from user u to the SBS s is then given by:

$$g_{us} = \sqrt{r_{us}^{-\alpha}} h_{us}, \tag{5.1}$$

where h_{us} represents the small scale fading coefficient modeled as Rayleigh fading i.e., $CN(0, 1)$ distributed random variable. We consider that the transmit power, used in both UL and DL, is defined according to channel inversion power control [177]. This is done so that the transmit power compensates the pathloss in order to keep the average signal power at the receiver (i.e., the SBS or the user terminal) equal to a certain constant value ρ_0. The transmit power used by user u to communicate with SBS s, according to channel inversion power control, is given by: $\rho_{us} = \rho_0 r_{us}^{\alpha}$. The channel inversion power control will ensure a limitation of the interference level since the power received at any BS from a typical user is upper bounded by $\rho_0 R^{\alpha}$, where R denotes the maximum communication radius. Controlling the level of interference, in both UL and DL, is a vital factor that guarantees an EE gain [76].

5.2.2 User scheduling and caching strategy

We consider a file catalog C containing F files with different sizes. Each file i has a size of L_i bits. Users are assumed to have heterogeneous file popularity distributions. Each user u is associated with a popularity vector $P_u = [p_{1u}...p_{Fu}]$, where p_{iu} denotes the probability that

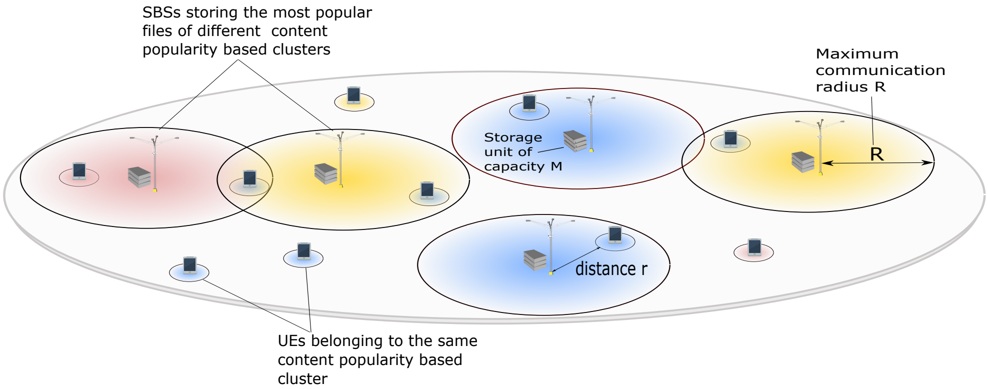
SBSs storing the most popular files of different content popularity based clusters
Maximum communication radius R
Storage unit of capacity M
R
distance r
UEs belonging to the same content popularity based cluster

Figure 5.1: System Model

user u requests file i from the catalog. We consider that these probabilities change slowly over time and that they are previously known by the network. Estimating the popularity distributions can be performed by learning from previously recorded requests [178]. In this work, we limit our analysis to the case of perfectly known popularity distributions. The study of the impact of estimation error in popularity distributions is considered in future work. Although users have heterogeneous popularity profiles, we assume that they can be grouped according to their interest into N_c clusters. This means that the users, forming each cluster, have correlated request patterns. Meaning that the distance between their content popularity vectors is small. Each SBS is equipped with a caching capacity of M bits.

Each individual SBS fills its memory device with the most popular files from a given cluster. In each cluster, the most popular files are selected based on the average of the popularity vectors associated with the users forming this cluster. Therefore, appropriately clustering the users based on the similarity in their popularity distributions is of paramount importance. Not all SBSs are required to be active. We consider the density vector of active SBSs $\Lambda_s \in \mathbb{R}^{N_c \times 1}$, where each of its coefficients $\lambda_{sk},\ k = 1..N_c$ represents the density of active SBS caching the most popular files of cluster k. Λ_s is defined such that $\Lambda_s^{\dagger}\mathbf{1} = \lambda_s$. Each user looks for the requested file in the cache of the SBSs within a radius R. The user starts with the closest SBS from his own cluster. If the requested file is available in a cache within this distance, a cache hit event occurs and the user will associate with the closest SBS storing the requested file. In the event of a cache miss, the user, simply, associates with the nearest SBS from its corresponding cluster and the requested content will be retrieved from the core network through the backhaul. If a user cannot find an SBS from its own cluster within a radius of R, it will only communicate with SBSs from other clusters within radius R, in the case of a cache hit event. An example of the considered model is represented in Figure 1 with three popularity based clusters represented, each, by a color.

5.3 Information theoretic approach to user clustering

Proactive caching systems require an efficient characterization of content popularity in order to correctly predict the files that are most likely to be requested. While most of the existing work assume similar popularity distributions for all users, we adopt a content based clustering approach. We argue that clustering users according to their popularity distributions enables to better assess social similarities [49, 51] and, consequently, devise a more efficient caching system. In fact, supposing that all users in the network have similar popularity distributions means that the resulting statistics are just an average of content popularity over all social groups. This leads to neglecting the diversity of social behavior. Contrary to traditional location based clustering methods, content aware clustering enables to identify the main request patterns in the network which leads to a better understanding of user preferences. In this work, content popularity based clustering is considered. Users are grouped such that, the correlation between the popularity profiles of users in the same cluster is maximized. This correlation is characterized by the euclidean distance between their content popularity vectors. While content based clustering was proposed in [51] using a spectral approach, we choose to adopt an information theoretic method, namely, Akaike information criterion. AIC allows to efficiently estimate the number of clusters and to assess the information loss that results from assuming a single popularity distribution.

5.3.1 Cluster estimation: Akaike information criterion

In the considered setting, the users have heterogeneous popularity profiles. However, the different social relations and interactions may result in some correlation in user request patterns. Consequently, content popularity based clustering is used in order to minimize the divergence among user content popularity distributions in each cluster. The number of content popularity based clusters is unknown a priori and should be estimated. Allowing the system to estimate this parameter periodically or whenever a substantial change in user interest is recorded, allows the network to cope with any modification in user request pattern.

In order to estimate the number of clusters, we use AIC [173] as a statistical model selection criterion. It allows to assess the quality of a statistical models for a given set of data. The data set to be modeled in our case is the collection of user content popularity distributions. Using AIC enables to estimate the expected Kullback-Leibler discrepancy between the data generating model and any candidate statistical model. It also addresses the tradeoff between the fitness of the statistical model, based on maximum likelihood estimation, and its complexity, which is given by the number of model-characterizing parameters to be estimated.

In our case, we aim at modeling the distribution that generates the user's popularity vectors. We consider the true generating distribution $A(P_1, ..., P_U) = \prod_{u=1}^{U} \mathbb{P}_u(P_u)$, where $\mathbb{P}_u(P_u)$ is the probability that user u has a popularity vector P_u. We assume that $A(P_1, ..., P_U)$

results from the aggregating of N_c user clusters where, N_c denotes the true number of clusters. Let $\xi_{N_i}, i \in [c_{min}, .., c_{max}]$ be a set of approximation models. Each approximation model ξ_{N_i} is characterized by N_i clusters and a popularity generating distribution $\prod_{u=1}^{U} \mathbb{P}_u(P_u|N_i)$. The popularity generating distribution depends on the number of clusters N_i. In fact, the average and variance of content popularity vectors in each cluster depend, primarily, on N_i. The Kullback-Leibler information, which characterizes the information lost when an approximating model is used, can be written, $\forall\ \xi_{N_i}, i \in [c_{min}, .., c_{max}]$, as:

$$d\big(N_i, N_c\big) = \int_{[0,1]^F} \prod_{u=1}^{U} \mathbb{P}_u(P_u) \log\Big(\frac{\prod_{u=1}^{U} \mathbb{P}_u(P_u)}{\prod_{u=1}^{U} \mathbb{P}_u(P_u|N_i)}\Big) \mathrm{d}\, p_1 ... \mathrm{d}\, p_F, \tag{5.2}$$

After simplification, the discrepancy between the two models is given by [174]:

$$d\big(N_i, N_c\big) = \mathbb{E}\left\{-2 \log\left(L_{\xi_{N_i}}(N_i|P_u, u = 1...U)\right)\right\}, \tag{5.3}$$

where $\mathbb{E}\{.\}$ denotes the expectation with respect to the available data, which is the collection of all users popularity profiles, knowing N_i. $L_{\xi_{N_i}}(N_i|P_u, u = 1...U)$ denotes the likelihood of having N_i clusters, knowing the popularity profiles of the users (the expression will be given later on in this section).

In [174], Akaike noted that $-2 \log\left(L_{\xi_{N_i}}(N_i|P_u, u = 1...U)\right)$ is a biased estimate of the average discrepancy. After bias adjustment, the expected discrepancy can be approximated by:

$$\mathbb{E}\left\{d\big(N_i, N_c\big)\right\} \approx 2k_i - 2\mathrm{log}\left(L_{\xi_{N_i}}(N_i|P_u, u = 1...U)\right). \tag{5.4}$$

Here k_i denotes the number of characterizing parameters in model ξ_{N_i} and $\mathbb{E}\{.\}$ denotes the expectation with respect to the available data. The expected value of the discrepancy is asymptotically equal to the expected AIC of the considered statistical model which is given by:

$$\mathrm{AIC}(\xi_{N_i}) = 2k_i - 2\mathrm{log}\left(L_{\xi_{N_i}}(N_i|P_u, u = 1...U)\right). \tag{5.5}$$

AIC allows to assess the truthfulness of any considered statistical model, and in our case, allows to estimate the number of content based clusters together with the characterizing parameters of each one. Each cluster is characterized by the average file popularity distribution and its variance within the cluster. In order to approximate the process generating the users probability vectors, we consider a set of statistical models $\Xi = \{\xi_{N_{cmin}} ... \xi_{N_{cmax}}\}$ where, $\{N_{cmin} ... N_{cmax}\}$ represents the range over which the search for the true number of clusters will be carried out. Each of the considered models will be typified by a number of defining parameters. In our case, each considered model ξ_{N_i} is characterized by $N_i \times (F + 1)$ parameters, $N_i \times F$ representing the average file popularity in each cluster and N_i variance estimates. We consider that the likelihood $L_{\xi_{N_i}}(N_i|P_u, u = 1...U)$ is computed based on a Gaussian Mixture model. This is a common assumptions for data generating models [188].

The log likelihood function log $(L_{\xi_{N_i}}(N_i|P_u, u = 1...U))$ is computed after clustering user with the assumption that they can be grouped into N_i clusters. log $(L_{\xi_{N_i}}(N_i|P_u, u = 1...U))$ can be written as follows:

$$\text{log } (L_{\xi_{N_i}}(N_i|P_u, u = 1...U)) = \sum_{u=1}^{U} \Big(\text{log}(\frac{1}{\sqrt{2\pi}\hat{\sigma}^F_{\psi(u)}}) - \frac{\|P_u - \hat{P}_{\psi(u)}\|^2}{2\hat{\sigma}^2_{\psi(u)}} + \text{log}(\frac{U_{\psi(u)}}{U})\Big), \tag{5.6}$$

where $\psi(u)$ represents the index of the cluster to which user u is assigned. $\hat{P}_{\psi(u)}$ denotes the average popularity vector in cluster $\psi(u)$. $U_{\psi(u)}$ refers to the number of users in cluster $\psi(u)$. $\hat{\sigma}^2_{\psi(u)}$ denotes the variance of content popularity vectors in cluster $\psi(u)$ and is given by:

$$\hat{\sigma}^2_{\psi(u)} = \frac{1}{(U_{\psi(u)})} \sum_{j \in U_{\psi(u)}} \|P_j - \hat{P}_{\psi(u)}\|^2. \tag{5.7}$$

Then the log-likelihood function can be written as:

$$\text{log } (L_{\xi_{N_i}}(N_i|P_u, u = 1...U)) = \sum_{k=1}^{N_i} -\frac{U_k}{2}(\text{log}(2\pi) - 1 + 2\text{log}(\frac{U_k}{U}) - F\text{log}(\hat{\sigma}^2_k)). \tag{5.8}$$

The resulting AIC for model ξ_{N_i} is given by:

$$\text{AIC}(\xi_{N_i}) = 2N_i(F + 1) + \sum_{k=1}^{N_i} U_k(\text{log}(2\pi) - 1 + 2\text{log}(\frac{U_k}{U}) - F\text{log}(\hat{\sigma}^2_k)). \tag{5.9}$$

The model that best describe the user popularity vectors is the one that minimizes the AIC and, consequently, the discrepancy. In order to find the best model, the user are clustered according to their content popularity vectors using the K-mean algorithm [189], while assuming different numbers of clusters from a search range $\{N_{cmin}...N_{cmax}\}$. The selected model ξ_{AIC} verifies:

$$\xi_{\text{AIC}} = \underset{\xi \in \Xi}{\text{argmin}} \text{ AIC}(\xi). \tag{5.10}$$

The selected model which minimizes the AIC, strikes the best trade-off between fitness and complexity. This results in a truthful modeling of content popularity based clusters. The resulting model guarantees minimum discrepancy among the request patterns of the users within each cluster. We now provide the detailed description of the content based user clustering algorithm.

5.3.2 User clustering algorithm

The proposed content popularity based clustering algorithm starts by defining a search interval $[N_{cmin}...N_{cmax}]$. The algorithm begins by assuming the existence of N_{cmin} clusters. It clusters the users accordingly using the K-mean algorithm [189]. K-mean allows to assign each user to the cluster with the nearest centroid which results in minimizing the disparity between users behaviors in the same cluster. The popularity profile of the cluster is then defined as the average of the popularity vectors of all users in the cluster as:

$$\hat{P}_k = \frac{\sum_{u,\psi(u)=k} P_u}{U_k}. \tag{5.11}$$

Each cluster k is then associated with a vector $\hat{P}_k = [\hat{p}_{1k}...\hat{p}_{Fk}]$, where $\hat{p}_{fk}$ denotes the average popularity of file f in cluster k. Once users are assigned to their respective clusters, AIC is computed. The number of clusters is incremented by adding a new centroid. The AIC is then recomputed until reaching a minimum. AIC is decreasing as a function of the number of clusters until reaching a minimum in the most accurate estimate. The AIC will then start increasing because of model complexity. Since the goal of the clustering is to reduce the divergence among users from the same cluster, a new centroids is added, at each step, in the cluster with the greatest popularity variance. The new center is selected as the user having the largest distance from the mean popularity vector of its cluster. This allows to reduce the discrepancy in user traffic pattern. The detailed clustering algorithm can be written as the following:

```
Initialize: Define search interval $[N_{cmin}...N_{cmax}]$, Set
$K = N_{cmin}$ Choose randomly the first $N_{cmin}$ centroids
from the available users
1. Run $K$-mean algorithm and compute $\mathrm{AIC}(\xi_K)$
2. Choose the user having the largest distance from its
   centroid in cluster $k^*$ with the greatest variance
   $(k^* = \underset{k=1...K}{\text{argmax}}\ \hat{\sigma}_k^2)$
3. Add a centroid with the popularity profile of
   the chosen user and set $K = K + 1$
4. Run step 1 to step 3 until AIC starts to increase.
5. Choose the model which minimizes the AIC and
   cluster the users accordingly
```

Table 5.1: **Content-popularity based user clustering algorithm**

Once content popularity based clustering is performed, the cached files of each cluster are selected based on its average popularity vector, which is given in (11). For each $\{k = 1...N_c\}$, the files in the catalog are ordered in a decreasing order of popularity according to $\hat{P}_k$. The set of cached files, in each cluster, $\{\Delta_k, k = 1...N_c\}$ is then selected as the most popular files, according to $\{\hat{P}_k, k = 1...N_c\}$, whose aggregate size is at maximum M. Apart from the maximum size constraint, we impose no restrictions on the set of cached files. Consequently, there may be some overlapping between the cached files of different clusters. Meaning that the same file can be selected in the cached sets of different clusters $(\Delta_k \cap \Delta_j \neq \emptyset, \text{for some } k \neq j)$. Files that are selected by different clusters are very popular across user. Consequently, it makes sense to increase the number of cached copies in the network. Given the considered model in section II, allowing overlapping between the cached files of different clusters provides better performance. Since user preference can change over time, the algorithm in Table I can be executed periodically or whenever substantial popularity profile modification is recorded. This allows the caching system to adapt the selected files accordingly. In order to investigate the performance of the proposed scheme, we assess its impact on the achievable EE of the system.

5.4 EE with content popularity clustering

Predicting which content is most likely to be requested and caching it in the edge can reduce the latency and backhaul load as well as increasing the overall throughput. Proactive caching also proved to be an effective technology that can improve another very important metric in future generation networks, namely, EE [76]. In what follows, we investigate the EE of cache enabled small cell networks with content popularity based user clustering. We consider the DL of the cache enabled network. Without loss of generality, we concentrate on a reference user located at the origin of the plane. The EE of the network is given by the ratio between the average achievable spectral efficiency and the average consumed power [76].

$$\Sigma = \frac{SE}{\rho^{c}_{total}}, \tag{5.12}$$

where Σ denotes the average energy efficiency, ρ^{c}_{total} denotes the average consumed power in the cache enabled small cell network and SE denotes its average achievable spectral efficiency. In order to derive the expressions of SE and ρ^{c}_{total} and, consequently, the achievable EE, we need to start by finding the expression of the cache hit probability.

5.4.1 Cache hit Probability

According to the considered system model, the cache hit probability refers to the probability of finding a requested file in the cache of a SBS within radius R from a given user [176]. Our context is different from the one in [176], since the users are clustered and the SBSs cache different files depending on their associated cluster. Considering the proposed clustering model, the cache hit probability can be expressed as follows (the derivations are skipped for brevity):

$$\mathbb{P}\{hit\} = \frac{1}{U}\sum_{k=1}^{N_c}\sum_{u=1}^{U}\big(\sum_{i\in\Delta_k} p_{iu}\big)\big(1 - e^{-\lambda_{sk}\pi R^2}\big), \tag{5.13}$$

where Δ_k represents the set of the most popular files of cluster k that fills the SBS caching capacity. This equation denotes the probability of finding of at least one SBS with the requested file stored in its cache within a radius R from a given user. The density of SBS caching the most popular content of a cluster $\{k, k = 1...N_c\}$ is given by λ_{sk} and, their average number N_{sk} is given by $N_{sk} = \lambda_{sk}\pi R_n^2$. The densities $\{\lambda_{sk}, k = 1...N_c\}$ are such that

$\sum_{k=1}^{N_c} \lambda_{sk} = \lambda_s$. One major upside of content popularity based user clustering is content diversity. While each SBS caches the most popular files of only one cluster, users can request any of the cached files in SBSs within radius R, which can be fetched without additional load on the backhaul. In fact, a given user can communicate with the closest SBS caching the files of a cluster different from his own whenever the requested content is already cached. Consequently, compared with the classical approach of caching the same popular content everywhere, the users covered by several SBSs from different clusters will see an increase in their cache hit probability.

5.4.2 Average total consumed power

In order to gain a useful insight into the achievable EE and capture the fundamental tradeoffs, we extend the power model in [76]. The average consumed total power in the considered network with caching capabilities can be modeled as follow:

$$\rho_{total}^{c} = \mathbb{E}\left\{\rho_I\right\} + \mathbb{E}\left\{\rho_T\right\} + \mathbb{E}\left\{\rho_f\right\}, \tag{5.14}$$

where ρ_I, ρ_T and ρ_f denote, respectively, the power consumed by the infrastructure of active base stations, the total transmit power and the used power to fetch files from the hard disc or the core network. The expectation $\mathbb{E}\left\{.\right\}$ is taken over the users and SBSs PPPs.
The average power consumed by the infrastructure is given by:

$$\mathbb{E}\left\{\rho_I\right\} = \rho \lambda_s \pi R_n^2, \tag{5.15}$$

ρ and $\lambda_s \pi R_n^2$ denote, respectively, the fix operational charge consumed by an active SBS and the average number of active SBSs. The average power used to retrieve a file either over the backhaul, when a cache miss event occurs, or from a SBS cache is given by:

$$\mathbb{E}\left\{\rho_f\right\} = \lambda_s \pi R_n^2 \big(\rho_{hd} \mathbb{P}\left\{hit\right\} + \rho_{bh}\big(1 - \mathbb{P}\left\{hit\right\}\big)\big), \tag{5.16}$$

where ρ_{hd} denotes the power needed to retrieve data from the local hard disk of a small BS when the requested content is already cached and a cache hit event occurs. ρ_{bh} denotes the power needed to retrieve data from the core network through the backhaul when a cache miss event occurs. Owing to channel inversion power control, the power used for transmission depends on the distance between the communicating SBS and users. Here we consider Υ_k as the set of users associated with cluster $k, \forall k = 1..N_c$. Each user looks for the requested file in the cache of the SBSs within a radius R, starting with the closest SBS from his own cluster. If the requested file is available in a cache within this distance, a cache hit event occurs and the user will associate with the closest SBS storing the requested file. In the event of a cache miss, the user associates with the nearest SBS from its corresponding cluster and the requested content is retrieved from the core network through the backhaul. If a user cannot find an SBS from its own cluster within a radius R, it only communicate with SBSs from other clusters within radius R, in the case of a cache hit event. The average total transmission power is given by:

$$\mathbb{E}\left\{\rho_T\right\} = \frac{\lambda_s \pi R_n^2}{U} \sum_{k=1}^{N_c} \sum_{u \in \Upsilon_k} \big(\mathbb{E}\left\{\rho_k\right\} + \sum_{j \neq k} \sum_{i \in \Delta_s} p_{iu}(1 - e^{-\lambda_{sj}\pi R^2})(\mathbb{E}\left\{\rho_j\right\} - \mathbb{E}\left\{\rho_k\right\})\big), \tag{5.17}$$

where $E\left\{\rho_k\right\}$ denotes the average transmit power that a typical user utilizes when communicating with SBSs associated with cluster $k, \forall k = 1..N_c$. Finally, the expression of the average consumed total power is derived by including the expressions of $\mathbb{E}\left\{\rho_k\right\}, \forall k = 1..N_c$.

Lemma 10. *The average consumed total power in the considered network with caching capabilities and content-popularity based user clustering can be modeled as follows:*

$$\rho^c_{total} = \lambda_s \pi R_n^2 (\rho_{hd} \mathbb{P}\left\{hit\right\} + \rho_{bh}\big(1 - \mathbb{P}\left\{hit\right\}\big) + \rho) + \frac{\lambda_s \pi R_n^2}{U} \sum_{k=1}^{N_c} \sum_{u \in \Upsilon_k} \Big(\frac{\rho_0 \gamma(\frac{\alpha}{2} + 1, \pi \lambda_{sk} R^2)}{(\lambda_{sk}\pi)^{\frac{\alpha}{2}}} \tag{5.18}$$

$$+ \sum_{j \neq k} \sum_{i \in \Delta_s} p_{iu}(1 - e^{-\lambda_{sj}\pi R^2})\Big(\frac{\rho_0 \gamma(\frac{\alpha}{2} + 1, \pi \lambda_{sj} R^2)}{(\lambda_{sj}\pi)^{\frac{\alpha}{2}}} - \frac{\rho_0 \gamma(\frac{\alpha}{2} + 1, \pi \lambda_{sk} R^2)}{(\lambda_{sk}\pi)^{\frac{\alpha}{2}}}\Big)\Big).$$

Proof. See C.1. □

Following the same reasoning, the average consumed total power in the network with no proactive caching capabilities at the SBSs, is given by:

$$\rho^{nc}_{total} = \lambda_s \pi R_n^2 \rho_{bh} + \rho \lambda_s \pi R_n^2 + \lambda_s \pi R_n^2 \frac{\rho_0 \gamma(\frac{\alpha}{2} + 1, \pi \lambda_s R^2)}{(\lambda_s \pi)^{\frac{\alpha}{2}}}. \tag{5.19}$$

ρ^{nc}_{total} is taken into consideration in order to guarantee an improvement in the average EE of the network, when proactive caching is implemented.

