

# High Performance Cache-Aided Downlink Systems: Novel Algorithms and Analysis

Dissertation

*submitted to*

Sorbonne University

*in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy*

*Author:*

**Hui ZHAO**

*Scheduled for defense on the 16th December, 2022, before a committee composed of:*

<i>Examiner/Reviewer</i>	<b>Prof. Antti Tölli</b>	University of Oulu, Finland
<i>Examiner/Reviewer</i>	<b>Prof. Meixia Tao</b>	Shanghai Jiao Tong University, China
<i>Examiner</i>	<b>Prof. Daniela Tuninetti</b>	University of Illinois Chicago, U.S.A.
<i>Examiner</i>	<b>Prof. Giuseppe Caire</b>	Technical University of Berlin, Germany
<i>Examiner</i>	<b>Prof. Marios Kountouris</b>	EURECOM, France
<i>Thesis Advisor</i>	<b>Prof. Petros Elia</b>	EURECOM, France



# Systèmes de Liaison Descendante Assistés par Cache Hautes Performances: Nouveaux Algorithmes et Analyses

Thèse

*soumise à*

Sorbonne Université

*pour l'obtention du Grade de Docteur*

*présentée par:*

**Hui ZHAO**

*Soutenance de thèse prévue le 16 Décembre 2022, devant le jury composé de:*

<i>Examineur/Rapporteur</i>	<b>Prof. Antti Tölli</b>	University of Oulu, Finlande
<i>Examineur/Rapporteur</i>	<b>Prof. Meixia Tao</b>	Shanghai Jiao Tong University, Chine
<i>Examineur</i>	<b>Prof. Daniela Tuninetti</b>	University of Illinois Chicago, États-Unis
<i>Examineur</i>	<b>Prof. Giuseppe Caire</b>	Technical University of Berlin, Allemagne
<i>Examineur</i>	<b>Prof. Marios Kountouris</b>	EURECOM, France
<i>Directeur de Thèse</i>	<b>Prof. Petros Elia</b>	EURECOM, France



# Abstract

In this thesis, we focus on the design and analysis of physical layer (PHY) coded caching schemes for content delivery in realistic wireless networks. The main aim of this thesis is to show that cache-aided PHY techniques can indeed be used to substantially boost the performance of existing cutting-edge wireless systems. To do this, we employ advanced analytical methods inspired by random-matrix theory, as well as provide novel solutions that substantially ameliorate some of the bottlenecks that have been known to diminish cache-aided gains in wireless settings.

The thesis first addresses the worst-user bottleneck of wireless coded caching, which is known to severely diminish cache-aided multicasting gains due to the fundamental worst-channel limitation of multicasting. We consider the quasi-static Rayleigh fading Broadcast Channel, for which we first evaluate the exact effective coded caching gain of the standard XOR-based coded caching scheme for any finite SNR value. The result reveals that this effective (real) gain is now very diminished, and that it in fact completely vanishes in the low SNR regime. Then though we present a novel scheme that we refer to as the *aggregated coded caching* (ACC) scheme, which can fully recover the coded caching gains by capitalizing on one aspect that has remained unexploited to date: the shared side information brought about by the effectively unavoidable file-size constraint. By analyzing this scheme, we reveal that — under practical considerations — this collapse is not intrinsic to coded caching, and that in fact, due to the ACC scheme, the worst-user effect is dramatically ameliorated, as it is substituted by a much softer worst-group-of-users effect, where the suggested grouping is fixed, and it is decided before the channel or the user-demands are known. This grouping requires no additional overhead or assumptions. Our analysis provides achievable rates expressions for the ACC, and derives approximations which prove to be extremely precise. Importantly, this novel ACC approach can be translated to other coded caching schemes and scenarios, including decentralized scenarios.

We further proceed to provide rigorous analysis of the performance of coded caching — both for the traditional so called MN approach, as well as for the new ACC invented method — in various realistic scenarios. Toward this we investigate the delivery performance of ACC in a single dense urban micro/macro cell, where the system parameters for wireless propagation are set according to 3GPP standards. Our analysis shows that ACC-aided delivery provides the equivalent of a spatial-averaging effect, thereby recovering most of the theoretic (nominal) coded caching gains. We also consider ACC in land mobile satellite (LMS) systems, where a satellite serves a set of terrestrial users. ACC

plays a particularly crucial role here because such LMS systems incur large distances, heavy shadowing and thus low SNR, which, in the absence of ACC, would severely exacerbate the worst-user bottleneck. By considering the widely adopted Rician-Shadowed fading model (or even the more general mixture Gamma (MG) distribution model) for the satellite-terrestrial channel, we quantify the aforementioned spatial averaging effect and reveal how ACC-aided delivery can recover the majority of the theoretical gains. The effective rates and effective gains derivations for the aforementioned broad class of channels, nicely reveal that our ACC approach covers substantial ground in better establishing the direct utility of coded caching in realistic single-stream wireless networks, both in cellular as well as wireless settings.