5.4.3 Average Spectral Efficiency

In order to derive the achievable EE, the expression of the average spectral efficiency should be derived. The DL SINR for a user u taken at the origin is given by:

$$\text{SINR} = \frac{\rho_0 \left\|h_u\right\|^2}{\sigma^2 + \sum_{k=1}^{N_c} I_k}, \tag{5.20}$$

where $I_k,\ \forall k = 1..N_c$ represents the interference coming from SBS from cluster k given by $I_k = \sum_{i \in \phi_{sk}} \rho_{ik} \|h_{ui}\|^2 r_{ui}^{-\alpha}$. Here ϕ_{sk} denotes the set of SBSs associated with cluster $k,\ k = 1..N_c$. ρ_{ik} refers to the power used in the DL by SBS i from cluster k. σ^2 represents the noise power. In order to compute the average spectral efficiency of the network, first we need to derive the achievable coverage probability which is given in the following Lemma:

Lemma 11. *The downlink coverage probability is given by:*

$$\mathbb{P}\{\mathit{SINR} \geq \theta\} = exp(-\frac{\theta}{\rho_0}\sigma^2) \times \prod_{k=1}^{N_c} exp\big(-\pi\lambda_{sk}\Gamma(1+\frac{2}{\alpha})\Gamma(1-\frac{2}{\alpha})(\frac{\theta}{\rho_0})^{\frac{2}{\alpha}}\mathbb{E}\left[\rho_k^{\frac{2}{\alpha}}\right]\big). \tag{5.21}$$

Proof. See C.2. □

We can see from Lemma 11, that increasing the SBS density enables to reduce the used transmit power. Nevertheless, we need to take into consideration the constant power consumed by the infrastructure of active SBSs which represents an important part of power consumption of the network. The average achievable spectral efficiency can be written as:

$$SE = \lambda_s \pi R_n^2 \log(1+\theta)\, \mathbb{P}\{\text{SINR} \geq \theta\}. \tag{5.22}$$

Given the average achievable spectral efficiency and average consumed power, we can derive a closed form expression of the energy efficiency Σ:

$$\Sigma = \frac{\lambda_s \pi R_n^2 \log(1+\theta)\, \mathbb{P}\{\text{SINR} \geq \theta\}}{\rho_{total}^c}. \tag{5.23}$$

By substituting (2.18) and (2.21) into (2.23), we obtain the average EEs. We can notice from (2.21) that the density of SBSs is a major defining parameter of Σ.

5.4.4 Analysis of Energy Efficiency

We can see, in (2.22), that increasing the SBS density results in a reduction in the interference. This is mainly due to the resulting decrease in transmit power since users are closer to their serving SBSs. Nevertheless, increasing SBS density results in more power consumption due to the strain of active infrastructure. We aim at finding the optimal active SBS density vector that maximizes the achievable EE, even when user positions are not taken into consideration. We can imagine a setting in which SBS are activated and shutdown based on user density and popularity profiles. We consider a constraint in which we aim at maintaining a power budget that is lower than that used when no proactive caching is enabled. The problem can be formulated as follows:

$$\underset{\Lambda_s}{\text{maximize}} \;\; \Sigma \tag{5.24}$$

$$\text{subject to} \;\; \rho_{total}^c - \rho_{total}^{nc} \leq 0, \tag{4.24a}$$

$$\Lambda_s^\dagger \mathbf{1} \leq \lambda_{s_{max}}. \tag{4.24b}$$

This optimization problem allows to derive the optimal density vector needed to maximize the average EE for a given user density, popularity profiles and cache size. Although the closed form expression of Σ is complex to analyze, it can be proven that Σ is quasi concave using an intelligent simplification by considering a composition of Σ with an affine mapping [200].

Theorem 12. *The considered optimization problem is quasi concave and the optimal SBS density Λ_s^* can be derived, with zero duality gap. If $\exists \Lambda_s^*$ such that $\nabla \Sigma(\Lambda_s^*) = 0$ then this vector is unique and it is the optimal solution. If this condition is not satisfied for any Λ_s such that $\Lambda_s^\dagger \mathbf{1} \leq \lambda_{s_{max}}$, then the optimal solution can be found using the Karush-Kuhn-Tucker (KKT) conditions:*

$$\begin{aligned} &\nabla L(\Lambda_s^*, \varsigma, \kappa) = \nabla \Sigma(\Lambda_s^*) + \varsigma \nabla C(\Lambda_s^*) + \kappa \nabla H(\Lambda_s^*) = 0, \\ &\varsigma C(\Lambda_s^*) = 0, \kappa H(\Lambda_s^*) = 0, \\ &H(\Lambda_s^*) \leq 0, C(\Lambda_s^*) \leq 0, \\ &\varsigma > 0, \kappa > 0, \end{aligned} \tag{5.25}$$

where L refers to the Lagrangian associated with problem (24), $C(\Lambda_s) = \rho_{total}^c - \rho_{total}^{nc}$, $H(\Lambda_s) = \Lambda_s^\dagger \mathbf{1}$. ς and κ refers to the Lagrangian multipliers associated, respectively, with $(24a)$ and $(24b)$.

Proof. See C.3. □

Finding the optimal density vector Λ_s^* based on the KKT conditions in (4.25), can be done using, for example, the sub-gradient descent method [200].

5.5 Exploiting spatial correlation in users demand

While, in the previous sections, EE was optimized with respect to the density vector of active SBSs, further EE gain can be achieved by including spatial information whenever it is available. The choice to decouple the two problems of cached file selection and content placement can be justified by the fact that popularity distributions change slower than user locations. Consequently, the selected cached content which depends on the average popularity distribution per cluster, is kept constant for long periods and the network can adapt its location based on user movement. Real life examples can also support this approach. While correlation in content popularity between users from the same social group stays for long periods of time, their location can change due to mobility. This motivates the need to adapt the cached content placement more often than the selected files in order to simplify the management of the network. Acquiring information on user location requires a non negligible processing and signaling overhead. Consequently, this information should be leveraged whenever it is available.

In the case of low mobility, where users do not change positions too often, it makes sense to adapt the files placement based on location information. Adapting the cached files placement can be done periodically or whenever the backhaul load allows it.

Practically, we may observe a spatial correlation in user file demand. This can be explained by the fact that people from the same social group (living or working in the same place) are most likely to have similar preferences. We aim at finding an effective allocation of the SBSs to the different clusters in order to minimize transmit power and, consequently, to improve the achievable EE. In fact, decreasing the distance between a given user and the SBS storing its requested file results in lower transmit power. Consequently, by effectively allocating the SBSs to the different clusters, we are able to lower the level of interference in the network, which results in increasing the EE [76].

The problem of cache placement can be tackled by adopting a hybrid approach where, a clustering based on both the location and content popularity is performed. This approach is more complex and do not necessarily produce better results since content popularity stays constant for long periods of time. In addition, the clustering that is done on the users enables also to group the files accordingly. Consequently, the complexity of the resulting optimal file placement problem is lower since the search space is reduced from the whole file catalog to groups of file of approximately equal total size. This simplifies the management of the caching system compared to existing work on location based optimization where, the complexity of the formulated problems is proportional to the number of files. This results in a considerable gain in running time which enables the network to be more reactive since it can adapt the content placement depending on user location very rapidly.

We consider a setting in which the location of all users and SBSs are known. This is implemented by considering a snapshot of the users and SBSs PPPs. We develop an integer optimization problem where we aim at minimizing the used power over the possible SBS-cluster affectation. Thanks to channel inversion power control, minimizing the used power is equivalent to reducing the distance between the users and the SBS caching the files they are most likely to request. We define $\omega_{u,s}$, the weight of the link between user u and the SBS s as follows:

$$\omega_{u,s} = \begin{cases} r_{us}^{-\alpha} & \text{if } r_{us} < R, \\ \omega_{\infty} & \text{otherwise,} \end{cases} \tag{5.26}$$

where r_{us} representing the distance between user u and the SBS s. ω_{∞} is an arbitrarily large value. ω_{∞} assures that no user can communicate with a SBS at a distance larger than R. We sort the links pathloss coefficients in decreasing order and denote by $(s)_u$ the SBS with the s-th greatest pathloss coefficient to user u. Based on the considered system model, less power is used when a user is served from a SBS within its neighborhood. Consequently, maximizing $\omega_{u,s}$ is equivalent to minimizing the transmit power and the distance between the user and its serving SBS. The average number of SBS associated with each cluster k is given by $N_{sk} = \lambda_{sk}\pi R_n^2$. Since $\lambda_{sk},\ k = 1..N_c$ are computed in (24) so that EE is maximized, it does not take into consideration their spatial repartition in the network. The transmit power increases when users from the same cluster are not located within a reduced area. In order to deal with this problem, we relax the constraint on the number of SBS per cluster and we replace N_{sk} by N'_{sk} where $N'_{sk} > N_{sk}$. We consider the adjacency matrix Y, where $y_{s,k}, \forall s = 1..N_s, k = 1..N_c$ is given by:

$$y_{s,k} = \begin{cases} 1 & \text{if SBS } s \text{ is associated with cluster } k. \\ 0 & \text{otherwise.} \end{cases} \tag{5.27}$$

The problem of optimal SBS allocation to their respective clusters can be formulated as:

$$\max_{Y} \sum_{k=1}^{N_c} \sum_{s=1}^{N_s} \sum_{u=1}^{U} \big(\sum_{f \in \Delta_k} p_{fu} \big) \omega_{u,(s)_u} \big(y_{(s)_u,k} \prod_{i=1}^{s-1} (1 - y_{(i)_u,k}) \big) \tag{5.28}$$

$$\text{subject to} \sum_{k=1}^{N_c} y_{s,k} \leq 1, \forall s = 1..N_s, \tag{4.28a}$$

$$\sum_{s=1}^{N_s} y_{s,k} \leq N'_{sk}, \forall k = 1..N_c. \tag{4.28b}$$

Here, $(4.28a)$ captures the fact that each SBS stores the most popular files of one unique cluster. $(4.28b)$ indicates that the number of SBSs allocated to each cluster should respect the density vector Λ_s^* which maximizes EE. The objective function in (4.28) guarantees that each SBS caches the files that are most likely to be requested by nearby users. In fact, $(y_{(s)_u,k} \prod_{i=1}^{s-1} (1 - y_{(i)_u,k}))$ is an indicator function that refers to the case where the most popular files of cluster k are cached in SBS $(s)_u$ and not in SBSs $(i)_u, i = 1, ..., s-1$. Consequently, the objective function value is equal to the expected pathloss between users and their serving SBSs. We show that the considered optimization problem is NP-hard. We then prove that it can be formulated as the maximization of a submodular function over matroid constraints and we provide an algorithm that enables to derive a $(1 - \frac{1}{e})$ approximation of the optimal solution of problem (4.28). This formulation looks somehow similar to the considered problem in [45] where, the authors aim at optimizing the allocation of each individual file to a set of femto access points in order to minimize the expected downloading time. Nevertheless, problem (4.28) consider a different objective function where, the aim is to minimize the transmit power. While [45] aims at optimizing the assignment of each individual file to the different femto access points, the objective in (4.28) is, actually, to assign predefined batches of files from each cluster to the SBSs. Consequently, the problem formulation here enables to considerably reduce the complexity of deriving a solution. In fact, the running time depends on the number of popularity based clusters rather than the number of files. This is an important impact of the present formulation in (4.28) since the number of files is typically very large. The considered setting enables to solve problem (4.28) using sophisticated algorithms that can be computationally prohibitive otherwise.

5.5.1 Computational Intractability

We start by showing the computational intractability of problem (4.28).

Theorem 13. *The considered optimization problem in* (4.28) *is NP-hard.*

Proof. In order to show that (4.28) is NP-hard, we consider a special case of our setting where $N'_{sk} = N \ \forall k = 1..N_c$ and $N_s = N_c$. This special case means that the number of SBS associated with each cluster is the same, which is the case when $\sum_{u=1}^{U} \sum_{f \in \Delta_k} p_{fu} = C, \ \forall k = 1..N_c$. In this case, the resulting optimization problem can be written as follow:

$$\begin{aligned} \max_{Y} & \sum_{k=1}^{N_c} \sum_{u=1}^{U} \sum_{s=1}^{N_s} C\omega_{u,(s)_u} (y_{(s)_u,k} \prod_{i=1}^{s-1} (1 - y_{(i)_u,k})) \\ \text{subject to} & \sum_{k=1}^{N_c} y_{s,k} \leq 1, \forall s = 1..N_s, \\ & \sum_{s=1}^{N_s} y_{s,k} \leq N, \forall k = 1..N_c. \end{aligned} \tag{5.29}$$

In order to show NP-hardness, we use a reduction from the following NP-hard problem:
Weighted K-Set Packing Problem: K-Set packing is an NP-hard combinatorial problem. It is one of the 21 problems of Karp [197]. The K-Set packing problem aims to find a maximum number of pairwise disjoint sets, with at most K elements, in a family S of subsets of a universal set V. The weighted version of the K-Set packing problem is obtained by assigning a real weight to each subset and maximizing the total weight.
We consider a collection of SBS sets $\{v_i, i = 1, ..., n\}$, associated each with a weight $\omega_{v_i} = \sum_{u=1}^{U} \max_{s \in v_i} \omega_{u,s}$. Problem (28) can then be formulated as a Weighted K-Set Packing Problem:

$$\begin{aligned} \underset{X}{\text{maximize}} & \sum_{i} C\omega_{v_i} x_i \\ \text{subject to } & v_i \cap v_j = \emptyset, \forall i, j, \\ & |v_i| \leq N, \forall k = 1..N_c, \\ & x_i \in \{0, 1\}. \end{aligned} \tag{5.30}$$

Solving (30) results in at most N_c sets of SBSs. Since the resulting sets are disjoint, each of them will be associated with a given cluster. The number of resulting sets could not exceed N_c since $N_s = N_c$. We can see that solving the weighted K-Set Packing Problem, for $K = N$ and where the weight of each subset is given by $C\omega_{v_i}$, is equivalent to solving the special case of the SBS allocation problem in (4.29). Knowing that the Weighted K-Set Packing Problem is NP-hard, we can then conclude that (4.28) is also NP-hard. □

5.5.2 Optimizing small base station allocation

In order to solve the considered optimization problem in (4.28), we start by showing that it is equivalent to the maximization of a sub-modular set function over matroid constraints. The definitions of matroids and sub-modular set functions can be found in [196]. This structure allows the use the randomized algorithm proposed in [198] which achieves, at least, $(1-\frac{1}{e})$ of the optimal value. Taking into consideration the problem constraints we have the following:

Lemma 14. *The Considered Optimization problem in* (4.28) *is equivalent to a maximization of a sub-modular set function over matroid constraints.*

Proof. See C.4. □

In order to solve the considered problem, we use the randomized algorithm proposed in [198]. This algorithm provides a $(1 - \frac{1}{e})$-approximation of the optimal solution for sub-modular set function maximization with matroid constraints. This algorithm consists of two steps. In the first one, a fractional solution of the relaxed problem is obtained using a continuous greedy process. In the second part of the algorithm, the derived fractional solution is rounded using a variant of the pipage rounding technique [199]. The detailed algorithm is given in table 5.2.
We define the ground set $G = \{g_{s,k}, \forall s = 1, \ldots, N_s, k = 1, \ldots, N_c\}$, where each element $g_{s,k}$ represents the allocation of SBS s to cluster k. We also define the function f over G as a set function that is equivalent to the objective function of (5.28). We also define $f_k, k = 1, \ldots, N_c$ as the set functions associated with each cluster $k = 1, \ldots, N_c$.

1. Run Continuous Greedy(f) to obtain a fractional solution $\tilde{Y}$

2. Run Pipage Rounding($\tilde{Y}$) to obtain a discrete solution Y

Table 5.2: **Randomized algorithm for Transmit power minimization over SBS-Cluster association**

```
1. Define $\delta = \frac{1}{9d^2}$ where $d = N_s$.
2. $t \longleftarrow 0$ and $\tilde{y}_{s,k}(0) = 0, \forall s, k$
3. For $k = 1, \ldots, N_c$
4.Define $R_k(t), k = 1...Nc$ a set containing each SBS $s$ with probability $\tilde{y}_{s,k}(t)$
5. For all $k = 1, \ldots, Nc, \quad s = 1, \ldots, Ns$
6. Define $\Xi_{s,k}(t) = \mathbf{E}[f_k(R_k(t) + g_{s,k}) - f_k(R_k(t))]$ obtained
by averaging $\frac{10}{\delta^2}(1 + \ln(N_c N_s))$ independent samples.
7. For $s = 1, \ldots, N_s$
8. $k_s(t) = \text{argmax}_k(\Xi_{s,k}(t))$ be the best cluster for SBS $s$.
9. Set $\tilde{y}_{s,k}(t+\delta) \longleftarrow \tilde{y}_{s,k}(t) + \delta$ for $k = k_s(t)$ and
$\tilde{y}_{s,k}(t+\delta) \longleftarrow \tilde{y}_{s,k}(t)$ for $k \neq k_s(t)$
10. $t \longleftarrow t + \delta$
11. If $t < 1$ go back to step (3)
```

Table 5.3: **Continuous Greedy(f)**

```
1. While $\tilde{Y}$ is not integral do
2. Pick $\tilde{y}_{s,k}$ and $\tilde{y}_{s',k}$ be two fractional variables
3. Find $A$ a minimal tight set such that $g_{s,k} \in A$ and $g_{s',k} \notin A$
4. Denote $\epsilon^+$ the greatest value that can be added to $\tilde{y}_{s,k}$ without violating any constraints.
5. Denote $\epsilon^-$ the greatest value that can be added to $\tilde{y}_{s',k}$ without violating any constraints.
6. $\tilde{Y}^+_{s,k}$ defined by setting $\tilde{y}_{s,k} \longleftarrow \tilde{y}_{s,k} + \epsilon^+$, $\tilde{y}_{s',k} \longleftarrow \tilde{y}_{s',k} - \epsilon^+$ and
   $\tilde{y}_{i,j} \longleftarrow \tilde{y}_{i,j}, \forall\{i,j\} \neq [\{s,k\},\{s,k'\}]$
7. $\tilde{Y}^-_{s,k}$ defined by setting $\tilde{y}_{s,k} \longleftarrow \tilde{y}_{s,k} - \epsilon^-$, $\tilde{y}_{s',k} \longleftarrow \tilde{y}_{s',k} + \epsilon^-$ and
   $\tilde{y}_{i,j} \longleftarrow \tilde{y}_{i,j}, \forall\{i,j\} \neq [\{s,k\},\{s,k'\}]$
8. If $F(\tilde{Y}^+_{s,k}) > F(\tilde{Y}^-_{s,k})$ then
9. $\tilde{Y} \longleftarrow \tilde{Y}^+_{s,k}$
10. else
11. $\tilde{Y} \longleftarrow \tilde{Y}^-_{s,k}$
12. End while
```

Table 5.4: **Pipage Rounding ($\tilde{Y}$)**

In a typical setting, randomly rounding a fractional solution of an optimization problem does not preserve the feasibility of the solution, in particular when equality constraints are considered. Nevertheless, the pipage rounding technique in [199] enables to round a fractional solution so that the problem constraints are not violated. This can be seen in steps (4) and (5) in 5.4.

In our case the running time of the algorithm is $O((N_sN_c)^8)$ [198], where $N_s = \lambda_s\pi R_n^2$. This is quite convenient since the running time of the algorithm does not depend on the number of files in the catalog which can be very large. This an interesting result of content based clustering since it reduces the search space from the whole catalog to bins of files of approximately equal size.

5.6 Numerical Results AND Discussion

In this section, we investigate the impact of the different system parameters on the Cache hit probability and EE. We then investigate the impact of the SBS allocation algorithm on the performances of the network. We consider a circular region with an area of $A = 10Km^2$. We simulate two PPP processes, one for the users and another for the SBSs over this area. The respective densities of these process are λ and $\lambda_{s_{max}}$ with $\lambda >> \lambda_{s_{max}}$. The considered SBS density values are defined based on the typical communication range of a SBS. We consider a catalog constituted of $F = 2000$ files with different randomly generated sizes $L_i,\ i = 1...F$ in the range [10 MB...100 MB] [76]. We also consider the normalized cache size $\eta = \frac{M}{\sum_{i=1}^{F} L_i}$. We characterize each cluster by a given popularity based file ordering. For each user u, P_u is generated according to a Zipf distribution with parameter 1 [165], after randomly selecting a cluster file ordering. This results in a random allocation of the users to the different clusters. In order to run the clustering algorithm, we only need an interval over which the search of the number of clusters is carried on. In our simulations we take $[N_{cmin}...N_{cmax}] = [5...30]$. We consider a pathloss exponent $\alpha = 2.5$. We consider the following power values [76]:

$\rho_o(dBm)$	21	$\rho_{bh}(W)$	$10W$
$\rho(W)$	10.16	$\rho_{hd}(W)$	12.5×10^-5

Figure 5.2 shows the evolution of the cache hit probability as a function of the communication radius R for different SBS densities. Figure 2 also shows a comparison between the content popularity based clustering approach and the classical method of supposing the same content popularity among all users. As an example, for a communication radius of $0.9\ Km$ and a SBS density of $\lambda_s = 1$, we notice a substantial gain with a cache hit probability of 0.736 for the clustering scheme compared with a probability equal to 0.41 when caching the most popular files in all SBSs. We notice that the hit probability for the scheme without clustering saturates at a low value. This is due to the fact that the same set of files is cached in all the SBSs, which is clearly a suboptimal approach, especially when users are covered by multiple SBSs. The increase in hit probability for the clustering method is mainly due to the diversity of files cached in the SBS. Increasing the SBS density results in reducing the average distance from mobile users, which, consequently, results in improving the cache hit probability.

Figure 5.3 shows the evolution of the achievable EE as a function of the normalized cache size for different SBS densities. Figure 3 also shows a comparison between the content

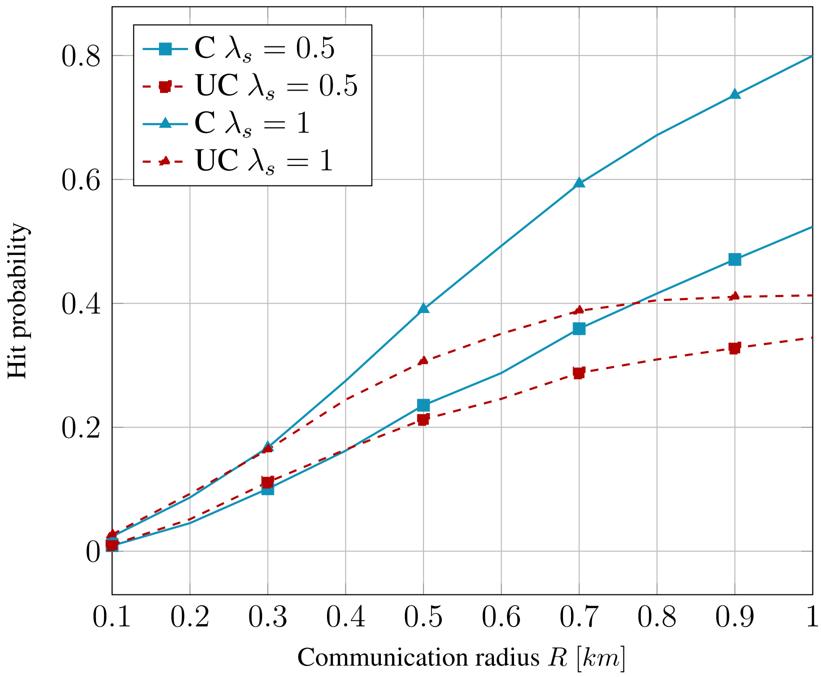

Figure 5.2: Hit probability versus communication radius with different SBS densities (C= content based clustering, UC= Unclustered approach), Normalized Cache Size $\eta = 0.25$

popularity based clustering approach and the classical method of supposing the same content popularity among all users. We can see that the proposed clustering method outperforms the classical approach of caching the same most popular files in all SBSs. For a normalized cache size of 0.4 and an SBS density of 1.6 SBS$/Km^2$, we notice an increase of 12.5% in the achievable EE. This gain is mainly due to the fact that the proposed method scores a higher hit probability than the unclustered approach. Consequently, the average energy needed to fetch the requested content is lower when user clustering is used. Even though restricting users to communicate with the closest SBS from their cluster, in the case of a cache miss event, can lead to an increase in the average transmit power, the observed gain in the energy used to fetch the desired content compensates for that. The gain in EE increases as a function of the SBS density. For a normalized cache size of 0.4 and an SBS density of 1.9 SBS$/Km^2$, we notice an increase of 14.2% in the achievable EE. This increase in EE gain can be explained by the fact that the average transmit power is a decreasing function of SBS density.

Figure 5.4 shows the performance of AIC model selection. We consider three settings in which, the true numbers of clusters are 10, 15 and 20, respectively. The figure represents the computed AIC per point over the estimation range. The lowest AIC value represents the model that strikes the best trade-off between fitness and complexity. Note that the negative values of the AIC are due to a negative bias which characterize the AIC with a small sample number.

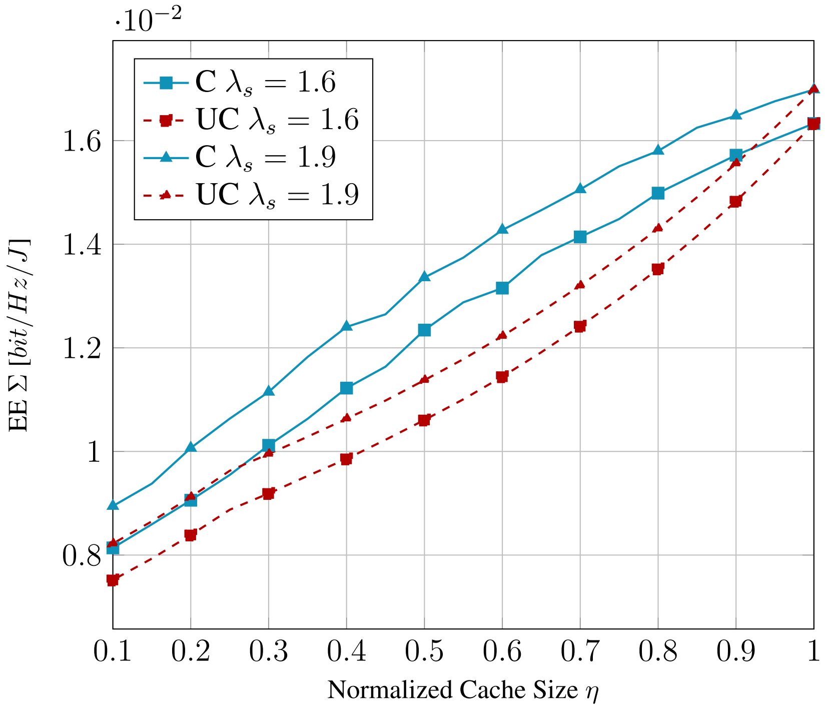

Figure 5.3: EE vs normalized cache size with different SBS densities (C= content based clustering, UC= Unclustered approach)

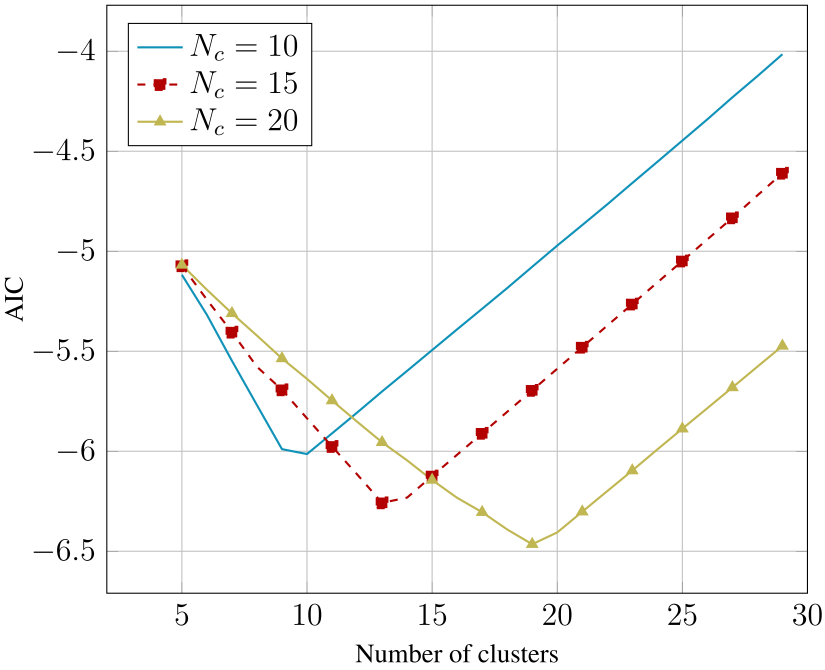

Figure 5.4: Akaike information criterion

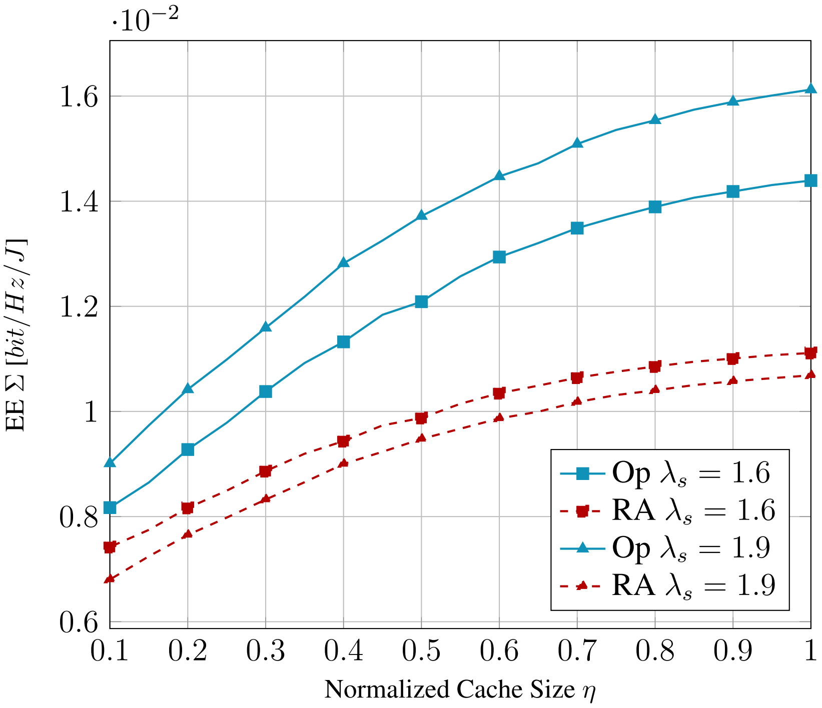

Figure 5.5: EE vs normalized cache size η with different SBS densities (RA= random SBS allocation, Op= optimized SBS allocation)

Figure 5.5 shows the impact of the SBS allocation algorithm on EE. We can see that, for different values of the SBS density, optimizing the SBS allocation results in a considerable gain in the EE. For a SBS density of $\lambda_s = 1.9$ and a normalized cache size of 0.4, optimizing the allocation of the SBSs results in an EE gain of 42.2%. As the SBSs density increases, the allocation algorithm results in greater improvement in EE. Optimizing the cluster-SBS association results in less average transmit power which reduces the interference and improves the achievable EE.

5.7 Closing Remarks

In this chapter, we studied a cache enabled small cell network with limited storage capability. We have proposed a novel proactive caching framework that utilizes content popularity based clustering in order to efficiently model traffic patterns and leverage the correlation between users demand. Our approach showed that taking into account heterogeneous popularity profiles and caching content accordingly yields a considerable increase in the networks capabilities to predict user traffic. This results in a more efficient proactive caching that can save substantial back-haul resources. We have also studied EE in cache enabled small cell networks. Using the proposed Content popularity based clustering, the optimal active SBS density vector that maximizes EE was derived. We then went one step further and optimized the networks EE with respect to cache placement. This allows to exploit any spatial correlation in user request patterns by bringing the cached files closer to the users that are most likely to request them. Numerical results shows that the proposed clustering framework considerably outperforms the scheme in which the files popularity are obtained by an averaging over all users. It also shows that optimized SBS allocation results in an improvement in the achievable hit probability and EE.

Chapter 6

Conclusions and Outlook

In this thesis, we have focused on improving the performance of 5G and beyond networks by leveraging all available information including, among others, traffic patterns, channel second order statistics and Doppler spread. The main idea was to exploit side information that can be obtained with low signaling overhead, thanks to its slow variation, in order to enable high efficiency wireless network operation. Indeed, filling the performance gap between LTE and 5G calls upon a more intelligent use of the available information and shifting part of the network intelligence down to the RAN side.

In this context, we proposed different novel schemes that aims at enabling a more service and user aware network. In particular, we focused on two key enablers of 5G, namely massive MIMO and proactive caching. Based on the observation that, already available information at the network is somehow underexploited, we devised novel procedures that allow for a more intelligent and efficient network.

In particular, in chapter 3 , we focused on TDD Massive MIMO systems and tried to address one of their major bottlenecks, namely CSI acquisition. As a matter of fact, it is now solidly established that massive MIMO can deliver a considerable increase in SE, EE and area capacity. Nevertheless, in order to achieve these gains, accurate CSI is needed. Consequently, if connection density is to be increased, more efficient CSI acquisition schemes are needed. In chapter 3, spatial diversity was leveraged in order to optimize UL training in TDD CSI systems. Based on the observation that, spatially independent users can be allowed to utilize the same pilot sequences, we proposed a smart UL training scheduling scheme that aims at increasing the connection density while, at the same time, improving the achievable SE. Indeed, we provide efficient algorithms that associate users in copilot groups based on their spatial information. In order to leverage a maximum of the Massive MIMO system DoFs, we form the copilot groups so that the users in each group provide a maximum coverage of the signal's spatial basis with minimum overlapping between user spatial signature. This enables to increase the pilot reuse in each cell which results in increasing connection density for the same training overhead. We provide two formulations of the user grouping problem. First, the problem of spatial basis coverage based copilot user selection is formulated as a *maximum coverage problem*. In the second case, the problem of copilot group generation is formulated as a *Generalized maximum coverage problem*. This allow us to provide efficient algorithms to perform the desired grouping. In addition, we analyze the performance of the proposed algorithms and we derive their achievable approximation ratios. We go one step further and address inter-cell copilot interference through an efficient pilot sequence allocation. Numerical results show that the provided schemes enable to mitigate copilot interference while increasing both connection density and SE.

In chapter 4, we focused on another underexploited information, namely Doppler spread. As a matter of fact, we noticed that current wireless systems assume the same time slot duration for all devices regardless of the fact that users are subject to heterogeneous Doppler spreads. This fact lead as to regard *CSI estimation periodicity* as an additional degree of freedom. The main idea in this chapter comes from the observation that users with low velocity are not required to send UL training sequences with the same periodicity as faster moving users due to the resulting heterogeneous coherence times. Consequently, a dynamic adaptation of the TDD frame based on heterogeneous Doppler spreads can lead to improvement in the network performance. In chapter 4, we proposed a planning framework where a network of Massive MIMO BSs is enabled to learn the best UL training policy for long time periods. Since channel changes result primarily from device mobility, location awareness was also included in the learning process. The resulting planning problem was modeled as a two time scale POMDP. Although complex, we proposed efficient algorithms that enables the network to optimize its training decisions for long time spans while reducing the required signaling overhead and processing complexity.

Traffic and service awareness is also a critical feature in high efficiency networks. As a matter of fact 5G networks are expected to be User-centric, trading the conventional service agnostic paradigm for a more intelligent and proactive one. This made proactive caching a key technology that can address a large range of 5G requirements, including low latency, energy efficiency and QoE. In chapter 5, we focused on improving proactive caching performance by leveraging the diversity in user traffic patterns. We provided an adaptive content popularity based clustering algorithm that enables the network to learn dominant traffic patterns. This approach provides more detailed information on the content that is most likely to be requested which enables a more efficient proactive caching. The proposed framework strikes the optimal trade-off between complexity and truthfulness in user behavior modeling. It also simplifies the management of the caching system since it groups files based on the cache capacity and on the dominant user behaviors. Building on the proposed model, content placement was optimized in order to maximize a major KPI of 5G networks, namely EE. Results show that a more detailed and truthful modeling of user preference can lead to an increase in the majority of proactive caching KPIs.

Despite the fact that the proposed schemes in this thesis enable considerable gains in 5G networks, there exists several challenges which need to be investigated in the future.

In particular, in chapter 3 of the thesis where we have focused on exploiting spatial diversity, we have the following future directions.

- *Include mobility and spatial information variation* : In chapter 3, constant spatial information were considered. This is true when we are dealing with time spans during which user mobility does not change the channel's second order statistics. Devising new training schemes that take into consideration mobility and that are able to track or predict the signal subspace may enable more efficient CSI acquisition.

- *Spatial division multiplexing and multicast communication in massive MIMO systems* : The potential of massive MIMO in the context of multicast transmission is considerable. Indeed, owing to its ability to efficiently shape the transmitted signal, massive MIMO is well suited to multicast communication. In this context, optimizing spatial division multiplexing for multicast communication can lead to considerable improvement of the network performance.

- *Spatial division in CRAN systems* :

  Wireless networks are evolving from a cellular to a cell-free topology. CRAN makes such distributed systems possible by centralizing part of the physical layer processing. In this context, Cell-free Massive MIMO systems, where users are served by a large number of distributed APs over the coverage area, is a very interesting architecture that can fully benefit from a spatial information based user scheduling and a centralized processing. Owing to the resulting macro-diversity, Cell-free Massive MIMO has an considerable potential in addressing the requirement of high density connection. Nevertheless, in order to be able to handle a large number of connected devices, efficient CSI acquisition schemes, that take into consideration the special architecture of Cell-free Massive MIMO, need to be developed.

For the fourth chapter, our future directions for enabling a more self organizing and intelligent massive MIMO network can be summarized as follows.

- *Include handover and more sophisticated mobility patterns*: Although user mobility was considered in our work, handovers were not included in the analysis. Designing more advanced learning algorithms for UL training with complex mobility patterns and taking into consideration handovers can lead to longer time spans optimization.

- *A fully dynamic TDD frame*: The proposed scheme supposes that the TDD frame can be defined in flexible manner depending on the required training overhead. The concept of dynamic TDD, where the split between UL and DL resources is defined dynamically is starting to attract more attention. However, the current state of development of the latter concept does not include a dynamic allocation of training resources. Consequently, combining the proposed scheme with the Dynamic TDD concept in a fully flexible frame where resources are allocated as a function of the actual UL, DL and training needs may enable substantial improvement of different network KPIs.

- *Efficient algorithms for high density connectivity*: The analysis in our work payed special attention to the planning algorithms complexity. Nevertheless, in high density scenarios, planning the network CSI estimation policy for long time periods becomes very challenging. Consequently, more sophisticated planning algorithms that are suited to these scenarios are needed. We can think about exploiting other side informations such as traffic patterns in order to reduce the complexity of deriving optimal training policies.

- *CSI estimation planning in the mmWave range*: mmWave communication is considered in the 5G standards. Although it enables huge SE gains, mmWave communications come with its toll of challenges, particularly, small coherence times. In this range of frequency, efficient training becomes vital. An extension of CSI planning for mmWave communications seems to be in order.

For the fifth chapter, our future directions for content popularity learning and algorithmic aspects of proactive caching can be summarized as follows.

- *Proactive caching for 5G vehicular networks*:

  Drivers and mobile users in general typically spend a non-negligible fraction of time in vehicles. The development of 5G vehicular networks with specific protocols provides unlimited peer-to-peer capabilities and additional communication opportunities that can be exploited in proactive caching. Nevertheless, specific proactive caching are needed in this case owing to a range of additional constraints that arise from high mobility.

- *A more detailed characterization of the demand*: An interesting future direction of this work is to conduct a more detailed characterization of the traffic which captures different spatio-temporal content access patterns for caching.

- *Proactive caching and multicast*: Proactive caching provides more multi-casting opportunities at the BSs. Distinguishing between multicast-suited content and other popular files may allow to better exploit multicast communication while satisfying other popular content demands.

- *Social networks and relations*: Involving social networks information can lead to better understanding of traffic patterns. In fact, social influence is of paramount importance when trying to predict human behavior. An intelligent learning framework for mutual influence between users may enable to better predict future popular content and to derive better cache placement strategies.

- *Leveraging high density networks through coded caching*: The expected high density deployment of connected devices can be leveraged in the framework of proactive content provisioning through coded caching. Intelligent coded caching schemes that take into consideration the particularly tight EE and batteries life span constraints of some devices need to be developed.

Appendix A

TDD Massive MIMO systems: Enhancing CSI estimation through Spatial Division based training

A.1 Proof of Theorem 3

In order to derive a tight bound on the performance of the proposed algorithm in table 3.1, we consider the worst case behavior of the greedy heuristic as in [198]. However in our case, the problem is more sophisticated since it includes the combination of the greedy heuristic with an approximation algorithm.

The optimization problem (3.12) can be decomposed into N_c independent problems, each of which is defined in a given cell $b = 1, \dots, N_c$. We start by defining $C_k^{[b]}, k = 1, \dots, \tau, b = 1, \dots, N_c$ as the maximum coverage at iteration k of the approximate algorithm in table I. We define $C_{k_{opt}}^{[b]}, k = 1, \dots, \tau, b = 1, \dots, N_c$ as the optimal maximum coverage that can be obtained at iteration k. We also consider $C_{opt}^{[b]}$ as the optimal solution the problem of (3.12) defined in each cell $b = 1, \dots, N_c$:

$$V(C_{opt}^{[b]}) = \max_{Y} \sum_{k=1}^{\tau} \sum_{s=1}^{M} y_{s,b}^{[k]} \tag{A.1}$$

$$\text{s.t} \sum_{i} x_{\{i,b\}}^{[k]} \leq U_b^{[k]} \quad \forall k = 1 \dots \tau \tag{A.1a}$$

$$\sum_{i, f_s \in F_{ib}} x_{\{i,b\}}^{[k]} \geq y_{s,b}^{[k]} \quad \forall k = 1 \dots \tau, \tag{A.1b}$$

where $V(C_{opt}^{[b]})$ represents the value of the coverage $C_{opt}^{[b]},\ b = 1, \dots, N_c$. The objective function in (A.1) is modular. Consequently, the following property holds $\forall b = 1, \ldots, N_c$:

$$V(C^{[b]}_{opt}) \leq \sum_{k=1}^{t-1} V(C^{[b]}_{k_{opt}}) + \tau V(C^{[b]}_{t_{opt}}) \ \ , \forall t = 1, ..., \tau, \tag{A.2}$$

In order to derive a bound on the achievable performance of the proposed algorithm, we consider the worst case behavior of the greedy heuristic. Doing so is equivalent to solving the following linear problems $\forall\ b = 1, ..., N_c$:

$$P(b) = \min \sum_{k=1}^{j} \frac{V(C^{[b]}_{k_{opt}})}{V(C^{[b]}_{opt})} \tag{A.3}$$

$$\text{s.t} \sum_{k=1}^{t-1} \frac{V(C^{[b]}_{k_{opt}})}{V(C^{[b]}_{opt})} + \tau \frac{V(C^{[b]}_{t_{opt}})}{V(C^{[b]}_{opt})} \geq 1 \ \ \forall t = 1...j \tag{A.3a}$$

Where the constraint $(A.3a)$ is obtained from (A.2). Since (A.3) is a linear problem, we can solve it by considering its dual $\forall b = 1, ..., N_c$ which can be written as follows

$$D(b) = \max \sum_{t=1}^{j+1} v_t \tag{A.4}$$

$$\text{s.t } \tau v_k + \sum_{t=k+1}^{j+1} v_t = 1 \ \ , \forall k = 1...j \tag{A.4a}$$

$$v_t \geq 0 \ \ \forall t = 1...j+1 \tag{A.4b}$$

We now proceed by solving (A.4), and, consequently, by linear programming duality, (A.3). Let $v = 1 - v_{j+1}$, where v_{j+1} is defined in (A.4). Consequently, $\forall k = 1...j$, we have:

$$\tau v_k + \sum_{t=k+1}^{j} v_t = v \text{ and } v_k = \frac{v - \sum_{t=k+1}^{j} v_t}{\tau} \tag{A.5}$$

Then v_k are calculated iteratively $\forall\ k = 1...j$. Indeed,

$$v_j = \frac{v}{\tau},\ v_{j-1} = \frac{v - v_j}{\tau} = \frac{v - \frac{v}{\tau}}{\tau} = \frac{v}{\tau}(\frac{\tau - 1}{\tau}), \ldots,\ v_1 = \frac{v - \sum_{t=2}^{j} v_t}{\tau} = \frac{v}{\tau}(\frac{\tau-1}{\tau})^{j-1}. \tag{A.6}$$

Consequently, $v_k, \forall\ k = 1...j$ are given by

$$v_k = \frac{v}{\tau}(\frac{\tau - 1}{\tau})^{j-k} \tag{A.7}$$

Therefore, finding $D(b), \forall b = 1, ..., N_c$ is equivalent to the following :

$$D(b) = \max_{0 \leq v \leq 1} (\sum_{t=1}^{j+1} v_t) = \max_{0 \leq v \leq 1} (v(1 - (\frac{\tau - 1}{\tau})^j)) \tag{A.8}$$

which is achieved for $v = 1$. It follows that, $P(b) = 1 - \frac{j}{\tau}(\frac{\tau-1}{\tau})^j$. Taking $j = \tau$, we obtain:

$$P(b) = 1 - (\frac{\tau - 1}{\tau})^\tau \tag{A.9}$$

Consequently, the obtained solution using a greedy sequential maximum coverage verifies:

$$\frac{V(C_{opt}^{[b]}) - \sum_{k=1}^{\tau} V(C_{k_{opt}}^{[b]})}{V(C_{opt}^{[b]})} \leq (\frac{\tau - 1}{\tau})^\tau \tag{A.10}$$

In order to derive the performance guarantee of the proposed algorithm in table 3.1, the approximation ratio of the used algorithm to perform maximum coverage, at each iteration, needs to be accounted for. We use the following result which is based on the approximation ratio given in [79].