The thesis then transitions to scenarios with transmitters with multi-antenna arrays, where such arrays are rightfully recognized as one of the most valuable resources in modern networks. In particular, we now consider the multi-antenna cache-aided multi-user (MU) scenario, where the multi-antenna transmitter delivers coded caching streams (for example MN-based XORs, or ACC based signals), thus being able to serve multiple users at a time, with a reduced (or just one) radio frequency (RF) chains. By doing so, coded caching can assist a simple analog (AG) beamformer (only a single RF chain), thus incurring considerable power and hardware savings that are particularly useful in various scenarios such as the mmWave setting. This is in contrast to conventional cacheless (multiplexing gain) methods which generally require multiple RF chains for hybrid/full-digital (FD) precoding to multiple users. As we see, the aforementioned worst-user bottleneck will persist in MN-based coded caching, and for this we will apply ACC-aided AG beamforming, *even without any beamforming optimization* to achieve the same delivery performance of the cacheless FD precoding (e.g., with zero-forcing, ZF) in realistic wireless networks corresponding to realistic mmWave channel assumptions and RF-chain requirements. Interestingly we will see that in the context of ACC-aided MU multicasting, and in the presence of sufficiently many users, a single-antenna transmitter effectively matches the performance of a similar system with a multi-antenna transmitter.

Finally, but perhaps most importantly, after removing the RF-chain limitation (perhaps being more in line with the assumptions in sub-6GHz bands), the thesis studies the properties and performance of the so-called vector coded caching technique, and reveals that this technique can achieve, under several realistic assumptions, a multiplicative sum-rate boost *over the optimized* cacheless multi-antenna counterpart that typically focuses on yielding a proper mixture of multiplexing and beamforming gains. In this thesis, we modify vector coded caching to account for the PHY properties, and analyze — under the assumption of symmetric Rayleigh fading channels — the corresponding sum-rate and effective vector coded caching gains (over optimized multi-antenna systems) with the help of large random matrix theory. In particular, for a given downlink MISO system already optimized to exploit both multiplexing and beamforming gains, and for a fixed set of antenna and SNR resources, our analysis answers a simple question: What is the multiplicative throughput boost obtained from introducing reasonably-sized receiver-side caches that can pre-store information content? The schemes are very simple (we simply collapse precoding vectors into a single vector), and the recorded gains are notable. For example, for 32 transmit antennas, a received SNR of 20 dB, a coher-

ence bandwidth of 300 kHz, a coherence period of 40 ms, and under realistic file-size and cache-size constraints, vector coded caching is here shown to offer a multiplicative throughput boost of about 310% with ZF (or Regularized ZF) precoding and a 430% boost in the performance of already optimized Matched Filtering (MF) based (cacheless) systems. This is after accounting for CSI costs. We further investigate the performance of vector coded caching aided MU-MIMO systems under more realistic considerations that include CSI costs but also variable path-loss, max-min fairness and the presence of multi-antenna receivers, where again the aforementioned large gains are maintained especially in micro-cell scenarios.



# Abrégé

Dans cette thèse, nous nous concentrons sur la conception et l'analyse de schémas de cache codés de la couche physique (PHY) pour la diffusion de contenu dans des réseaux sans fil en réalité. L'objectif principal de cette thèse est de montrer que les techniques PHY assistées par cache peuvent augmenter considérablement les performances des systèmes sans fil de pointe. Pour ce faire, nous utilisons des méthodes analytiques avancées inspirées de la théorie des matrices aléatoires, ainsi que des solutions novatrices qui améliorent considérablement certains des goulots d'étranglement connus pour diminuer les gains assistés par le cache dans les paramètres sans fil.