Lemma 15. *For any given cell index $b = 1, \ldots, N_c$ and iteration $k = 1, \ldots, \tau$, the implemented algorithm provides a $(1 - \frac{1}{e})$ approximation of the optimal cover, i.e $C_k^{[b]} \geq (1 - \frac{1}{e}) C_{k_{opt}}^{[b]}, k = 1, \ldots, \tau, b = 1, \ldots, N_c$.*

Since at each iteration of the algorithm, in table 3.1, we obtain a $(1 - \frac{1}{e})$ approximation of the optimal maximum coverage, we obtain the following for $b = 1, \ldots, N_c$

$$\begin{aligned} \sum_{k=1}^{\tau} V(C_k^{[b]}) &\geq (1 - \frac{1}{e}) \sum_{k=1}^{\tau} V(C_{k_{opt}}^{[b]}) \\ \sum_{k=1}^{\tau} V(C_k^{[b]}) &\geq (1 - \frac{1}{e})(1 - (\frac{\tau - 1}{\tau})^\tau) V(C_{opt}^{[b]}) \end{aligned} \tag{A.11}$$

we can deduce that the algorithm in table I provides a $(1 - (\frac{\tau-1}{\tau})^\tau)(1 - \frac{1}{e})$-approximation for each subproblem of (3.12). Taking the sum over $b = 1, ..., N_c$ finishes the proof.

A.2 Proof of Theorem 5

The proof for the performance bound of the proposed algorithm in table 3.2 follows the same reasoning as the proof of Theorem 3. The main idea is also to consider the worst case behavior of the greedy heuristic with a change in the achievable approximation ratio at each iteration. The optimization problem (3.13) can be decomposed into N_c independent problems, each of which is defined in a given cell $b = 1, \ldots, N_c$. We start by defining $C_k^{[b]}, k = 1, \ldots, \tau, b = 1, \ldots, N_c$ as the maximum coverage at iteration k of the algorithm in table 3.2. We define $C_{k_{opt}}^{[b]}, k = 1, \ldots, \tau, b = 1, \ldots, N_c$ as the optimal maximum coverage that can be obtained at iteration k. We also consider $C_{opt}^{[b]}$ as the optimal solution the problem of (3.13) defined in each cell $b = 1, \ldots, N_c$:

$$V(C_{opt}^{[b]}) = \max_{Y} \sum_{k=1}^{\tau} \sum_{i \in \Gamma(b)} \sum_{f_s \in \mathbf{F}} \zeta_{ib}^{[s]} y_{\{i,b\}}^{[s,k]} \tag{A.12}$$

$$\text{subject to} \quad \sum_{i \in \Gamma(b), f_s \in F_{ib}} y_{\{i,b\}}^{[s,k]} \leq 1 \ \ \forall k = 1 ... \tau, \tag{A.12a}$$

$$\sum_{i \in \Gamma(b), f_s \in F_{ib}} x_{\{i,b\}}^{[k]} \geq y_{\{i,b\}}^{[s,k]} \ \ \forall k = 1 ... \tau \tag{A.12b}$$

$$\sum_{i \in \Gamma(b)} x_{\{i,b\}}^{[k]} \leq U_b^{[k]} \ \ \forall k = 1 ... \tau, \tag{A.12c}$$

The objective function in (A.12) is modular. Consequently, the following property holds $\forall b = 1, \ldots, N_c$:

$$V(C_{opt}^{[b]}) \leq \sum_{k=1}^{t-1} V(C_{k_{opt}}^{[b]}) + \tau V(C_{t_{opt}}^{[b]}) \ \ , \forall t = 1, ..., \tau, \tag{A.13}$$

In order to derive a bound on the achievable performance of the proposed algorithm, we consider the worst case behavior of the greedy heuristic. Doing so is equivalent to solving the following linear problems $\forall \ b = 1, ..., N_c$:

$$P(b) = \min \sum_{k=1}^{j} \frac{V(C_{k_{opt}}^{[b]})}{V(C_{opt}^{[b]})} \tag{A.14}$$

$$\text{s.t} \sum_{k=1}^{t-1} \frac{V(C_{k_{opt}}^{[b]})}{V(C_{opt}^{[b]})} + \tau \frac{V(C_{t_{opt}}^{[b]})}{V(C_{opt}^{[b]})} \geq 1 \ \ \forall t = 1 ... j \tag{A.14a}$$

Where the constraint $(30a)$ is obtained from (A.13). Here also (A.14) is solved using its dual. It follows that, $P(b) = 1 - \frac{j}{\tau}(\frac{\tau-1}{\tau})^j$. Taking $j = \tau$, we obtain:

$$P(b) = 1 - (\frac{\tau - 1}{\tau})^{\tau}, \forall b = 1, \ldots, N_c. \tag{A.15}$$

Consequently, the obtained solution using a greedy sequential maximum coverage verifies:

$$\frac{V(C_{opt}^{[b]}) - \sum_{k=1}^{\tau} V(C_{k_{opt}}^{[b]})}{V(C_{opt}^{[b]})} \leq (\frac{\tau - 1}{\tau})^{\tau} \tag{A.16}$$

In order to derive the performance guarantee of the proposed algorithm in table 3.2, the approximation ratio of the used algorithm , at each iteration, needs to be considered. We use the following result which is based on the approximation ratio given in [80].

Lemma 16. *For any given cell index $b = 1, \dots, N_c$ and iteration $k = 1, \dots, \tau$, using an algorithm of approximation ratio β at step* 3 *of the algorithm in table 3.3, the implemented algorithm provides a $\frac{1+\beta-\beta e^{-\frac{1}{\beta}}}{1-e^{-\frac{1}{\beta}}}$ approximation of the optimal cover, i.e $C_k^{[b]} \geq \frac{1+\beta-\beta e^{-\frac{1}{\beta}}}{1-e^{-\frac{1}{\beta}}} C_{k_{opt}}^{[b]}, k = 1, \dots, \tau, b = 1, \dots, N_c$.*

In this work, we use the greedy algorithm for knapsack problems in step 3 of the algorithm in table 3.3. Consequently, in our case, $\beta = \frac{1}{2}$ and the approximation ratio, at each iteration, becomes $\frac{\frac{3}{2}-\frac{e^{-2}}{2}}{1-e^{-2}}$

Since at each iteration of the algorithm, in table 3.2, we obtain a $\frac{\frac{3}{2}-\frac{e^{-2}}{2}}{1-e^{-2}}$ approximation of the optimal maximum coverage, we obtain the following for $b = 1, \dots, N_c$

$$\sum_{k=1}^{\tau} V(C_k^{[b]}) \geq \frac{\frac{3}{2}-\frac{e^{-2}}{2}}{1-e^{-2}} \sum_{k=1}^{\tau} V(C_{k_{opt}}^{[b]}) \tag{A.17}$$

$$\sum_{k=1}^{\tau} V(C_k^{[b]}) \geq \frac{\frac{3}{2}-\frac{e^{-2}}{2}}{1-e^{-2}} (1 - (\frac{\tau-1}{\tau})^{\tau}) V(C_{opt}^{[b]}) \tag{A.18}$$

we can deduce that the algorithm in table 3.2 provides a $(1 - (\frac{\tau-1}{\tau})^{\tau})\frac{\frac{3}{2}-\frac{e^{-2}}{2}}{1-e^{-2}}$-approximation for each subproblem of (3.13). Taking the sum over $b = 1, ..., N_c$ finishes the proof.

Appendix B

Dynamic TDD: Enhancing performance by long term CSI estimation planning

B.1 Proof of Theorem 6

The network serves N_g copilot groups, τ of which are scheduled for UL training. At the reception, each BS uses a matched filter receiver that is based on the latest available CSI estimates. BS l detects the signal of user g in cell l by applying the following filter $u_{gl}(t) = \frac{\hat{g}_{gl}^{[l]}(t-d_g)}{\|\hat{g}_{gl}^{[l]}(t-d_g)\|}, t \geq d_g$, where $\hat{g}_{gl}^{[l]}(t-d_g)$ denotes the latest available CSI estimate for user g in cell l. Consequently, the detected signal of user g in cell l is given by the following

$$\begin{aligned}
u_{gl}^{\dagger}(t)\frac{Y_u^{[l]}(t)}{\sqrt{P_u}} &= u_{gl}^{\dagger}(t)(\sum_{k=1}^{N_g}\sum_{c=1}^{C} g_{kc}^{[l]}(t)S_{kc} + \frac{W_u(t)}{\sqrt{P_u}}) \qquad \text{(B.1)}\\
&= u_{gl}^{\dagger}(t)((\rho_{gl}^{[l]})^{d_g}\hat{g}_{gl}^{[l]}(t-d_g)S_{gl} + \sum_{c\neq l}^{C}(\rho_{gc}^{[l]})^{d_g}\hat{g}_{gc}^{[l]}(t-d_g)S_{gc}\\
&+ \sum_{c=1}^{C}(\rho_{gc}^{[l]})^{d_g}\tilde{g}_{gc}^{[l]}(t-d_g)S_{gc} + \sum_{c=1}^{C}\sum_{j=0}^{d_g-1}(\rho_{gc}^{[l]})^{j}\sqrt{\beta_{gc}^{[l]}}\varepsilon_{gc}^{[l]}(t-j)S_{gc}\\
&+ \frac{W_u(t)}{\sqrt{P_u}} + \sum_{k\neq g}^{N_g}\sum_{c=1}^{C} g_{kc}^{[l]}(t)S_{kc})\\
&= u_{il}^{\dagger}(t)(I_1(t) + I_2(t) + I_3(t)),
\end{aligned}$$

with

$$I_1(t) = (\rho_{gl}^{[l]})^{d_g} \hat{g}_{gl}^{[l]}(t-d_g) S_{gl}, \tag{B.2}$$

$$I_2(t) = \sum_{c \neq l}^{C} (\rho_{gc}^{[l]})^{d_g} \hat{g}_{gc}^{[l]}(t-d_g) S_{gc}, \tag{B.3}$$

$$\begin{aligned} I_3(t) &= \sum_{c=1}^{C} (\rho_{gc}^{[l]})^{d_g} \tilde{g}_{gc}^{[l]}(t-d_g) S_{gc} + \sum_{c=1}^{C} \sum_{j=0}^{d_g-1} (\rho_{gc}^{[l]})^j \sqrt{\beta_{gc}^{[l]}} \varepsilon_{gc}^{[l]}(t-j) S_{gc} \\ &+ \sum_{k \neq g}^{N_g} \sum_{c=1}^{C} g_{kc}^{[l]}(t) S_{kc} + \frac{W_u(t)}{\sqrt{P_u}} \end{aligned} \tag{B.4}$$

The third equality in Equation B.1 follows from the fact that $g_{kc}^{[l]}(t) = \sqrt{\beta_{kc}^{[l]}} h_{kc}^{[l]}(t)$, $h_{kc}^{[l]}(t) = \rho_{kc}^{[l]} h_{kc}^{[l]}(t-1) + \varepsilon_{kc}^{[l]}(t)$ for all t and $g_{kc}^{[l]}(t) = \hat{g}_{kc}^{[l]}(t) + \tilde{g}_{kc}^{[l]}(t)$ for all t. We note that $I_1(\cdot)$ refers to the useful signal, $I_2(\cdot)$ represents the impact of pilot contamination and $I_3(\cdot)$ regroups the impact of the white noise, channel estimation error, non correlated interference due to users with different pilot sequences and the impact of channel aging.
The instant spectral efficiency attained by user g in cell l is:

$$R_{g,l} = \left(1 - \frac{\tau}{T}\right) \log \left(1 + \frac{|u_{gl}^{\dagger}(t) I_1(t)|^2}{|u_{gl}^{\dagger}(t) I_2(t)|^2 + |u_{gl}^{\dagger}(t) I_3(t)|^2}\right). \tag{B.5}$$

We now define $\overline{R}_{g,l}$ to be the average achievable sum rate of user g in cell l, namely,

$$\overline{R}_{g,l} = \mathbb{E}\left(\mathbb{E}\left(\left(1 - \frac{\tau}{T}\right) \log \left(1 + \frac{|u_{gl}^{\dagger}(t) I_1(t)|^2}{|u_{gl}^{\dagger}(t) I_2(t)|^2 + |u_{gl}^{\dagger}(t) I_3(t)|^2}\right) \middle| \hat{g}_{gl}^{[l]}(t-d_g)\right)\right), \tag{B.6}$$

the last equality follows from the law of total expectation. Let us define $\overline{R}_{g,l}^{0}$ such that

$$\overline{R}_{g,l}^{0} = \mathbb{E}\left(\left(1 - \frac{\tau}{T}\right) \log \left(1 + \frac{|u_{gl}^{\dagger}(t) I_1(t)|^2}{|u_{gl}^{\dagger}(t) I_2(t)|^2 + |u_{gl}^{\dagger}(t) I_3(t)|^2}\right) \middle| \hat{g}_{gl}^{[l]}(t-d_g)\right), \tag{B.7}$$

therefore, $\overline{R}_{g,l} = \mathbb{E}(\overline{R}_{g,l}^{0})$. Based on the convexity of $\log(1 + \frac{1}{x+a})$, and Jensen's inequality we obtain the following

$$\overline{R}_{g,l}^{0} \geq \left(1 - \frac{\tau}{T}\right) \log \left(1 + \frac{|u_{gl}^{\dagger}(t) (\rho_{gl}^{[l]})^{d_g} \hat{g}_{gl}^{[l]}(t-d_g)|^2}{\mathbb{E}(|u_{gl}^{\dagger}(t) I_2(t)|^2 | \hat{g}_{gl}^{[l]}(t-d_g)) + \mathbb{E}(|u_{gl}^{\dagger}(t) I_3(t)|^2 | \hat{g}_{gl}^{[l]}(t-d_g))}\right), \tag{B.8}$$

since

$$\mathbb{E}(|u_{gl}^{\dagger}(t) I_1(t)|^2 | \hat{g}_{gl}^{[l]}(t-d_g)) = |u_{gl}^{\dagger}(t) (\rho_{gl}^{[l]})^{d_g} \hat{g}_{gl}^{[l]}(t-d_g)|^2, \tag{B.9}$$

by the property $\mathbb{E}(f(Z)|Z) = f(Z)$ for a random variable Z.
We now aim at computing $\mathbb{E}(|u_{gl}^{\dagger}(t)I_j(t)|^2|\hat{g}_{gl}^{[l]}(t-d_g))$ for $j = 2, 3$. In order to do so, we are first going to obtain an alternative expression for $I_2(t)$, that is,

$$I_2(t) = \sum_{c \neq l}^{C} \hat{g}_{gc}^{[l]}(t-d_g)S_{gc} = \hat{g}_{gl}^{[l]}(t-d_g)\sum_{c \neq l}^{C} \frac{\beta_{gc}^{[l]}}{\beta_{gl}^{[l]}} S_{gc}, \tag{B.10}$$

since $\hat{g}_{gc}^{[l]}(t-d_g) = \hat{g}_{gl}^{[l]}(t-d_g)\frac{\beta_{gc}^{[l]}}{\beta_{gl}^{[l]}}$. Therefore, $I_2(t)$ and $\hat{g}_{gl}^{[l]}(t-d_g)$ are correlated. Consequently, we obtain

$$\mathbb{E}\left[|u_{gl}^{\dagger}(t)I_2(t)|^2|\hat{g}_{gl}^{[l]}(t-d_g)\right] = \left|u_{gl}^{\dagger}(t)\hat{g}_{gl}^{[l]}(t-d_g)\right|^2 \sum_{c \neq l}^{C} (\rho_{gc}^{[l]})^{2d_g} \frac{\beta_{gc}^{[l]^2}}{\beta_{gl}^{[l]^2}}. \tag{B.11}$$

We will now compute $\mathbb{E}(|u_{gl}^{\dagger}(t)I_3(t)|^2|\hat{g}_{gl}^{[l]}(t-d_g))$. First note that, $I_3(t)$ is independent of $\hat{g}_{gl}^{[l]}(t-d_g)$ and since $u_{gl}^{\dagger}(t)$ has unit norm, we have that $\mathbb{E}(|u_{gl}^{\dagger}(t)I_3(t)|^2|\hat{g}_{gl}^{[l]}(t-d_g)) = \mathbb{E}(|I_3(t)|^2)$, therefore we obtain

$$\begin{aligned}
\mathbb{E}(|I_3(t)|^2) = \mathbb{E}\bigg(| \sum_{c=1}^{C} (\rho_{gc}^{[l]})^{d_g} \tilde{g}_{gc}^{[l]}(t-d_g)S_{gc} + \sum_{c=1}^{C}\sum_{j=0}^{d_g-1} \sqrt{\beta_{gc}^{[l]}}(\rho_{gc}^{[l]})^j \varepsilon_{gc}^{[l]}(t-j)S_{gc} \qquad \text{(B.12)}\\
+ \sum_{k \neq g}^{N_g}\sum_{c=1}^{C} g_{kc}^{[l]}(t)S_{kc} + \frac{W_u(t)}{\sqrt{P_u}}|^2 \bigg) \\
= \mathbb{E}\bigg(\sum_{c=1}^{C} |(\rho_{gc}^{[l]})^{d_g} \tilde{g}_{gc}^{[l]}(t-d_g)|^2 + \sum_{c=1}^{C}\sum_{j=0}^{d_g-1} |\sqrt{\beta_{gc}^{[l]}}(\rho_{gc}^{[l]})^j \varepsilon_{gc}^{[l]}(t-j)|^2 \\
+ \sum_{k \neq g}^{N_g}\sum_{c=1}^{C} |g_{kc}^{[l]}(t)|^2 + |\frac{W_u(t)}{\sqrt{P_u}}|^2 \bigg),
\end{aligned}$$

where the last equality follows from noting the following four properties; (i) $S_{kc} \cdot S_{ic'} = 0$ for all $k \neq i$ and all $c, c' \in \{0, \ldots, C\}$, (ii) $\mathbb{E}(ZW_u(t)) = \mathbb{E}(Z)\mathbb{E}(W_u(t)) = 0$ for all random variables Z that are independent of $W_u(t)$ (zero mean complex Gaussian noise), (iii) similar to the previous property, $\mathbb{E}(Z\varepsilon_{ic}^{[l]}(t)) = \mathbb{E}(Z)\mathbb{E}(\varepsilon_{ic}^{[l]}(t)) = 0$ for all Z independent of $\varepsilon_{ic}^{[l]}(t)$ (zero mean complex white Gaussian noise) and finally (iv) $g_{kc}^{[l]}$ and $\tilde{g}_{k'c'}^{[l]}$ are independent for all $(k, c) \neq (k', c')$.

We now compute the four terms in Equation B.12. The last term, i.e.,

$$\mathbb{E}(|W_u(t)/\sqrt{P_u}|^2) = \frac{1}{P_u}. \tag{B.13}$$

We now compute the third term in Equation B.12, that is,

$$\mathbb{E}\left(\sum_{c=1}^{C}\sum_{j=0}^{d_g-1}|\sqrt{\beta_{gc}^{[l]}}(\rho_{gc}^{[l]})^j\varepsilon_{gc}^{[l]}(t-j)|^2\right)=\sum_{c=1}^{C}\sum_{j=0}^{d_g-1}\beta_{gc}^{[l]}(\rho_{gc}^{[l]})^{2j}(1-(\rho_{gc}^{[l]})^2) \tag{B.14}$$
$$=\sum_{c=1}^{C}\beta_{gc}^{[l]}\frac{1-(\rho_{gc}^{[l]})^{2d_g}}{1-(\rho_{gc}^{[l]})^2}(1-(\rho_{gc}^{[l]})^2)$$
$$=\sum_{c=1}^{C}\beta_{gc}^{[l]}(1-(\rho_{gc}^{[l]})^{2d_g}),$$

for the second equality we have used the expression of finite geometric sums since $(\rho_{gc}^{[l]})^2<1$ for all g and c. Next we compute the second term in Equation B.12, namely,

$$\mathbb{E}(\sum_{c=1}^{C}|(\rho_{gc}^{[l]})^{d_g}\tilde{g}_{gc}^{[l]}(t-d_g)|^2)=\sum_{c=1}^{C}(\rho_{gc}^{[l]})^{2d_g}\left(\beta_{gc}^{[l]}-\frac{(\beta_{gc}^{[l]})^2}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]}}\right). \tag{B.15}$$

the latter is satisfied due to the fact that the variance of $\tilde{g}_{gc}^{[l]}(t-d_g)$ is given by $\beta_{gc}^{[l]}-\frac{(\beta_{gc}^{[l]})^2}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]}}$ for all g and c.

We are left with the first term in Equation B.12, that is,

$$\sum_{k\neq g}^{N_g}\sum_{c=1}^{C}\mathbb{E}(|g_{kc}^{[l]}(t)|^2)=\sum_{k\neq g}^{N_g}\sum_{c=1}^{C}\mathbb{E}(|\sqrt{\beta_{kc}^{[l]}}h_{kc}^{[l]}(t)|^2)=\sum_{k\neq g}^{N_g}\sum_{c=1}^{C}\beta_{kc}^{[l]}, \tag{B.16}$$

Combining all four terms, that is, Equations B.13, B.14, B.15, B.16 and B.12, we obtain

$$\mathbb{E}\left[|u_{gl}^{\dagger}(t)I_3(t)|^2\right]=\sum_{k\neq g}^{N_g}\sum_{c=1}^{C}\beta_{kn}^{[l]}+\sum_{c=1}^{C}\beta_{gc}^{[l]}(1-(\rho_{gc}^{[l]})^{2d_g})+\sum_{c=1}^{C}(\rho_{gc}^{[l]})^{2d_g}(\beta_{gc}^{[l]}-\frac{(\beta_{gc}^{[l]})^2}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]}}) \tag{B.17}$$
$$=\sum_{k\neq g}^{N_g}\sum_{c=1}^{C}\beta_{kc}^{[l]}+\sum_{c=1}^{C}(\beta_{gc}^{[l]}-\rho_{gc}^{[l]^{2d_g}}\frac{\beta_{gc}^{[l]^2}}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]}})+\frac{1}{P_u}.$$

Substituting the results in Equations B.17 and B.11 in Equation B.8, we obtain

$$\overline{R}_{g,l}^{0}\geq\left(1-\frac{\tau}{T}\right)\log\left(1+\frac{(\rho_{gl}^{[l]})^{2d_g}|u_{gl}^{\dagger}(t)\hat{g}_{gl}^{[l]}(t-d_g)|^2}{F}\right), \tag{B.18}$$

with

$$F=|u_{gl}^{\dagger}(t)\hat{g}_{gl}^{[l]}(t-d_g)|^2\sum_{c\neq l}^{C}(\rho_{gc}^{[l]})^{2d_g}\frac{\beta_{gc}^{[l]^2}}{\beta_{gl}^{[l]^2}}+\sum_{k\neq g}^{N_g}\sum_{c=1}^{C}\beta_{kc}^{[l]}+\sum_{c=1}^{C}(\beta_{gc}^{[l]}-\rho_{gc}^{[l]^{2d_g}}\frac{\beta_{gc}^{[l]^2}}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]}})+\frac{1}{P_u}. \tag{B.19}$$

From Equation B.6 and B.19 we obtain

$$\overline{R}_{g,l} = \mathbb{E}(\overline{R}^0_{g,l}) = \mathbb{E}\left(\left(1-\frac{\tau}{T}\right)\log\left(1+\frac{(\rho^{[l]}_{il})^{2d_g}|u^{\dagger}_{il}(t)\hat{g}^{[l]}_{il}(t-d_g)|^2}{F}\right)\right) \tag{B.20}$$
$$= \mathbb{E}\left(\left(1-\frac{\tau}{T}\right)\log\left(1+\frac{(\rho^{[l]}_{gl})^{2d_g}}{G}\right)\right),$$

where

$$G = \frac{F}{|u^{\dagger}_{gl}(t)\hat{g}^{[l]}_{gl}(t-d_g)|^2}. \tag{B.21}$$

In order to compute the final expression of the bound on the average rate, it now suffices to compute the explicit expression of the right hand side (RHS) of Equation B.6. In order to do so, we apply Jensen's inequality to the RHS in Eq. B.6, that is,

$$\overline{R}_{g,l} \geq \left(1-\frac{\tau}{T}\right)\log\left(1+\frac{(\rho^{[l]}_{g,l})^{2d_g}}{\mathbb{E}(G)}\right), \tag{B.22}$$

with

$$\mathbb{E}(G) = \sum_{c\neq l}^{C}(\rho^{[l]}_{gc})^{2d_g}\frac{\beta^{[l]^2}_{gc}}{\beta^{[l]^2}_{gl}} + \mathbb{E}\left(\frac{1}{|u^{\dagger}_{gl}(t)\hat{g}^{[l]}_{gl}(t-d_g)|^2}\right) \tag{B.23}$$
$$\cdot\left(\sum_{k\neq g}^{N_g}\sum_{c=1}^{C}\beta^{[l]}_{kc} + \sum_{c=1}^{C}(\beta^{[l]}_{gc} - \rho^{[l]^{2d_g}}_{gc}\frac{\beta^{[l]^2}_{gc}}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta^{[l]}_{gb}}) + \frac{1}{P_u}\right).$$

Note that $\left|u^{\dagger}_{gl}(t)\hat{g}^{[l]}_{gl}(t-d_g)\right|^2$ has a Gamma distribution with parameters $(M, \frac{\beta^{[l]^2}_{gl}}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta^{[l]}_{gb}})$. Consequently, the mean value of $\frac{1}{\left|u^{\dagger}_{gl}(t)\hat{g}^{[l]}_{gl}(t-d_g)\right|^2}$ (that has an inverse Gamma distribution) is equal to $\frac{1}{(M-1)\times\frac{\beta^{[l]^2}_{gl}}{\frac{1}{P_p}+\sum_{b=1}^{C}\beta^{[l]}_{gb}}}$. Combining this together with the results in Equations B.20 and B.23 we obtain the desired lower bound on the average achievable spectral efficiency of user g, l, that is,

$$\overline{R}_{g,l} \geq \left(1-\frac{\tau}{T}\right)\log\left(1+\frac{(M-1)(\beta^{[l]}_{gl})^2(\rho^{[l]}_{gl})^{2d_g}}{(M-1)I^p_{gl}+I^n_{gl}}\right), \tag{B.24}$$

where I^p_{gl} and I^n_{gl} are given by

$$I^p_{gl} = \sum_{c\neq l}^{C} (\rho^{[l]}_{gc})^{2d_g} (\beta^{[l]}_{gc})^2, \text{ and} \tag{B.25}$$

$$I^n_{gl} = (\frac{1}{P_p} + \sum_{b=1}^{C} \beta^{[l]}_{gb}) \cdot \left(\sum_{k\neq g}^{N_g} \sum_{c=1}^{C} \beta^{[l]}_{kc} + \sum_{c=1}^{C} (\beta^{[l]}_{gc} - \rho^{[l]^{2d_g}}_{gc} \frac{\beta^{[l]^2}_{gc}}{\frac{1}{P_p} + \sum_{b=1}^{C} \beta^{[l]}_{gb}}) + \frac{1}{P_u} \right). \tag{B.26}$$

Summing the achievable spectral efficiency of all grouped users concludes the proof.

B.2 Proof of Theorem 7

The network serves N_g copilot groups, τ of which are scheduled for UL training. At the reception, each BS uses a Zero Forcing receiver that is based on the latest available CSI estimates.

BS l detects the signal of the users within the same cell by applying the following filter $U^{zf}_l(t) = (\hat{G}^{o\dagger}_l(t)\hat{G}^o_l(t))^{-1}\hat{G}^{o\dagger}_l(t),\ t \geq d_g$. Consequently, the detected signal of the users in cell l is given by the following

$$\begin{aligned} U^{zf}_l(t)\frac{Y^{[l]}_u(t)}{\sqrt{P_u}} &= U^{zf}_l(t)(\sum_{c=1}^{C} G^{[l]}_c(t) S_c + \frac{W_u(t)}{\sqrt{P_u}}) \\ &= U^{zf}_l(t)((C_l(t)\hat{G}^o_l(t) + C_l(t)\tilde{G}^o_l(t) + \epsilon_l(t))S_l + \sum_{c\neq l}^{C} \\ &\quad (C_c(t)\hat{G}^o_c(t) + C_c(t)\tilde{G}^o_c(t) + \epsilon_c(t))S_c + \frac{W_u(t)}{\sqrt{P_u}}) \\ &= C_l(t)\cdot I_{N_g} S_l + U^{zf}_l(t)((C_l(t)\tilde{G}^o_l(t) + \epsilon_l(t))S_l + \sum_{c\neq l}^{C} \\ &\quad (C_c(t)\hat{G}^o_c(t) + C_c(t)\tilde{G}^o_c(t) + \epsilon_c(t))S_c + \frac{W_u(t)}{\sqrt{P_u}}), \end{aligned} \tag{B.27}$$

where $C_c(t) \in \mathbb{C}^{M\times N_g}$ and $\epsilon_c(t) \in \mathbb{C}^{M\times N_g}$, $\forall c = 1,\ldots,C$, are given by

$$C_c(t) = diag(\rho^{[l]^{2d_k}}_{kc}, k = 1,\ldots,N_g),\ \forall c = 1,\ldots,C, \tag{B.28}$$

$$\epsilon_c(t) = [\sum_{j=0}^{d_1-1} (\rho^{[l]}_{1c})^j \sqrt{\beta^{[l]}_{1c}} \varepsilon^{[l]}_{1c}(t-j), \ldots, \sum_{j=0}^{d_{N_g}-1} (\rho^{[l]}_{N_g c})^j \sqrt{\beta^{[l]}_{N_g c}} \varepsilon^{[l]}_{N_g c}(t-j)],\ \forall c = 1,\ldots,C \tag{B.29}$$

We define Δ as

$$\Delta = U_l^{zf}(t)\big((C_l(t)\tilde{G}_l^o(t) + \epsilon_l(t))S_l + \sum_{c\neq l}^{C}(C_c(t)\hat{G}_c^o(t) + C_c(t)\tilde{G}_c^o(t) + \epsilon_c(t))S_c + \frac{W_u(t)}{\sqrt{P_u}}) \tag{B.30}$$

$$\big((C_l(t)\tilde{G}_l^o(t) + \epsilon_l(t))S_l + \sum_{c\neq l}^{C}(C_c(t)\hat{G}_c^o(t) + C_c(t)\tilde{G}_c^o(t) + \epsilon_c(t))S_c + \frac{W_u(t)}{\sqrt{P_u}})^{\dagger}U_l^{[zf]^{\dagger}}(t)$$

and

$$\Delta_1 = U_l^{zf}(t)(C_l(t)\tilde{G}_l^o(t) + \epsilon_l(t))(C_l(t)\tilde{G}_l^o(t) + \epsilon_l(t))^{\dagger}U_l^{[zf]^{\dagger}}(t) \tag{B.31}$$

$$\Delta_2 = U_l^{zf}(t)\sum_{c\neq l}^{C}(C_c(t)\hat{G}_c^o(t) + C_c(t)\tilde{G}_c^o(t) + \epsilon_c(t))(C_c(t)\hat{G}_c^o(t) + C_c(t)\tilde{G}_c^o(t) + \epsilon_c(t))^{\dagger}U_l^{[zf]^{\dagger}}(t) \tag{B.32}$$

$$\Delta_3 = U_l^{zf}(t)(\frac{W_u(t)}{\sqrt{P_u}})(\frac{W_u(t)}{\sqrt{P_u}})^{\dagger}U_l^{[zf]^{\dagger}}(t) \tag{B.33}$$

For user g in cell l, the SINR is given by $([C_l(t)]_{gg})^2[\Delta]_{gg}^{-1}$. Consequently, the average rate of this user is

$$\begin{aligned} \bar{R}_{g,l} &= \mathbb{E}\left(\left(1 - \frac{\tau}{T_s}\right)\log\left(1 + ([C_l(t)]_{gg})^2[\Delta]_{gg}^{-1}\right)\right) \qquad \text{(B.34)} \\ &\geq \left(1 - \frac{\tau}{T_s}\right)\log\left(1 + \mathbb{E}\left(([C_l(t)]_{gg})^2[\Delta]_{gg}^{-1}\right)\right), \end{aligned}$$

where the last inequality is obtained using the convexity of $\log(1 + x^{-1})$. We, now, compute each component of $\mathbb{E}\left([\Delta]_{gg}\right)$.

$[\Delta_1]_{gg}$ is the intra-cell interference due to the MMSE estimation error and the impact of channel aging. Since $\hat{G}_l^o(t)$, $\tilde{G}_l^o(t)$ and $\epsilon_l(t)$ are mutually independent, we obtain

$$\begin{aligned} \mathbb{E}\left([\Delta_1]\right) &= \mathbb{E}\left(U_l^{zf}(t)(C_l(t)\tilde{G}_l^o(t) + \epsilon_l(t))(C_l(t)\tilde{G}_l^o(t) + \epsilon_l(t))^{\dagger}U_l^{[zf]^{\dagger}}(t)\right) \qquad \text{(B.35)} \\ &= \mathbb{E}\left(U_l^{zf}(t)(C_l^2(t)\tilde{G}_l^o(t)\tilde{G}_l^{o^{\dagger}}(t) + \epsilon_l(t)\epsilon_l^{\dagger}(t))U_l^{[zf]^{\dagger}}(t)\right) \\ &= \mathbb{E}\left(U_l^{zf}(t)(C_l^2(t)\tilde{G}_l^o(t)\tilde{G}_l^{o^{\dagger}}(t) + \epsilon_l(t)\epsilon_l^{\dagger}(t))U_l^{[zf]^{\dagger}}(t)\right) \\ &= \mathbb{E}\left(U_l^{zf}(t)U_l^{[zf]^{\dagger}}(t)\right)\cdot\left(C_l^2(t)\mathbb{E}\left(\tilde{G}_l^o(t)\tilde{G}_l^{o^{\dagger}}(t)\right) + \mathbb{E}\left(\epsilon_l(t)\epsilon_l^{\dagger}(t)\right)\right), \end{aligned}$$

In addition, we have $[\tilde{G}_l^o(t)]_k \sim \mathcal{CN}\left(0, \left(\beta_{kl}^{[l]} - \frac{{\beta_{kl}^{[l]}}^2}{\frac{1}{P_p} + \sum_{b,b\neq l}^{C}\beta_{kl}^{[l]}}\right)I_M\right)$, and $[\epsilon_l(t)]_k$ $\sim \mathcal{CN}\left(0, \beta_{kl}^{[l]}(1 - (\rho_{kl}^{[l]})^{2d_k})I_M\right)$. Consequently, we obtain

$$\begin{aligned}\mathbb{E}\left([\Delta_1]_{gg}\right) = \mathbb{E}\left([\hat{G}_l^{o^\dagger}(t)\hat{G}_l^o(t)]^{-1}_{gg}\right)\cdot \\ \left(\sum_{k,k\neq g}^{N_g} \rho_{kl}^{[l]^{2d_k}}\left(\beta_{kl}^{[l]} - \frac{\beta_{kl}^{[l]^2}}{\frac{1}{P_p}+\sum_{b,b\neq l}^{C}\beta_{kl}^{[l]}}\right) + \sum_{k,k\neq g}^{N_g}\beta_{kl}^{[l]}(1-(\rho_{kl}^{[l]})^{2d_k})\right),\end{aligned} \tag{B.36}$$

Note that $[\hat{G}_l^{o^\dagger}(t)\hat{G}_l^o(t)]^{-1}_{gg}$ has a Gamma distribution with $M - N_g + 1$ degrees of freedom. Consequently, $[\hat{G}_l^{o^\dagger}(t)\hat{G}_l^o(t)]^{-1}_{gg}$ has an inverse Gamma distribution with mean $\frac{(\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]})}{\beta_{gl}^{[l]^2}(M-N_g)}$ and we obtain the following

$$\mathbb{E}\left([\Delta_1]_{gg}\right) = \frac{(\frac{1}{P_p}+\sum_{b=1}^{C}\beta_{gb}^{[l]})}{\beta_{gl}^{[l]^2}(M-N_g)}\cdot\sum_{k,k\neq g}^{N_g}\left(\beta_{kl}^{[l]} - \rho_{kl}^{[l]^{2d_k}}\frac{\beta_{kl}^{[l]^2}}{\frac{1}{P_p}+\sum_{b,b\neq l}^{C}\beta_{kl}^{[l]}}\right), \tag{B.37}$$

$[\Delta_2]_{gg}$ is the inter-cell interference. It includes the impact of coherent interference (pilot contamination). Since $C_c(t)\hat{G}_c^o(t)$, $C_b(t)\tilde{G}_b^o(t)$ and $\epsilon_u(t)$ are mutually independent $\forall u, c, b$, we have

$$\begin{aligned}\mathbb{E}\left([\Delta_2]\right) = \sum_{c\neq l}^{C}\mathbb{E}\left(U_l^{zf}(t)\left(C_c^2(t)\hat{G}_c^o(t)\hat{G}_c^{o^\dagger}(t)\right)U_l^{[zf]^\dagger}(t)\right) \\ +\,\mathbb{E}\left(U_l^{zf}(t)\left(C_c^2(t)\tilde{G}_c^o(t)\tilde{G}_c^{o^\dagger}(t)\right)U_l^{[zf]^\dagger}(t)\right) + \mathbb{E}\left(U_l^{zf}(t)\left(\epsilon_c(t)\epsilon_c^\dagger(t)\right)U_l^{[zf]^\dagger}(t)\right)\end{aligned} \tag{B.38}$$

Owing to pilot reuse, $\hat{G}_l^o(t)$ and $C_c(t)\hat{G}_c^o(t)$ are correlated for $l \neq c$. Since, for each copilot group g, we have $\hat{g}_{gc}^{[l]}(t-d_g) = \hat{g}_{gl}^{[l]}(t-d_g)\frac{\beta_{gc}^{[l]}}{\beta_{gl}^{[l]}}$, we obtain $\hat{G}_l^{o^\dagger}(t) = N_c^l\hat{G}_c^{o^\dagger}(t)$ where $N_c^l \in \mathbb{C}^{N_g\times N_g}$, $[N_c^l] = \text{diag}(\frac{\beta_{kc}^{[l]}}{\beta_{kl}^{[l]}},\ k = 1,\ldots,N_g)$. Consequently,

$$\sum_{c\neq l}^{C}\mathbb{E}\left(U_l^{zf}(t)\left(C_c^2(t)\hat{G}_c^o(t)\hat{G}_c^{o^\dagger}(t)\right)U_l^{[zf]^\dagger}(t)\right) = \sum_{c\neq l}^{C}C_c^2(t)N_c^{[l]^2} \tag{B.39}$$

Since $U_l^{zf}(t)$, $\tilde{G}_c^o(t)$ and $\epsilon_c(t)$ are mutually independent we have

$$\begin{aligned}\sum_{c\neq l}^{C}\mathbb{E}\left(U_l^{zf}(t)\left(C_c^2(t)\tilde{G}_c^o(t)\tilde{G}_c^{o^\dagger}(t)\right)U_l^{[zf]^\dagger}(t)\right) + \mathbb{E}\left(U_l^{zf}(t)\left(\epsilon_c(t)\epsilon_c^\dagger(t)\right)U_l^{[zf]^\dagger}(t)\right) \\ = \sum_{c\neq l}^{C}\mathbb{E}\left(U_l^{zf}(t)U_l^{[zf]^\dagger}(t)\right)\cdot\left(\mathbb{E}\left(\epsilon_c(t)\epsilon_c^\dagger(t)\right) + \mathbb{E}\left(C_c^2(t)\tilde{G}_c^o(t)\tilde{G}_c^{o^\dagger}(t)\right)\right)\end{aligned} \tag{B.40}$$

Combining the results of the two previous equations, we finally obtain $\mathbb{E}\left([\Delta_2]_{gg}\right)$ as

$$\mathbb{E}\left([\Delta_2]_{gg}\right) = \sum_{c\neq l}^{C} \left(\frac{(\frac{1}{P_p} + \sum_{b=1}^{C} \beta_{gb}^{[l]})}{\beta_{gl}^{[l]^2}(M-N_g)} (\sum_{k}^{N_g} (\beta_{kc}^{[l]} - \rho_{kc}^{[l]^{2d_k}} \frac{\beta_{kc}^{[l]^2}}{\frac{1}{P_p} + \sum_{b,b\neq c}^{C} \beta_{kb}^{[l]}})) + (\rho_{gc}^{[l]})^{2d_g} \frac{\beta_{gc}^{[l]^2}}{\beta_{gl}^{[l]^2}} \right) \tag{B.41}$$

$[\Delta_3]_{gg}$ denotes the impact of UL noise. Using the same aforementioned reasoning, we obtain

$$\mathbb{E}\left([\Delta_3]_{gg}\right) = \mathbb{E}\left(U_l^{zf}(t) \frac{W_u(t)}{\sqrt{P_u}} (\frac{W_u(t)}{\sqrt{P_u}})^{\dagger} U_l^{[zf]^{\dagger}}(t) \right) = \frac{1}{P_u} \mathbb{E}\left(U_l^{zf}(t) U_l^{[zf]^{\dagger}}(t) \right) \tag{B.42}$$
$$= \frac{1}{P_u} \frac{(\frac{1}{P_p} + \sum_{b=1}^{C} \beta_{gb}^{[l]})}{\beta_{gl}^{[l]^2}(M-N_g)}$$

Combining the obtained results for $\mathbb{E}\left([\Delta_1]_{gg}\right)$, $\mathbb{E}\left([\Delta_2]_{gg}\right)$ and $\mathbb{E}\left([\Delta_3]_{gg}\right)$, we obtain

$$([C_l(t)]_{gg})^2 [\Delta]_{gg}^{-1} = \tag{B.43}$$
$$\frac{(M-N_g)(\rho_{gl}^{[l]})^{2d_g} \beta_{gl}^{[l]^2}}{(M-N_g) \sum_{c\neq l}^{C} (\rho_{gc}^{[l]})^{2d_g} \beta_{gc}^{[l]^2} + \left(\frac{1}{P_u} + \sum_{c}^{C} \sum_{k}^{N_g} \beta_{kc}^{[l]} - \rho_{kc}^{[l]^{2d_k}} \frac{\beta_{kc}^{[l]^2}}{\frac{1}{P_p} + \sum_{b,b\neq c}^{C} \beta_{kb}^{[l]}} \right) (\frac{1}{P_p} + \sum_{b=1}^{C} \beta_{gb}^{[l]})}$$

Which finishes the proof.

B.3 Proof of Theorem 8

In order to prove Theorem 8, we consider the asymptotic regime where the number of BS antennas M grows very large. In this case the lower bound on the spectral efficiency of each user g, l converges to the following limit:

$$\left(1 - \frac{\tau}{T_s}\right) \log \left(1 + \frac{\beta_{gl}^{[l]^2} \rho_{gl}^{[l]^{2d_g}}}{\sum_{b\neq l}^{C} \rho_{gb}^{[l]^{2d_g}} \beta_{gb}^{[l]^2}} \right). \tag{B.44}$$

The proposed framework is compared with a reference massive MIMO system where, all scheduled users are required to perform UL training. In the asymptotic regime, the lower bound on the achievable spectral efficiency of each user g, l in the reference system converges to the following limit:

$$\left(1 - \frac{N_g}{T_s}\right) \log \left(1 + \frac{\beta_{gl}^{[l]^2}}{\sum_{b\neq l}^{C} \beta_{gb}^{[l]^2}} \right). \tag{B.45}$$

The aim here, is to improve the achievable spectral efficiency of each scheduled users. Consequently, the spectral efficiency of each user in the two considered systems should verify, $\forall\ g = 1...N_g, l = 1...C$:

$$\left(1 - \frac{\tau}{T_s}\right) \log\left(1 + \frac{\beta_{gl}^{[l]^2} \rho_{gl}^{[l]^{2d_g}}}{\sum_{b\neq l}^{C} \rho_{gb}^{[l]^{2d_g}} \beta_{gb}^{[l]^2}}\right) \geq \left(1 - \frac{N_g}{T_s}\right) \log\left(1 + \frac{\beta_{gl}^{[l]^2}}{\sum_{b\neq l}^{C} \beta_{gb}^{[l]^2}}\right). \tag{B.46}$$

(41) is equivalent to the following condition:

$$\frac{\beta_{gl}^{[l]^2} \rho_{gl}^{[l]^{2d_g}}}{\sum_{b\neq l}^{C} \rho_{gb}^{[l]^{2d_g}} \beta_{gb}^{[l]^2}} \geq \left(1 + \frac{\beta_{gl}^{[l]^2}}{\sum_{b\neq l}^{C} \beta_{gb}^{[l]^2}}\right)^{\frac{T_s - N_g}{T_s - \tau}} - 1. \tag{B.47}$$

We consider the extreme case where $\rho_{gl}^2 = \bar{\rho}_g^{[min]^2}$ and $\forall b \neq l, \rho_{gb}^2 = \bar{\rho}_g^{[max]^2}$. Here $\bar{\rho}_g^{[min]}$ and $\bar{\rho}_g^{[max]}$ denote respectively the minimum and maximum channel autocorrelation coefficients in group g. This means that we assume the worst case scenario for each user where, its coherence time is always lower than its fellow copilot users. Finally, by considering $SINR_{g,l}^{[\infty]} = \frac{\beta_{gl}^{[l]^2}}{\sum_{b\neq l}^{C} \beta_{gb}^{[l]^2}}$, we finish the proof.

B.4 Proof of Theorem 9

In order to prove theorem 9, we start by demonstrating that the objective function of problem (4.38), is submodular. In order to do so, we note that the sum of submodular functions is submodular. Consequently, it is enough to prove the submodularity of f_g for a given copilot group g, where f_g is given by

$$f_g(\vec{a}(t_0), \ldots, \vec{a}(t_{H-1}), x, \vec{u}) = \sum_{t=t_0}^{t_{H-1}} \sum_{l=1}^{C} \big(1 - \frac{1}{T_s} \sum_{i=1}^{N_G} a_i(t)\big) \log\big(1 + \text{SINR}_{gl}^{MRC}(\vec{d}(t), x, \vec{u})\big), \tag{B.48}$$

We consider two sets of action vectors, $\{\vec{a}(t) \in A, , t = t_0, \ldots, t_{H-1}\}$ and $\{\vec{a}'(t) \in A, t = t_0, \ldots, t_{H-1}\}$ such that, $\forall t = t_0, \ldots, t_{H-1},\ \sum_{i=1}^{N_G} a_i(t) \leq \sum_{i=1}^{N_G} a_i'(t)$, and $\forall i = 1, \ldots, N_G,\ a_i(h) = 1 \Rightarrow a_i'(h) = 1$.

These two sets of action vectors result, respectively, in two sets of delay vectors $\{\vec{d}(t), t = t_0, \ldots, t_{H-1}\}$ and $\{\vec{d}'(t), t = t_0, \ldots, t_{H-1}\}$ that can be obtained from $\vec{a}(t)$ and $\vec{a}'(t)$ according to (4.37).

In order to prove the submodularity of f_g, we need to prove that, for a given h and j such that $a_j(h) = a_j'(h) = 0$, the marginal values of setting $a_j(h) = 1$ is higher than that of $a_j'(h) = 1$

,i.e

$$f_g(\vec{a}(t_0), \dots, \vec{a}(h) \oplus a_j(h), \dots, \vec{a}(t_{H-1}), x, \vec{u}) - f_g(\vec{a}(t_0), \dots, \vec{a}(h), \dots, \vec{a}(t_{H-1}), x, \vec{u}) \geq \tag{B.49}$$
$$f_g(\vec{a}'(t_0), \dots, \vec{a}'(h) \oplus a'_j(h), \vec{a}'(t_{H-1}), x, \vec{u}) - f_g(\vec{a}'(t_0), \dots, \vec{a}'(h), \dots, \vec{a}'(t_{H-1}), x, \vec{u}).$$

We will distinguish between two cases, $j = g$ and $j \neq g$. For the first case, where $j = g$, the difference between the two marginal values is given by:

$$\begin{aligned}\Lambda - \Lambda' = &\sum_{l=1}^{C} \log\big(1 + \text{SINR}_{gl}^{MRC}(0, x, \vec{u})\big)\big(\frac{\sum_{i=1}^{N_G} a'_i(h) - \sum_{i=1}^{N_G} a_i(h)}{T_s}\big) + \big(1 - \frac{\sum_{i=1}^{N_G} a'_i(h)}{T_s}\big) \\ &\log\big(1 + \text{SINR}_{gl}^{MRC}(d'_g(h), x, \vec{u})\big) - \big(1 - \frac{\sum_{i=1}^{N_G} a_i(h)}{T_s}\big) \log\big(1 + \text{SINR}_{gl}^{MRC}(d_g(h), x, \vec{u})\big) \\ &+ \sum_{t=h+1}^{t_{H-1}} \sum_{l=1}^{C} \big(1 - \frac{\sum_{i=1}^{N_G} a_i(t)}{T_s}\big) \log\big(\frac{1 + \text{SINR}_{gl}^{MRC}(d_g(t) - 1, x, \vec{u})}{1 + \text{SINR}_{gl}^{MRC}(d_g(t), x, \vec{u})}\big) - \big(1 - \frac{\sum_{i=1}^{N_G} a'_i(t)}{T_s}\big) \\ &\log\big(\frac{1 + \text{SINR}_{gl}^{MRC}(d'_g(t) - 1, x, \vec{u})}{1 + \text{SINR}_{gl}^{MRC}(d'_g(t), x, \vec{u})}\big)\end{aligned} \tag{B.50}$$

The difference in marginal values is positive as $\log\big(\frac{1+\text{SINR}_{gl}^{MRC}(d_g(t)-1,x,\vec{u})}{1+\text{SINR}_{gl}^{MRC}(d_g(t),x,\vec{u})}\big)$ is decreasing as a function of $d_g(t)$. We, now, consider the case where $j \neq g$. In this case, we have

$$\Lambda - \Lambda' = \sum_{l=1}^{C} \frac{1}{T_s} \log\big(\frac{1 + \text{SINR}_{gl}^{MRC}(d'_g(h), x, \vec{u})}{1 + \text{SINR}_{gl}^{MRC}(d_g(h), x, \vec{u})}\big) \tag{B.51}$$

From the definition of $\vec{a}(h)$ and $\vec{a}'(h)$, we have $\sum_{i=1}^{N_G} a_i(h) \leq \sum_{i=1}^{N_G} a'_i(h)$.
Hence $\text{SINR}_{gl}^{MRC}(d'_g(h), x, \vec{u}) \geq \text{SINR}_{gl}^{MRC}(d_g(h), x, \vec{u})$ and the difference in marginal values is also positive, in this case. Consequently, f_g is submodular.
Concerning the matroid constraints, let us consider the ground set $G = \{v_{1t_1}, \dots, v_{N_G t_1}, \dots, v_{1t_{H-1}}, \dots, v_{N_G t_H}$ where each element v_{gt} represents the scheduling of copilot group g for training at slot t. It is clear that the constraints of (4.38) form a partition matroid on G [206]. Consequently, problem (4.38) is a maximization of a submodular function subject to matroid constraints.

Appendix C

User-centric 5G networks: Energy Efficiency under popularity based Clustering in cache enabled SCN

C.1 Proof of Lemma 10

We derive the expression of the average consumed power in the network with cache enabled SBSs. The average total power ρ^c_{total} is given by:

$$\rho^c_{total} = \mathbb{E}\left\{\rho_I\right\} + \mathbb{E}\left\{\rho_T\right\} + \mathbb{E}\left\{\rho_f\right\}, \tag{C.1}$$

where $\mathbb{E}\left\{\rho_I\right\} = \rho\lambda_s\pi R_n^2$ and $\mathbb{E}\left\{\rho_f\right\} = \lambda_s\pi R_n^2\left(\rho_{hd}\mathbb{P}\left\{hit\right\} + \rho_{bh}\left(1 - \mathbb{P}\left\{hit\right\}\right)\right)$. Taking into account the considered system model, the average transmit power used by a given user from cluster k, $\mathbb{E}\left\{\rho_T^{[k]}\right\}$ can be written as follows:

$$\mathbb{E}\left\{\rho_T^{[k]}\right\} = \mathbb{E}\left\{\rho_k\right\}(1 - \sum_{j\neq k}\sum_{i\in\Delta_s} p_{iu}(1 - e^{-\lambda_{sj}\pi R^2})) + \sum_{j\neq k}\sum_{i\in\Delta_s} p_{iu}(1 - e^{-\lambda_{sj}\pi R^2})\mathbb{E}\left\{\rho_j\right\}. \tag{C.2}$$

After averaging over all users in the network, the average consumed transmit power is given by:

$$\mathbb{E}\left\{\rho_T\right\} = \frac{\lambda_s\pi R_n^2}{U}\sum_{k=1}^{N_c}\sum_{u\in\Upsilon_k}\left(\mathbb{E}\left\{\rho_k\right\} + \sum_{j\neq k}\sum_{i\in\Delta_s} p_{iu}(1 - e^{-\lambda_{sj}\pi R^2})(\mathbb{E}\left\{\rho_j\right\} - \mathbb{E}\left\{\rho_k\right\})\right). \tag{C.3}$$

We need then to compute the average power used by the users to communicate with the nearest SBS from any given cluster k , $k = 1..N_c$.
According to the PPP assumption for the location of the SBSs, the distance from a user to its nearest SBS from cluster k, denoted by r_k, has the following *pdf* [177]:

$$f_{r_k}(r) = 2\pi\lambda_{sk} r e^{-\lambda_{sk}\pi r^2}. \tag{C.4}$$

The transmit power used by the user in this case is given by $\rho_k = \rho_0 r_k^{\alpha}$. Then:

$$\mathbb{E}\left[\rho_k\right] = \int_0^R 2\pi\lambda_{sk} r^{\alpha+1} exp(-\lambda_{sk}\pi r^2) dr = \frac{\rho_0\gamma(\frac{\alpha}{2}+1, \pi\lambda_{sk}R^2)}{(\lambda_{sk}\pi)^{\frac{\alpha}{2}}}. \tag{C.5}$$

Following the same calculus for $\mathbb{E}\left[\rho_k\right],\ k = 1..N_c$, we obtain the final expression of the average consumed power in the network:

$$\begin{aligned}\rho_{total}^{c} = \lambda_s \pi R_n^2 (\rho_{hd}\mathbb{P}\left\{hit\right\} + \rho_{bh}\left(1 - \mathbb{P}\left\{hit\right\}\right) + \rho)\frac{\lambda_s \pi R_n^2}{U}\sum_{k=1}^{N_c}\sum_{u\in\Upsilon_k}(\frac{\rho_0\gamma(\frac{\alpha}{2}+1, \pi\lambda_{sk}R^2)}{(\lambda_{sk}\pi)^{\frac{\alpha}{2}}} \\ + \sum_{j\neq k}\sum_{i\in\Delta_s} p_{iu}(1 - e^{-\lambda_{sj}\pi R^2})(\frac{\rho_0\gamma(\frac{\alpha}{2}+1, \pi\lambda_{sj}R^2)}{(\lambda_{sj}\pi)^{\frac{\alpha}{2}}} - \frac{\rho_0\gamma(\frac{\alpha}{2}+1, \pi\lambda_{sk}R^2)}{(\lambda_{sk}\pi)^{\frac{\alpha}{2}}})).\end{aligned} \tag{C.6}$$

C.2 Proof of Lemma 11

We derive the achievable coverage probability when using channel inversion power control:

$$\begin{aligned}\mathbb{P}\left\{\text{SINR} \geq \theta\right\} &= \mathbb{E}\left[\mathbb{P}(\|h_u\|^2 \geq (\frac{\sigma^2 + \sum_{k=1}^{N_c} I_k}{\rho_0})\theta) | I_k \forall k\right] \\ &= \mathbb{E}\left[exp(-\frac{\theta}{\rho_0}(\sigma^2 + \sum_{k=1}^{N_c} I_k)) | I_k \forall k\right] = exp(-\frac{\theta}{\rho_0}\sigma^2)\prod_{k=1}^{N_c}\mathcal{L}_{I_k}(\frac{\theta}{\rho_0}).\end{aligned} \tag{C.7}$$

We use the fact that $\|h_u\|^2$ is exponentially distributed and $\mathcal{L}_{I_k}(s)$ is the Laplace transform of I_k at s. To prove Lemma 5.21, we need to compute the Laplace transform of $I_k, \forall k$. The interfering base stations constitute multiple PPP processes $\phi_{sk}, k = 1...N_c$, each associated with a given cluster. The Laplace transform of I_k for a given k is obtained as:

$$\begin{aligned}\mathcal{L}_{I_k}(\frac{\theta}{\rho_0}) = \mathbb{E}\left[exp\left(-\sum_{i\in\phi_{sk}} \rho_{ik}\|h_{ui}\|^2 r_{ui}^{-\alpha}\right)\right] = exp(-2\pi\lambda_{sk}\int_0^\infty \left(1 - \mathbb{E}\left[e^{-\theta\|h\|^2\rho_k r^{-\alpha}}\right]\right) rdr) \\ = exp\left(-\pi\lambda_{sk}(\frac{\theta}{\rho_0})^{\frac{2}{\alpha}}\mathbb{E}\left[\rho_k^{\frac{2}{\alpha}}\right]\Gamma(1+\frac{2}{\alpha})\Gamma(1-\frac{2}{\alpha})\right).\end{aligned} \tag{C.8}$$

$\mathbb{E}\left[\rho_k^{\frac{2}{\alpha}}\right]$ depends on the density of the small cells caching files from cluster k. $\mathbb{E}\left[\rho_k^{\frac{2}{\alpha}}\right]$ can be deduced from C.1 as; $\mathbb{E}\left[\rho_k^{\frac{2}{\alpha}}\right] = \frac{\rho_0^{\frac{2}{\alpha}}\gamma(2,\pi\lambda_{sk}R^2)}{\lambda_{sk}\pi}$. Based on Slivnyak's Theorem for Poisson Point Processes [177], the obtained Expression is valid for any user within the network.

C.3 Proof of Theorem 12

We start by showing that the objective function is quasi-concave. Given the expression of Σ as a function of the density vector Λ_s, it is difficult to prove its quasi-concavity by using its gradient or Hessian matrix.

However, using the fact that a composition with an affine function preserves quasi-concavity [200], this proof can be considerably simplified. To prove the quasi-concavity of Σ as a function of the density vector Λ_s, we consider an affine function $f(t)$ given by:

$$f(t) = tZ + \Lambda_s^0, \tag{C.9}$$

where $\Lambda_s^0 \in \mathbb{R}^{N_c\times 1}$ such that $\sum_{k=1}^{N_c} \lambda_{sk}^0 \leq \lambda_{s_{max}}$, $Z \in \mathbb{R}^{N_c\times 1}$ and $t \in \mathbb{R}$. Since a composition with an affine function preserves quasi-concavity, it is sufficient to prove the quasi-concavity of $\Sigma(tZ+\Lambda_s^0)$ with respect to t in order to show the quasi-concavity of Σ with respect to Λ_s. The objective $\Sigma(tZ+\Lambda_s^0)$ can be written as a product of two nonnegative functions: $U(tZ+\Lambda_s^0) = \sum_{k=1}^{N_c}(tz_k+\lambda_{sk}^0)\pi R_n^2 \log(1+\theta)\,\mathbb{P}\left\{\text{SINR} \geq \theta\right\}$ and $V(tZ+\Lambda_s^0) = \frac{1}{\rho_{total}^c}$.
We start by computing the derivatives of $U(tZ+\Lambda_s^0)$ and $V(tZ+\Lambda_s^0)$ with respect to t:

$$U' = \frac{\mathrm{dU(tZ+\Lambda_s^0)}}{\mathrm{dt}} = (\sum_{k=1}^{N_c} z_k)\pi R_n^2 \log(1+\theta) exp(-\frac{\theta}{\rho_0}\sigma^2)\prod_{k=1}^{N_c}\mathcal{L}_{I_{ku}}(\frac{\theta}{\rho_0}) \tag{C.10}$$
$$\left(1-\sum_{k=1}^{N_c}\Gamma(1+\frac{2}{\alpha})\Gamma(1-\frac{2}{\alpha})\theta^{\frac{2}{\alpha}}\pi z_k R^2 e^{-(tz_k+\lambda_{sk}^0)\pi R^2}\right).$$

Then $\frac{\mathrm{dU(tZ+\Lambda_s^0)}}{\mathrm{dt}} > 0$. We do the same to $V(tZ+\Lambda_s^0) = \frac{1}{\rho_{total}^c}$. We have:

$$V' = \frac{\mathrm{dV(tZ+\Lambda_s^0)}}{\mathrm{dt}} = \frac{-\chi}{P_{total}^{c2}}, \tag{C.11}$$

where χ is given by:

$$\chi = (\sum_{k=1}^{N_c} z_k)\pi R_n^2((\rho_{hd}-\rho_{bh})\mathbb{P}\left\{hit\right\}+\rho+\rho_{bh}) + \frac{\pi R_n^2}{U}((\sum_{k=1}^{N_c} z_k)\chi_1 + (\sum_{k=1}^{N_c} tz_k+\lambda_{sk}^0)\chi_2) \tag{C.12}$$
$$+(\sum_{k=1}^{N_c} tz_k+\lambda_{sk}^0)\pi R_n^2\frac{1}{U}\sum_{u=1}^{U}\sum_{k=1}^{N_c}\left(\sum_{i\in\Delta_k} p_{iu}\right) z_k\pi R^2 e^{-(tz_k+\lambda_{sk}^0)\pi R^2}.$$

Here χ_1 and χ_2 are respectively given by:

$$\chi_1 = \sum_{k=1}^{N_c} \sum_{u\in\Upsilon_k} \Big(\frac{\rho_0 \gamma(\frac{\alpha}{2}+1, \pi(tz_k+\lambda_{sk}^0)R^2)}{((tz_k+\lambda_{sk}^0)\pi)^{\frac{\alpha}{2}}} + \sum_{j\neq k} \sum_{i\in\Delta_s} p_{iu}(1-e^{-(tz_j+\lambda_{sj}^0)\pi R^2}) \tag{C.13}$$
$$\rho_0\Big(\frac{\gamma(\frac{\alpha}{2}+1, \pi(tz_k+\lambda_{sj}^0)R^2)}{((tz_j+\lambda_{sj}^0)\pi)^{\frac{\alpha}{2}}} - \frac{\gamma(\frac{\alpha}{2}+1, \pi(tz_k+\lambda_{sk}^0)R^2)}{((tz_k+\lambda_{sk}^0)\pi)^{\frac{\alpha}{2}}} \Big)\Big),$$

$$\chi_2 = \rho_0 \sum_{k=1}^{N_c} \sum_{u\in\Upsilon_k} \Big((z_k \pi R^\alpha e^{-(tz_k+\lambda_{sk}^0)\pi R^2} - \frac{\gamma(\frac{\alpha}{2}+1, \pi(tz_k+\lambda_{sk}^0)R^2)\frac{\alpha z_k \pi}{2}((tz_k+\lambda_{sk}^0)\pi)^{\frac{\alpha}{2}-1}}{((tz_k+\lambda_{sk}^0)\pi)^\alpha} \Big) \tag{C.14}$$
$$+ \sum_{j\neq k} \sum_{i\in\Delta_s} p_{iu}(1-e^{-(tz_j+\lambda_{sj}^0)\pi R^2}) \Big(\frac{\gamma(\frac{\alpha}{2}+1, \pi(tz_k+\lambda_{sk}^0)R^2)\frac{\alpha z_k \pi}{2}((tz_k+\lambda_{sk}^0)\pi)^{\frac{\alpha}{2}-1}}{((tz_k+\lambda_{sk}^0)\pi)^\alpha}$$
$$+ \pi(z_j R^\alpha e^{-(tz_j+\lambda_{sj}^0)\pi R^2} - \frac{\gamma(\frac{\alpha}{2}+1, \pi(tz_j+\lambda_{sj}^0)R^2)\alpha z_j((tz_j+\lambda_{sj}^0)\pi)^{\frac{\alpha}{2}-1}}{2((tz_j+\lambda_{sj}^0)\pi)^\alpha} \Big)$$
$$- z_k R^\alpha e^{-(tz_k+\lambda_{sk}^0)\pi R^2}) + \sum_{j\neq k} \sum_{i\in\Delta_s} p_{iu} z_j \pi R^2 e^{-(tz_j+\lambda_{sj}^0)\pi R^2} \Big(\frac{\gamma(\frac{\alpha}{2}+1, \pi(tz_j+\lambda_{sj}^0)R^2)}{((tz_j+\lambda_{sj}^0)\pi)^{\frac{\alpha}{2}}}$$
$$- \frac{\gamma(\frac{\alpha}{2}+1, \pi(tz_k+\lambda_{sk}^0)R^2)}{((tz_k+\lambda_{sk}^0)\pi)^{\frac{\alpha}{2}}} \Big)\Big).$$

Consequently $\frac{\mathrm{dV(tZ+\Lambda_s^0)}}{\mathrm{dt}} < 0$. In what follows we distinguish two cases depending on the existence of a point t^* such that $\frac{\mathrm{d\Sigma(tZ+\Lambda_s^0)}}{\mathrm{dt}}_{t=t^*} = 0$. If $\exists t^*$ such that $\frac{\mathrm{d\Sigma(tZ+\Lambda_s^0)}}{\mathrm{dt}}_{t=t^*} = 0$ then:

$$\frac{\mathrm{d\Sigma(tZ+\Lambda_s^0)}}{\mathrm{dt}}_{t=t^*} = 0 \Leftrightarrow U'V + V'U = 0 \Leftrightarrow \frac{-U'V}{V'U} = 1. \tag{C.15}$$

We compute the derivative of $L(t) = \frac{-U'V}{V'U}$ with respect to t. The expression of the derivative is omitted here for brevity. We find that $L'(t) > 0$. Since $L(t)$ is a strictly increasing function then, according the Theorem of intermediate value, if $\exists t^*$ such that $L(t^*) = 1$ then this point is unique. Finally, depending on the existence of t^*, we have two cases:

- If $\exists\ t^*$ such that $L(t^*) = 1$ then this point is unique and $\Sigma(f(t))$ is increasing for $t < t^*$ and decreasing for $t > t^*$.
- If, on the other hand, t^* does not exists, then $\Sigma(f(t))$ is a strictly monotone function.

This proves that $\Sigma(f(t))$ is a quasi-concave function of t. Since composition with an affine function preserves quasi-concavity, we can deduce that $\Sigma(\Lambda_s)$ is a quasi-concave function of Λ_s and that, if $\exists \Lambda_s^*$ such that $\nabla\Sigma(\Lambda_s^*) = 0$ then this vector is unique .

In the second step of the proof, we need to show that the constraint $C(\Lambda_s) = \rho^{c}_{total} - \rho^{nc}_{total}$ is also quasi concave. This is done in a similar way as in the first step by considering $C(f(t))$. After computing the derivative of $C(f(t))$ with respect to t, we find that: $\frac{\mathrm{dC(f(t))}}{\mathrm{dt}} < 0$.
Then the first constraint is quasi concave. Using the same method, it is trivial to show that the second constraint is also quasi-concave. In order to finish the proof, we need to show that the optimal solution can be found with zero duality gap. This will be done using results on quasi-concave programming from [195]. Since $\frac{\mathrm{d}(\rho^{\mathrm{c}}_{\mathrm{total}} - \rho^{\mathrm{nc}}_{\mathrm{total}})}{\mathrm{d}\lambda_{\mathrm{sk}}} \neq 0, \forall k = 1...N_c$ then, according to the Necessity Theorem in [195], any solution of the optimization problem 5.24 satisfies the KKT conditions.
We, now, distinguish between two case:

- If $\exists\ \Lambda^*_s$ such that $\nabla\Sigma(\Lambda_s)_{\Lambda_s = \Lambda^*_s} = 0$ and Λ^*_s satisfies the constraints then, Λ^*_s is unique and it is a global optimum of (23). The uniqueness of Λ^*_s, if it exists, was shown in the first step of the proof.
- If $\nabla\Sigma(\Lambda_s) \geq 0$, $\forall\Lambda_s$ such that $\Lambda^{\dagger}_s \mathbf{1} \leq \lambda_{s_{max}}$, the sufficiency Theorem in [195] is verified. Consequently, by combining the necessity and sufficiency results, the optimal SBS density vector can be derived using KKT.

C.4 Proof of Theorem 13

First we need to prove that the objective function Ω is sub-modular. We consider two SBSs allocations X and Y such that $X \subseteq Y$ and we need to prove that the marginal value of adding a new allocated SBS l to cluster i in X and Y verifies:

$$\Omega\left(X \cup \{y_{li}\}\right) - \Omega\left(X\right) \geq \Omega\left(Y \cup \{y_{li}\}\right) - \Omega\left(Y\right). \tag{C.16}$$

Monotonicity is trivial since any new SBS allocation cannot decrease the value of the objective function. In order to show submodularity of the function, we compare the marginal values of adding y_{li} to X and Y.
Here we consider $\Pi_i(X \cup \{y_{li}\})$ referring to the users that change their serving SBS from cluster i. $\mu(u,i)$ refers to the index of the SBS from cluster i serving user u.

A user changes its serving SBS when the new allocated one is closer which induces less transmit power. Consequently, the marginal values of adding y_{li} to X and Y are as follows:

$$\Omega\left(X \cup \{y_{li}\}\right) - \Omega\left(X\right) = \sum_{u \in \Pi_i(X \cup \{y_{li}\})} \left(\sum_{f \in \Delta_i} p_{fu}\right) \left(\omega^{(X \cup \{y_{li}\})}_{u\mu(u,i)} - \omega^{(X)}_{u\mu(u,i)}\right), \tag{C.17}$$

$$\Omega\left(Y \cup \{y_{li}\}\right) - \Omega\left(Y\right) = \sum_{u \in \Pi_i(Y \cup \{y_{li}\})} \left(\sum_{f \in \Delta_i} p_{fu}\right) \left(\omega^{(Y \cup \{y_{li}\})}_{u\mu(u,i)} - \omega^{(Y)}_{u\mu(u,i)}\right). \tag{C.18}$$

Since $X \subseteq Y$ we can deduce that $\Pi_i(Y \cup \{y_{li}\}) \subseteq \Pi_i(X \cup \{y_{li}\})$. Since a user changes its serving SBS only when a closer allocated one is available then $\omega_{u\mu(u,i)}^{(Y \cup \{y_{li}\})} - \omega_{u\mu(u,i)}^{(Y)} > 0$ which proves that $\Omega\left(X \cup \{y_{li}\}\right) - \Omega\left(X\right) \geq \Omega\left(Y \cup \{y_{li}\}\right) - \Omega\left(Y\right)$.

Consequently, Ω is a sub-modular set function. It is simple to verify that the constraints $\sum_{s=1}^{N_s} y_{sk} \leq N_{sk}, \forall k = 1..N_c$ are equivalent to a matroid constraints [45]. Then the considered optimization problem is equivalent to maximizing a sub-modular function subject to matroid constraints.

Bibliography

[1] Issam Toufik, Stefania Sesia, *LTE: The UMTS Long Term Evolution*, John Wiley & Sons, 2011.

[2] White Paper: *Making 5G NR a reality*, Qualcomm Technologies Inc , December 2016.

[3] Wei Xiang, Kan Zheng, Xuemin Shen, *5G Mobile Communications*, Springer International Publishing Switzerland 2017.

[4] Report, *IMT Traffic estimates for the years 2020 to 2030*, ITU-R.

[5] White Paper: *5G Network Architecture - A High-Level Perspective*, HUAWEI Technologies Inc.

[6] T. L. Marzetta, , *Noncooperative cellular wireless with unlimited numbers of base station antennas*, IEEE Transactions on Wireless Communications, vol. 9, no. 11, pp. 3590-3600, November 2010.

[7] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, *Energy and spectral efficiency of very large multiuser MIMO systems*, IEEE Trans. Commun. , vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[8] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, *Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?*, IEEE Trans. Wireless Commun., vol. 14, no. 6, pp. 3059-3075, Jun. 2015

[9] C. Suh, M. Ho, and D. N. C. Tse, *Downlink interference alignment*, IEEE Trans. Commun., vol. 59, no. 9, pp. 2616-2626, Sep. 2011.

[10] S. Verdú, *Multiuser Dectection*, Cambridge, UK: Cambridge University Press, 1998.

[11] N. Jindal and A. Goldsmith, *Dirty-paper coding vs. TDMA for MIMO broadcast channels*, IEEE Trans. Inf. Theory, vol. 51, no. 5, pp. 1783-1794, May 2005.

[12] E. Björnson, E. G. Larsson, and T. L. Marzetta, *Massive MIMO: Ten myths and one critical question*, IEEE Commun. Mag., vol. 54, no. 2, pp. 114-123, Feb. 2016.

[13] B. Gopalakrishnan and N. Jindal, *An analysis of pilot contamination on multi-user MIMO cellular systems with many antennas*, in Proc. Int. Workshop Signal Process. Adv. Wireless Commun., June 2011, pp. 381 385.

[14] F. Fernandes, A. Ashikhmin, and T. Marzetta, *Inter-cell interference in noncooperative TDD large scale antenna systems*, IEEE J. Sel. Areas Commun., vol. 31, no. 2, pp. 192–201, Feb. 2013.

[15] K. Appaiah, A. Ashikhmin, and T. L. Marzetta, *Pilot contamination reduction in multi-user TDD systems*, in Proc. IEEE Int. Conf. Commun. (ICC), 2010, pp. 1–5.

[16] A. Ashikhmin and T.Marzetta, *Pilot contamination precoding in multicell large scale antenna systems*, in Proc. IEEE Int. Symp. Inf. Theory (ISIT), 2012, pp. 1137–1141.

[17] A. Ashikhmin, T. L. Marzetta, and L. Li, *Interference reduction in multi-cell massive MIMO systems I: Large-scale fading precoding and decoding*, 2014 [Preprint], arXiv:1411.4182.

[18] E. Björnson, J. Hoydis, and L. Sanguinetti, *Massive MIMO has unlimited capacity*, IEEE Transactions on Wireless Communications (Volume: 17, Issue: 1, Jan. 2018)

[19] H. Q. Ngo and E. G. Larsson, *EVD-based channel estimation in multicell multiuser MIMO systems with very large antenna arrays*, in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2012, pp. 3249–3252.

[20] R. R. Müller, M. Vehkaperä, and L. Cottatellucci, *Blind pilot decontamination*, in Proc. 17th Int. ITG Workshop Smart Antennas (WSA), 2013, pp. 1–6.

[21] R. R. Muller, M. Vehkapera, and L. Cottatellucci, *Analysis of blind pilot decontamination*, in Proc. Asilomar Conf. Signals Syst. Comput., Nov. 2013, pp. 1016–1020.

[22] D. Neumann, A. Gruendinger, M. Joham, and W. Utschick, *Pilot coordination for large-scale multi-cell TDD systems*, in Proc. 18th Int. ITG Workshop Smart Antennas (WSA), Mar. 2014, pp. 1–6.

[23] M. Latif Sarker and M. H. Lee, *A fast channel estimation and the reduction of pilot contamination problem for massive MIMO based on a diagonal jacket matrix*, in Proc. 4th Int. Workshop Fiber Opt. Access Netw. (FOAN), Sep. 2013, pp. 26–30.

[24] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, *Joint spatial division and multiplexing: the large-scale array regime*, IEEE Trans. Inf. Theory, vol. 59, no. 10, pp. 6441-6463, 2013.

[25] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, *Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling*, IEEE J. of Sel. Topics in Sig. Proc. (JSTSP), vol. 8, no. 5, pp. 876–890, 2014.

[26] Y. Xu, G. Yue, and S. Mao, *User grouping for massive MIMO in FDD systems: New design methods and analysis*, IEEE Access,vol.2, pp. 947–959, Sep. 2014.

[27] Ali Maatouk, Salah Eddine Hajri, Mohamad Assaad, Hikmet Sari, Serdar Sezginer, *Graph Theory Based Approach to Users Grouping and Downlink Scheduling in FDD Massive MIMO*, IEEE ICC, 2018.

[28] Love, D., Choi, J., Bidigare, P.*A closed-loop training approach for massive MIMO beamforming systems*, Proc. of 47th Annual Conference on Information Sciences and Systems (CISS'2013), 2013, pp. 1-5.

[29] Donoho, D.L.*Compressed sensing*, IEEE Transactions on Information Theory, 2006, 52, (4), pp.1289-1306.

[30] Ramasamy, D.,Venkateswaran, S.,Madhow, U.*Compressive tracking with 1000-element arrays: A framework for multi-Gbps mm wave cellular downlinks*, Proc. of 50th Annual Allerton Conference on Communication, Control, and Computing, 2012, pp. 690-697.

[31] Bajwa, W., Haupt, J., Sayeed, A., Nowak, R.*Compressed Channel Sensing: A New Approach to Estimating Sparse Multipath Channels*, Proc. of the IEEE, 2010, 98, (6), pp. 1058-1076.

[32] Araujo, D.C., de Almeida, A.L.F., Axnas, J.,Mota, J.C.M.*Channel estimation for millimeter-wave Very-Large MIMO systems*, Proc. of the 22nd European Signal Processing Conference (EUSIPCO), 2014, pp. 81-85.

[33] L. Thiele, M. Olbrich, M. Kurras, and B. Matthiesen, *Channel aging effects in CoMP transmission: gains from linear channel prediction*,in Proc. of Asilomar Conf. Signals Systems Computers, Nov. 2011, pp. 1924-1928.

[34] K. T. Truong and R. W. Heath Jr., *Effects of channel aging in massive MIMO systems*, J. Commun. Netw., vol. 16, no. 4, pp. 338-351, Aug. 2013.

[35] A. K. Papazafeiropoulos and T. Ratnarajah, *Linear precoding for downlink massive MIMO with delayed CSIT and channel prediction*, in Proc. IEEE WCNC, Apr. 2014, pp. 809-914.

[36] A. K. Papazafeiropoulos and T. Ratnarajah, *Uplink performance of massive MIMO subject to delayed CSIT and anticipated channel prediction*, in Proc. IEEE ICASSP, May 2014, pp. 3162-3165.

[37] Chuili Kong, Caijun Zhong, Anastasios K. Papazafeiropoulos, Michail Matthaiou and Zhaoyang Zhang, *Sum-Rate and Power Scaling of Massive MIMO Systems with Channel Aging*, Information Theory (ISIT), 2015.

[38] Hassan, N.; Fernando, X. *Massive MIMO Wireless Networks: An Overview*. Electronics 2017, 6, 63. [Google Scholar] [CrossRef]

[39] Lu, L.; Li, G.Y.; Swindlehurst, A.L.; Ashikhmin, A.; Zhang, R. An overview of massive MIMO: Benefits and challenges. IEEE J. Sel. Top. Signal Process. 2014, 8, 742–758.

[40] Dai, L.; Gao, X.; Quan, J.; Han, S.; Chih-Lin, I. Near-optimal hybrid analog and digital precoding for downlink mmWave massive MIMO Systems. In Proceedings of the 2015 IEEE International Conference on Communications (ICC 2015), London, UK, 8–12 June 2015; pp. 1334–1339.

[41] Belady, Laszlo A., *A study of replacement algorithms for a virtual-storage computer*, IBM Syst. J., 5, 78-101, 1996.

[42] Wang, Jia, *A survey of web caching schemes for the internet*, ACM SIGCOMM Computer Communication Review, 29, 36-46, 1999.

[43] Song, Chaoming and Qu, Zehui and Blumm, Nicholas and Barabási, Albert-László, *Limits of Predictability in Human Mobility*, Science, 327, 1018-1021, 2010.

[44] Ejder Baştuğ and Mehdi Bennis and Mérouane Debbah, *Living on the Edge: The role of Proactive Caching in 5G Wireless Networks*, IEEE Communications Magazine, 52, 82-89, 2014.

[45] K. Shanmugam, N. Golrezaei, A. G. Dimakis, Andreas F. Molisch, Giuseppe Caire, *FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers*, Proceedings IEEE INFOCOM, 2012.

[46] Ejder Baştuğ and Mehdi Bennis and Mérouane Debbah, *Anticipatory Caching in Small Cell Networks: A Transfer Learning Approach*, 1st KuVS Workshop on Anticipatory Networks, 2014.

[47] Ejder Baştuğ, Kenza Hamidouche and Walid Saad and Mérouane Debbah, *Centrality-Based Caching for Mobile Wireless Networks*, 1st KuVS Workshop on Anticipatory Networks, 2014.

[48] Kenza Hamidouche and Walid Saad and Mérouane Debbah, *Many-to-many matching games for proactive social-caching in wireless small cell networks*, 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 569-574, 2014.

[49] Salah Eddine HAJRI and Mohamad Assaad, *Caching improvement using adaptive user clustering*, 17th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Edinburgh, UK, 3-6 July 2016.

[50] Salah Eddine HAJRI and Mohamad Assaad, *Energy Efficiency in Cache Enabled Small Cell Networks With Adaptive User Clustering*, IEEE Transactions on Wireless Communications (Volume: PP, Issue: 99), 17 November 2017.

[51] M. S. ElBamby, M. Bennis, W. Saad, M. Latva-aho, *Content-Aware User Clustering and Caching in Wireless Small Cell Networks*, 11th International Symposium on Wireless Communications Systems (ISWCS), 2014.

[52] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, *Fundamental Limits of Caching*, IEEE Transactions on Information Theory, 60, 2856-2867, 2014.

[53] Niesen, Urs and Maddah-Ali, Mohammad Ali, *Coded caching with nonuniform demands*, arXiv preprint arXiv: 1308.0178, 2013.

[54] Hachem, Jad and Karamchandani, Nikhil and Diggavi, Suhas, *Multi-level Coded Caching*, arXiv preprint arXiv: 1404.6563, 2014.

[55] Hachem, Jad and Karamchandani, Nikhil and Diggavi, Suhas, *Content Caching and Delivery over Heterogeneous Wireless Networks*, arXiv preprint arXiv: 1404.6560, 2014.

[56] Sinong Wang and Wenxin Li and Xiaohua Tian and Hui Liu, *Fundamental Limits of Heterogenous Cache*, arXiv preprint arXiv: 1504.01123, 2015.

[57] Pedarsani, Ramtin and Maddah-Ali, Mohammad Ali and Niesen, Urs, *Online coded caching*, arXiv preprint arXiv: 1311.3646, 2013.

[58] Karamchandani, Nikhil and Niesen, Urs and Maddah-Ali, Mohammad Ali and Diggavi, Suhas, *Hierarchical coded caching*, IEEE International Symposium on Information Theory (ISIT'14), 2142-2146, 2014.

[59] Shariatpanahi, Seyed Pooya and Motahari, Seyed Abolfazl and Khalaj, Babak Hossein, *Multi-Server Coded Caching*, arXiv preprint arXiv: 1503.00265, 2015.

[60] Ji, Mingyue and Caire, Giuseppe and Molisch, Andreas F, *Wireless device-to-device caching networks: Basic principles and system performance*, arXiv preprint arXiv: 1305.5216, 2014.

[61] Ji, Mingyue and Caire, Giuseppe and Molisch, Andreas F, *Fundamental Limits of Caching in Wireless D2D Networks*, arXiv preprint arXiv: 1405.5336, 2014.

[62] Mingyue Ji, Giuseppe Caire, Andreas F. Molisch, *On the Average Performance of Caching and Coded Multicasting with Random Demands*,11th International Symposium on Wireless Communication Systems (ISWCS'14), 2014.

[63] Ji, Mingyue and Tulino, Antonia M and Llorca, Jaime and Caire, Giuseppe, *Order-Optimal Rate of Caching and Coded Multicasting with Random Demands*, arXiv preprint arXiv: 1502.03124, 2015.

[64] Pääkkönen, Joonas and Hollanti, Camilla and Tirkkonen, Olav, *Device-to-device data storage for mobile cellular systems*, IEEE Globecom Workshops (GC Wrokshops), 2013.

[65] Ostovari, Pouya and Khreishah, Abdallah and Wu, Jie, *Cache content placement using triangular network coding*, IEEE Wireless Communications and Networking Conference (WCNC), 1375-1380, 2013.

[66] Wei Huang and Sinong Wang and Lianghui Ding and Feng Yang and Wenjun Zhang, *The Performance Analysis of Coded Cache in Wireless Fading Channel*,arXiv preprint arXiv: 1504.01452, 2015.

[67] Molisch, Andreas F and Caire, Giuseppe and Ott, David and Foerster, Jeffrey R and Bethanabhotla, Dilip and Ji, Mingyue, *Caching Eliminates the Wireless Bottleneck in Video-Aware Wireless Networks*, arXiv preprint arXiv: 1405.5864, 2014.

[68] Golrezaei, Negin and Molisch, Andreas F and Dimakis, Alexandros G and Caire, Giuseppe, *Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution*, IEEE Communications Magazine, 142-149, 2013.

[69] Blasco, Pol and Gündüz, Deniz, *Learning-based optimization of cache content in a small cell base station*, arXiv preprint arXiv: 1402.3247, 2014.

[70] Blasco, Pol and Gündüz, Deniz, *Content-Level Selective Offloading in Heterogeneous Networks: Multi-armed Bandit Optimization and Regret Bounds*, arXiv preprint arXiv: 1407.6154, 2014.

[71] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, *Wireless device-to-device communication with distributed caching*, IEEE International Symposium on Information Theory Proceedings (ISIT), July 2012.

[72] Ejder Baştuğ and Mehdi Bennis and Marios Kountouris and Mérouane Debbah, *Cache-enabled Small Cell Networks: Modeling and Tradeoffs*,EURASIP Journal on Wireless Communications and Networking, 2015.

[73] Bartlomiej Blaszczyszyn and Anastasios Giovanidis, *Optimal Geographic Caching In Cellular Networks*, IEEE International Conference on Communications (ICC), 2015.

[74] Mihaela Mitici and Jasper Goseling and Maurits de Graaf and Richard J. Boucherie, *Deployment vs. data retrieval costs for caches in the plane*, University of Twente, Department of Applied Mathematics, 2013.

[75] Bhanukiran Perabathini and Ejder Baştuğ and Marios Kountouris and Mérouane Debbah and Alberto Conte, *Caching on the Edge: a Green Perspective for 5G Networks*,IEEE International Conference on Communications (ICC'15), 2015.

[76] D. Liu and C. Yang, *Energy Efficiency of Downlink Networks With Caching at Base Stations*, IEEE Journal on Selected Areas in Communications (Volume: 34, Issue: 4), April 2016.

[77] Zhou, Sheng and Gong, Jie and Zhou, Zhenyu and Chen, Wei and Niu, Zhisheng, *GreenDelivery: Proactive content caching and push with energy harvesting based small cells*, arXiv preprint arXiv: 1503.04254, 2015.

[78] Poularakis, Konstantinos and Iosifidis, George and Tassiulas, Leandros, *Joint Caching and Base Station Activation for Green Heterogeneous Cellular Networks*, IEEE International Conference on Communications (ICC'15), 2015.

[79] D. S. Hochbaum, editor. *Approximation algorithms for NP-hard problems*. PWS Publishing Co., Boston, MA, USA, 1997.

[80] R. Cohen and L. Klatzir, *The generalized maximum coverage problem*. Inf. Process. Lett., 108(1):15–22, 2008.

[81] S. Sahni and T. Gonzalez, "P-complete approximation problems," Journal of the Association for Computing Machinery, vol.23, No.3, pp.555-565, July 1976.

[82] A. Adhikary, E. Al Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, *Joint spatial division and multiplexing for mm-wave channels*, IEEE J. on Sel. Areas on Commun. (JSAC), vol. 32, no. 6, pp. 1239–1255, 2014.

[83] S. Haghighatshoar and G. Caire, *Massive MIMO channel subspace estimation from low-dimensional projections*, IEEE Transactions on Signal Processing, vol. 65, pp. 303–318, 2017

[84] H. S. Chang, P. J. Fard, S. I. Marcus, and M. Shayman, *Multitime scale markov decision processes*, Automatic Control, IEEE Transactions on, vol. 48, jun 2003.

[85] I. M. Taylor and M. A. Labrador, *Improving the Energy Consumption in Mobile Phones by Filtering Noisy GPS Fixes with Modified Kalman Filters*, Proceedings of IEEE Wireless Communications and Networking Conference, March 28-31 2011, pp. 2006-2011.

[86] LaValle, S. M.,*Planning Algorithms*, University of Illinois 1999-2004.

[87] D. BRAZIUNAS, *POMDP solution methods*. Tech. rep., 2003

[88] J. Hoydis, S. Ten Brink, and M. Debbah, *Massive MIMO in the ul/dl of cellular networks: How many antennas do we need?* IEEE J. on Sel. Areas on Commun. (JSAC), vol. 31, no. 2, pp. 160–171, 2013.

[89] O. Elijah, C. Leow, T. Rahman, S. Nunoo, and S. Iliya, *A Comprehensive Survey of Pilot Contamination in Massive MIMO-5G System*, IEEE Commun. Surv. Tut. , vol. 99, Nov. 2015.

[90] H. Yin, D. Gesbert, M. Filippou, Y. Liu, *A Coordinated Approach to Channel Estimation in Large-Scale Multiple-Antenna Systems*, Selected Areas in Communications, IEEE Journal on, vol. 31, no. 2, 2013

[91] METIS Deliverable D3.3, *Final Performance Results and Consolidated View on the Most Promising Multi-Node/Multi-Antenna Transmission Technologies*, February 2015.

[92] Xie, F. Gao, S. Zhang, and S. Jin, *A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model*, IEEE Transactions on Vehicular Technology (Volume: 66, Issue: 4, April 2017).

[93] O. Elijah, C. Leow, T. Rahman, S. Nunoo, and S. Iliya, “A Comprehensive Survey of Pilot Contamination in Massive MIMO-5G System,” IEEE Commun. Surv. Tut. , vol. 99, Nov. 2015.

[94] Geoff Varrall, *5G Spectrum and Standards*, Artech House, 31 mai 2016

[95] E. Björnson, J. Hoydis, and L. Sanguinetti, *Massive MIMO networks: Spectral, energy, and hardware efficiency*, Foundations and Trends in Signal Processing ,vol.11,no.3-4,pp.154–655,2017.[Online].Available:http://dx.doi.org/10.1561/2000000093.

[96] White Paper:Gabriel Brown, *Exploring the potential of mmWave for 5G Mobile Access* , Qualcomm Technologies Inc , June 2016.

[97] YANG Shan, CHEN Peng, LIANG Lin, ZHU Jianchi, and SHE Xiaoming , *Uplink Multiple Access Schemes for 5G: A Survey*. ZTE Communications.

[98] ETSI, *Final report of 3GPP TSG RAN1 86 v1.0.0*, R1-1608562, 3GPP TSG RAN WG1 Meeting 86, Gothenburg, Sweden, Aug.2016.

[99] Samsung, *Non-orthogonal multiple access candidate for NR*, R1-163992, 3GPP TSG RAN WG1 Meeting 85, Nanjing, China, May 2016.

[100] Nokia, Alcatel-Lucent Shanghai Bell, *Performance of Interleave Division Multiple Access (IDMA) in combination with OFDM family waveforms*, R1-165021, 3GPP TSG RAN WG1 Meeting 85, Nanjing, China, May 2016.

[101] FANTASTIC 5G Deliverable D3.2, *Final report on the holistic link solution adaptation*, Apr. 2017.

[102] 3GPP RP-171517, *Status report for RAN WG1 to TSG-RAN 77*, Docomo, Sep. 2017

[103] 3GPP T.S. 38.211, *Physical channels and modulation*, V1.0.0, Sep. 2017.

[104] 3GPP T.R. 38.802, *Study on New Radio Access Technology–Physical Layer Aspects*, V14.2.0, Sep. 2017.

[105] 3GPP T.R. 38.913, *Study on Scenarios and Requirements for Next Generation Access Technologies*, V14.2.0, Mar. 2017

[106] Salah Eddine HAJRI and Mohamad Assaad, *A spatial basis coverage approach for uplink training and scheduling in Massive MIMO systems*, submitted to IEEE Transactions on Wireless Communications 2018.

[107] Salah Eddine HAJRI, Maialen Larranaga and Mohamad Assaad, *Heterogeneous Doppler Spread-based CSI Estimation Planning for TDD Massive MIMO* , submitted to IEEE Transactions on Wireless Communications 2017.

[108] Ali Maatouk, Salah Eddine Hajri, Mohamad Assaad, Hikmet Sari, *Optimal Scheduling for Joint Spatial Division and Multiplexing: Complexity Results and Approximation Algorithm*, submitted to IEEE Transactions on Wireless Communications 2018.

[109] Salah Eddine HAJRI and Mohamad Assaad, *An Exclusion zone for Massive MIMO With Underlay D2D Communication*, International Symposium on Wireless Communication Systems (ISWCS), Aug. 2015 .

[110] Salah Eddine HAJRI, Mohamad Assaad and Giuseppe Caire, *Scheduling in Massive MIMO: User clustering and pilot assignment*, 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2016.

[111] Salah Eddine Hajri, Mohamad Assaad, Maialen Larranaga, *Enhancing massive MIMO: A new approach for Uplink training based on heterogeneous coherence time*, accepted in ICT'18, 2018.

[112] Salah Eddine Hajri, Juwendo Denis and Mohamad Assaad, *Enhancing Favorable Propagation in Cell-Free Massive MIMO Through Spatial User Grouping*, accepted in IEEE SPAWC 2018.

[113] Mohamad Assaad, Salah Eddine Hajri, Thomas Bonald, Anthony Ephremides, *Power Control in Massive MIMO with Dynamic User Population*, submitted to WiOpt'18.

[114] Salah Eddine HAJRI and Mohamad Assaad, *Relay operations in a cellular network*, gb GB1710498.5, UK, TCL Communication Limited, 2017.

[115] Salah Eddine HAJRI and Mohamad Assaad, *IMPROVEMENTS IN OR RELATING TO DYNAMIC CHANNEL AUTOCORRELATION BASED ON USER SCHEDULING*, gb GB1710487.8, UK, TCL Communication Limited, 2017.

[116] Salah Eddine HAJRI and Mohamad Assaad, *Sparse uplink CSI estimation for MTC devices in massive MIMO systems*, TCL Communication Limited, 2018.

[117] E. G. Larsson, O. Edfors, and T. L. Marzetta, *Massive MIMO for next generation wireless systems*, IEEE Commun. Mag. , vol. 52, no. 2, pp. 186-195, Feb. 2014.

[118] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, *Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation and capacity limits*, IEEE Trans. Inf. Theory, vol. 60, no. 11, pp. 7112-7139, Nov. 2014.

[119] V. Pauli, Y. Li, and E. Seidel. *Dynamic TDD for LTE-A and 5G* . Nomor Research GmbH, Munich, Germany, September 20150.

[120] Akhil Gupta, Rakesh Kumar Jha, *A survey of 5G network: Architecture and emerging technologies*, IEEE access (3), pp. 1206–1232, 2015.

[121] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, *Energy and spectral efficiency of very large multiuser MIMO systems*, IEEE Trans. Commun., 2012, submitted. [Online]. Available: http://arxiv.org/abs/1112.3810

[122] A. Ashikhmin and T. L. Marzetta, *Pilot contamination and precoding in multi-cell large scale antenna systems*, in Proc. IEEE Int. Symp. Inf. Theory, Cambridge, MA, 2012, pp. 1142-1146.

[123] Thang X. Vu, Trinh Anh Vu, Symeon Chatzinotas and Bjorn Ottersten, *Spectral-Efficient Model for Multiuser Massive MIMO: Exploiting User Velocity*, IEEE ICC'17, Mai 2017.

[124] W. Jakes and D. Cox, *Microwave Mobile Communications*, Wiley-IEEE Press, 1994.

[125] Tamás Szálka and Péter Fulop and Sándor Imre, *General mobility modeling and location prediction based on markovian approach constructor framework*, Periodica Polytechnica Electrical Engineering (Archives), 55, 2011.

[126] Yang Li, Young-Han Nam, Boon Loong Ng, Jianzhong Zhang*A non-asymptotic throughput for massive MIMO cellular uplink with pilot reuse*,Global Communications Conference (GLOBECOM), 2012 IEEE.

[127] Volker Pauli, Yi Li, Eiko Seidel *Dynamic TDD for LTE-A and 5G*, Nomor Research GmbH, Munich, Germany, September 2015.

[128] H. Xie, F. Gao, and S. Jin, *An overview of low-rank channel estimation for massive MIMO systems*, IEEE Access , vol. 4, pp. 7313–7321, 2016

[129] A. Ashikhmin and T. L. Marzetta, *Pilot contamination and precoding in multi-cell large scale antenna systems*, in Proc. IEEE Int. Symp. Inf. Theory, Cambridge, MA, 2012, pp. 1142-1146.

[130] Xu, G. Yue, N. Prasad, S. Rangarajan, and S. Mao, *User grouping and scheduling for large scale MIMO systems with two-stage precoding*, in Proc. IEEE ICC, Sydney, Australia, Jun. 2014, pp. 5208–5213

[131] Alexei Ashikhmin, Thomas L. Marzetta and Liangbin Li,*Interference Reduction in Multi-Cell Massive MIMO Systems I: Large-Scale Fading Precoding and Decoding*, November 2014.

[132] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, *5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice*, IEEE Journal on Selected Areas in Communications (Volume: 35, Issue: 6, June 2017) .

[133] G. Wunder, H. Boche, T. Strohmer, and P. Jung. , *Sparse Signal Processing Concepts for Efficient 5G System Design*, EEE Access , accepted, 2014 .

[134] P. P. G. Wunder, C. Stefanovic , *Compressive coded random access for massive MTC traffic in 5G systems*, 49th Asilomar Conference on Signals, Systems and Computers, 2015 .

[135] L. Sanguinetti, A. A. D'Amico, M. Morelli, and M. Debbah, *Random access in uplink massive MIMO systems: How to exploit asynchronicity and excess antennas*, in IEEE GLOBECOM , 2016.

[136] Vineeth S. Varma ; Salah Eddine Elayoubi ; Merouane Debbah ; Samson Lasaulce, *On the energy efficiency of virtual MIMO systems*, IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops), 2013.

[137] Hien Quoc Ngo ; Le-Nam Tran ; Trung Q. Duong ; Michail Matthaiou ; Erik G. Larsson, *Energy efficiency optimization for cell-free massive MIMO*, IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) , 2017.

[138] Stefano Buzzi ; Carmen D'Andrea, *Cell-Free Massive MIMO: User-Centric Approach*, IEEE Wireless Communications Letters (Volume: 6, Issue: 6, Dec. 2017).

[139] Hien Quoc Ngo ; Le-Nam Tran ; Trung Q. Duong ; Michail Matthaiou ; Erik G. Larsson, *On the Total Energy Efficiency of Cell-Free Massive MIMO*, IEEE Transactions on Green Communications and Networking (Volume: PP, Issue: 99).

[140] Hien Quoc Ngo ; Alexei Ashikhmin ; Hong Yang ; Erik G. Larsson ; Thomas L. Marzetta, *Cell-Free Massive MIMO Versus Small Cells*, IEEE Transactions on Wireless Communications (Volume: 16, Issue: 3, March 2017).

[141] Hien Quoc Ngo ; Alexei Ashikhmin ; Hong Yang ; Erik G. Larsson ; Thomas L. Marzetta, *Cell-Free Massive MIMO: Uniformly great service for everyone*, IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2015.

[142] Khem Narayan Poudel ; Shankar Gangaju, *Spectral Efficiency, Diversity Gain and Multiplexing Capacity Analysis for Massive MIMO, 5G Communications System*, International Conference on Networking and Network Applications (NaNA), 2017.

[143] Meysam Sadeghi ; Luca Sanguinetti ; Romain Couillet ; Chau Yuen, *Large System Analysis of Power Normalization Techniques in Massive MIMO*, IEEE Transactions on Vehicular Technology (Volume: 66, Issue: 10, Oct. 2017).

[144] Emil Björnson ; Jakob Hoydis ; Luca Sanguinetti, *Pilot contamination is not a fundamental asymptotic limitation in massive MIMO*, IEEE International Conference on Communications (ICC), 2017.

[145] Luca Sanguinetti ; Abla Kammoun ; Merouane Debbah, *Asymptotic analysis of multicell massive MIMO over Rician fading channels*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[146] Emil Björnson ; Luca Sanguinetti ; Merouane Debbah, *Massive MIMO with imperfect channel covariance information*, 50th Asilomar Conference on Signals, Systems and Computers, 2016.

[147] Luca Sanguinetti ; Antonio A. D'Amico ; Michele Morelli ; Merouane Debbah, *Random Access in Uplink Massive MIMO Systems: How to Exploit Asynchronicity and Excess Antennas*, IEEE Global Communications Conference (GLOBECOM), 2016 .

[148] Matha Deghel ; Mohamad Assaad ; Mérouane Debbah ; Anthony Ephremides, *Queueing Stability and CSI Probing of a TDD Wireless Network With Interference Alignment*, IEEE Transactions on Information Theory (Volume: 64, Issue: 1, Jan. 2018).

[149] Rasha Al Khansa ; Jean J. Saade ; Hassan A. Artail ; Mohamad Assaad, *A small cell approach to optimizing the coverage of MTC systems with massive MIMO and random access using stochastic geometry*, Wireless and Mobile Computing, Networking and Communications (WiMob), 2017.

[150] Jérôme Gaveau ; Christophe J. Le Martret ; Mohamad Assaad, *Grouping of subcarriers and effective SNR statistics in wideband OFDM systems using EESM*, Wireless and Mobile Computing, Networking and Communications (WiMob), 2017.

[151] Maialen Larrañaga ; Mohamad Assaad ; Apostolos Destounis ; Georgios S. Paschos, *Asymptotically optimal pilot allocation over Markovian fading channels*, IEEE Transactions on Information Theory (Volume: PP, Issue: 99), 2017.

[152] Rita Ibrahim ; Mohamad Assaad ; Berna Sayrac ; Anthony Ephremides, *Overlay D2D vs. Cellular communications: A stability region analysis*, International Symposium on Wireless Communication Systems (ISWCS), 2017.

[153] Amira Akra; Mohamad Assaad, *Energy efficient transmit beamforming under queueing stability constraints*, IEEE Wireless Communications and Networking Conference, 2016.

[154] Ayaz Ahmad; Naveed Ul Hassan; Mohamad Assaad; Hamidou Tembine, *Joint Power Control and Rate Adaptation for Video Streaming in Wireless Networks With Time-Varying Interference*, IEEE Transactions on Vehicular Technology, 2016.

[155] Maialen Larrañaga ; Mohamad Assaad ; Apostolos Destounis ; Georgios S. Paschos, *Dynamic pilot allocation over Markovian fading channels: A restless bandit approach*, Information Theory Workshop (ITW), 2016 IEEE.

[156] Matha Deghel ; Mohamad Assaad ; Merouane Debbah, *Queuing stability and CSI probing of a TDD wireless network with interference alignment*, IEEE International Symposium on Information Theory (ISIT), 2015.

[157] Matha Deghel ; Mohamad Assaad ; Mérouane Debbah, *System performance of interference alignment under TDD mode with limited backhaul capacity*, IEEE International Conference on Communications (ICC), 2015.

[158] Subhash Lakshminarayana ; Mohamad Assaad ; Mérouane Debbah, *Coordinated Multicell Beamforming for Massive MIMO: A Random Matrix Approach*, IEEE Transactions on Information Theory (Volume: 61, Issue: 6, June 2015).

[159] W. LI, M. Assaad, P. Duhamel, *Distributed Stochastic Optimization in Networks with Low Information Exchange*, in Proc. of IEEE Allerton conference, 2017

[160] Muller, R.R.,Cottatellucci, L. and Vehkapera, M.*Blind Pilot Decontamination*, IEEE Journal of Selected Topics in Signal Processing, 2014.

[161] Taesang Yoo, Andrea Goldsmith*on the optimality of multi antenna broadcast scheduling using zero-forcing beamforming*, IEEE Journal on selected areas in communications March 2006.

[162] Ejder Baştuğ, M. Bennis and M. Debbah, *Social and Spatial Proactive Caching for Mobile Data Offloading*, in Proc. IEEE International Conference on Communications (ICC) 2014, Sydney, Australia, June 2014.

[163] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, *Wireless caching: technical misconceptions and business barriers*, IEEE Communications Magazine, vol. 54, no. 8, pp. 16–22, August 2016.

[164] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, *A survey of information-centric networking*, IEEE Communications Magazine (Volume: 50, Issue: 7), July 2012.

[165] M. Deghel, E. Bastug, M. Assaad and M. Debbah, *On the benefits of Edge Caching for MIMO Interference Alignment*, IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2015.

[166] K. Poularakis, G. Iosifidis, and L. Tassiulas, *Approximation algorithms for mobile data caching in small cell networks*, IEEE Transactions on Communications (Volume: 62, Issue: 10), Oct. 2014.

[167] M. Maddah-Ali and U. Niesen, *Fundamental limits of caching*, IEEE Transactions on Information Theory (Volume: 60, Issue: 5), May 2014.

[168] E. Bastug, M. Bennis, and M. Debbah, *Living on the edge: The role of proactive caching in 5G wireless networks*, IEEE Commun. Mag., vol. 52, no. 8, pp. 82-89, Aug. 2014.

[169] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, *Hierarchical coded caching*, IEEE Transactions on Information Theory (Volume: 62, Issue: 6), June 2016.

[170] S. Zhou, J. Gong, Z. Zhou, W. Chen and Z. Niu, *GreenDelivery: proactive content caching and push with energy-harvesting-based small cells*, in IEEE Communications Magazine, vol. 53, no. 4, pp. 142-149, April 2015.

[171] 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 Journal on Selected Areas in Communications.

[172] Zhongyuan Zhao, Mugen Peng, Zhiguo Ding, Wenbo Wang, and H. Vincent Poor, *Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks*, http://arxiv.org/abs/1603.07052.

[173] H. Aikaike, *A New Look at the Statistical Model Identification*, IEEE Transactions on Automatic Control, Dec. 1974.

[174] J. E. Cavanaugh, *Unifying the derivations for the Akaike and corrected Akaike information criteria*, Statistics & Probability Letters, April 1997.

[175] T. Y. Young and T. W. Calvert, *Classification, estimation and pattern recognition*, American Elsevier Publisher, 1974.

[176] S. Tamoor-ul-Hassan, M. Bennis, P. H. J. Nardelli, M. Latva-Aho, *Modeling and Analysis of Content Caching in Wireless Small Cell Networks*, International Symposium on Wireless Communication Systems (ISWCS), 2015.

[177] H. ElSawy, E. Hossain and M. S. Alouini, *Analytical Modeling of Mode Selection and Power Control for Underlay D2D Communication in Cellular Networks*, IEEE Transactions on Communications (Volume: 62, Issue: 11), Nov. 2014.

[178] P. Blasco and D. Gunduz, *Multi-armed bandit optimization of cache content in wireless infostation networks*, IEEE International Symposium on Information Theory (ISIT), 2014

[179] 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*, IIEEE Journal on Selected Areas in Communications , vol. 35, no. 5, pp. 1046–1061, May 2017.

[180] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, *Optimal Transport Theory for Cell Association in UAV-Enabled Cellular Networks*, IEEE Communications Letters , vol. 21, no. 9, pp. 2053–2056, Sep. 2017.

[181] K. Hamidouche, W. Saad, M. Debbah, J. B. Song, and C. S. Hong, *The 5G Cellular Backhaul Management Dilemma: To Cache or to Serve*, submitted to IEEE Transactions on Wireless Communications , 2016.

[182] Syed Tamoor-ul-Hassan ; Sumudu Samarakoon ; Mehdi Bennis ; Matti Latva-aho ; Choong Seon Hong, *Learning-Based Caching in Cloud-Aided Wireless Networks*, IEEE Communications Letters (Volume: 22, Issue: 1, Jan. 2018).

[183] Konglin Zhu ; Wenting Zhi ; Lin Zhang ; Xu Chen ; Xiaoming Fu, *Social-Aware Incentivized Caching for D2D Communications*, IEEE Access (Volume: 4), 2016.

[184] Andriana Ioannou ; Stefan Weber, *A Survey of Caching Policies and Forwarding Mechanisms in Information-Centric Networking*, IEEE Communications Surveys & Tutorials (Volume: 18, Issue: 4, Fourthquarter 2016)

[185] Meng Zhang ; Hongbin Luo ; Hongke Zhang, *A Survey of Caching Mechanisms in Information-Centric Networking*, IEEE Communications Surveys & Tutorials (Volume: 17, Issue: 3, thirdquarter 2015)

[186] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Springer,2004

[187] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 3rd ed. New York: Springer-Verlag, 2006.

[188] D. Pelleg, A. Moore, *X-means: extending k-means with efficient estimation of the number of clusters*, ICML-2000, 2000.

[189] T. Y. Young and T. W. Calvert, *Classification, estimation and pattern recognition*, American Elsevier Publishing Company, 1974.

[190] J. Lee, V. Mirrokni, V. Nagarajan and M. Sviridenko. *Maximizing Non-Monotone Submodular Functions under Matroid and Knapsack Constraints*. IBM Research Report RC24679. 2008.

[191] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrak, *Maximizing a submodular set function subject to a matroid constraint*, Integer programming and combinatorial optimization, pp. 182-196, 2007.

[192] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific. 1995.

[193] Aberdeen, D. *A survey of approximate methods for solving partially observable Markov decision processes*, Technical report, Research School of Information Science and Engineering, Australia National University, 2002.

[194] I. Nourbakhsh, R. Powers, and S. Birchfield, *DERVISH an office-navigating robot*, AI Magazine 16, pp. 53-60, 1995.

[195] K. J. Arrow and A.C. Enthoven. *Quasi-concave programming*, Econometrica, 29:779-800, 1961.

[196] G. Nemhauser and L. Wolsey, *Integer and combinatorial optimization*, Wiley New York, 1988, vol. 18.

[197] R. M. Karp, *Reducibility among Combinatorial Problems*, The IBM Research Symposia Series, Springer, New York, pp. 85-103, 1972.

[198] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrak, *Maximizing a submodular set function subject to a matroid constraint*, Integer programming and combinatorial optimization, pp. 182-196, 2007.

[199] A. Ageev and M. Sviridenko. , *Pipage rounding: a new method of constructing algorithms with proven performance guarantee*,J. of Combinatorial Optimization, 307-328, 2004

[200] Boyd and Vandenberghe *Convex Optimization*, Cambridge university press, 2004.

[201] Ronald Y. Chang, Zhifeng Tao, Jinyun Zhang, and C.-C. Jay Kuo, *Multicell OFDMA Downlink Resource Allocation Using a Graphic Framework*, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, VOL. 58, NO. 7, SEPTEMBER 2009.

[202] Liang Zhao, Hiroshi Nagamochi, Toshihide Ibaraki *Greedy splitting algorithms for approximating multiway partition problems*, Mathematical programming (2005)

[203] Goldschmidt, O., Hochbaum, D.S. (1994),*A polynomial algorithm for the k-cut problem for fixed k* Math. Oper. Res. 1,9 2437

[204] Maurice Queyranne, *Minimizing symmetric submodular functions*, Mathematical programming 82 (1998)

[205] S. H. Tung, *On lower and Upper Bounds of the Difference Between the Arithmetic and the Geometric Mean*, MATHEMATICS OF COMPUTATION, VOLUME 29, NUMBER 131, JULY 1975

[206] J. G. Oxley, *Matroid theory*, Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1992.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies
de l'information
et de la communication (STIC)

Titre : L'amélioration des performances des systèmes sans fil 5G par groupements adaptatifs des utilisateurs

Mots clés : Mise en cache proactive, MIMO massif, apprentissage automatique, réseaux cellulaires, 5G

Résumé: 5G est prévu pour s'attaquer, en plus d'une augmentation considérable du volume de trafic, la tâche de connecter des milliards d'appareils avec des exigences de service hétérogènes. Afin de relever les défis de la 5G, nous préconisons une utilisation plus efficace des informations disponibles, avec plus de sensibilisation par rapport aux services et aux utilisateurs, et une expansion de l'intelligence du RAN. En particulier, nous nous concentrons sur deux activateurs clés de la 5G, à savoir le MIMO massif et la mise en cache proactive. Dans le troisième chapitre, nous nous concentrons sur la problématique de l'acquisition de CSI dans MIMO massif en TDD. Pour ce faire, nous proposons de nouveaux schémas de regroupement spatial tels que, dans chaque groupe, une couverture maximale de la base spatiale du signal avec un chevauchement minimal entre les signatures spatiales des utilisateurs est obtenue. Ce dernier permet d'augmenter la densité de connexion tout en améliorant l'efficacité spectrale. MIMO massif en TDD est également au centre du quatrième chapitre. Dans ce cas, en se basant sur les différents taux de vieillissement des canaux sans fil, la périodicité d'estimation de CSI est supplémentaire. Nous le faisons en proposant un exploité comme un degré de liberté supplémentaire. Nous le faisons en proposant une adaptation dynamique de la trame TDD en fonction des temps de cohérence des canaux hétérogènes. Les stations de bases MIMO massif sont capables d'apprendre la meilleure politique d'estimation sur le uplink pour de longues périodes. Comme les changements de canaux résultent principalement de la mobilité de l'appareil, la connaissance de l'emplacement est également incluse dans le processus d'apprentissage. Le problème de planification qui en a résulté a été modélisé comme un POMDP à deux échelles temporelles et des algorithmes efficaces à faible complexité ont été fournis pour le résoudre. Le cinquième chapitre met l'accent sur la mise en cache proactive. Nous nous concentrons sur l'amélioration de l'efficacité énergétique des réseaux dotes de mise en cache en exploitant la corrélation dans les modèles de trafic en plus de la répartition spatiale des demandes. Nous proposons un cadre qui établit un compromis optimal entre la complexité et la véracité dans la modélisation du comportement des utilisateurs grâce à la classification adaptative basée sur la popularité du contenu. Il simplifie également le problème du placement de contenu, ce qui se traduit par un cadre d'allocation de contenu rapidement adaptable et économe en énergie.

Title: Performance improvement of 5G Wireless Systems through adaptive grouping of users

Keywords: Massive MIMO, proactive caching, machine learning, cellular networks, 5G

Abstract: 5G is envisioned to tackle, in addition to a considerable increase in traffic volume, the task of connecting billions of devices with heterogeneous service requirements. In order to address the challenges of 5G, we advocate a more efficient use of the available information, with more service and user awareness, and an expansion of the RAN intelligence. In particular, we focus on two key enablers of 5G, namely massive MIMO and proactive caching. In the third chapter, we focus on addressing the bottleneck of CSI acquisition in TDD Massive MIMO. In order to do so, we propose novel spatial grouping schemes such that, in each group, maximum coverage of the signal's spatial basis with minimum overlapping between user spatial signatures is achieved. The latter enables to increase connection density while improving spectral efficiency. TDD Massive MIMO is also the focus of the fourth chapter. Therein, based on the different rates of wireless channels aging, CSI estimation periodicity is exploited as an additional DoF. We do so by proposing a dynamic adaptation of the TDD frame based on the heterogeneous channels coherence times. The Massive MIMO BSs are enabled to learn the best uplink training policy for long periods. Since channel changes result primarily from device mobility, location awareness is also included in the learning process. The resulting planning problem was modeled as a two-time scale POMDP and efficient low complexity algorithms were provided to solve it. The fifth chapter focuses on proactive caching. We focus on improving the energy efficiency of cache-enabled networks by exploiting the correlation in traffic patterns in addition to the spatial repartition of requests. We propose a framework that strikes the optimal trade-off between complexity and truthfulness in user behavior modeling through adaptive content popularity-based clustering. It also simplifies the problem of content placement, which results in a rapidly adaptable and energy efficient content allocation framework.