La thèse aborde d'abord le pire goulot d'étranglement pour les utilisateurs de cache codée sans fil, qui est connue pour diminuer considérablement les gains de multidiffusion assistée par cache en raison de la limitation fondamentale de la multidiffusion sur les pires canaux. Nous considérons le canal de diffusion quasi-statique à évanouissement de Rayleigh, pour lequel nous évaluons d'abord le gain de cache codé effectif exact du schéma de cache codé basé sur XOR standard pour toute valeur SNR finie. Le résultat révèle que ce gain effectif (réel) est maintenant très diminué, et qu'il s'annule en fait complètement dans le régime bas SNR. Ensuite, bien que nous présentions un nouveau schéma que nous appelons le schéma de cache codée agrégée (ACC), qui peut récupérer entièrement les gains de cache codée en capitalisant sur un aspect qui est resté inexploité à ce jour : les informations secondaires partagées apportées par la contrainte de taille de fichier inévitable. En analysant ce schéma, nous révélons que — sous des considérations pratiques — cet effondrement n'est pas intrinsèque à la cache codée, et qu'en fait, grâce au schéma ACC, l'effet du pire utilisateur est considérablement amélioré, car il est remplacé par un beaucoup plus doux l'effet du pire groupe d'utilisateurs, où le groupement suggéré est fixé, et il est décidé avant que le canal ou les demandes des utilisateurs ne soient connus. Ce regroupement ne nécessite aucune surcharge ou hypothèse supplémentaire. Notre analyse fournit des expressions de taux réalisables pour l'ACC et en déduit des approximations qui s'avèrent extrêmement précises. Il est important de noter que cette nouvelle approche ACC peut être traduite dans d'autres schémas et scénarios de cache codée, y compris des scénarios décentralisés.

Nous procédons ensuite à une analyse rigoureuse des performances de la cache codée — à la fois pour l'approche traditionnelle dite MN, ainsi que pour la nouvelle méthode inventée par ACC — dans divers scénarios réalistes. Dans cette optique, nous étudions les performances de livraison de l'ACC dans une seule cellule micro/macro urbaine dense, où les paramètres du système pour la propagation sans fil sont définis conformément aux

---

normes 3GPP. Notre analyse montre que la livraison assistée par ACC fournit l'équivalent d'un effet de moyenne spatiale, récupérant ainsi la plupart des gains théoriques (nominaux) de cache codée. Nous considérons également l'ACC dans les systèmes mobiles terrestres par satellite (LMS), où un satellite dessert un ensemble d'utilisateurs terrestres. L'ACC i Abstrait joue ici un rôle particulièrement crucial car de tels systèmes LMS entraînent de grandes distances, une forte ombrage et donc un faible SNR, ce qui, en l'absence d'ACC, aggraverait gravement le goulot d'étranglement des pires utilisateurs. En considérant le modèle d'évanouissement Rician-Shadowed largement adopté (ou même le modèle de distribution Gamma (MG) de mélange plus général) pour le canal satellite-terrestre, nous quantifions l'effet de moyenne spatiale susmentionné et révélons comment la livraison assistée par ACC peut récupérer la majorité des gains théoriques. Les taux effectifs et les dérivations des gains effectifs pour la large classe de canaux susmentionnée révèlent bien que notre approche ACC couvre un terrain substantiel en établissant mieux l'utilité directe de la cache codée dans des réseaux sans fil à flux unique réalistes, à la fois dans les paramètres cellulaires et sans fil.

La thèse passe ensuite à des scénarios avec des émetteurs avec des réseaux multi-antennes, où ces réseaux sont à juste titre reconnus comme l'une des ressources les plus précieuses des réseaux modernes. En particulier, nous considérons maintenant le scénario multi-utilisateurs assisté par cache multi-antennes (MU), où l'émetteur multi-antennes fournit des flux de cache codés (par exemple des XOR basés sur MN ou des signaux basés sur ACC), pouvant ainsi servir plusieurs utilisateurs. à la fois, avec une chaîne radiofréquence (RF) réduite (ou une seule). Ce faisant, la cache codée peut aider un simple formateur de faisceau analogique (AG) (une seule chaîne RF), entraînant ainsi des économies considérables d'énergie et de matériel qui sont particulièrement utiles dans divers scénarios tels que le réglage mmWave. Cela contraste avec les méthodes conventionnelles sans cache (gain de multiplexage) qui nécessitent généralement plusieurs chaînes RF pour le précodage hybride/entièrement numérique (FD) pour plusieurs utilisateurs. Comme nous le voyons, le goulot d'étranglement du pire utilisateur mentionné ci-dessus persistera dans la cache codée basée sur MN, et pour cela, nous appliquerons la formation de faisceau AG assistée par ACC, *même sans aucune optimisation de beamforming* de faisceau pour obtenir les mêmes performances de livraison que le précodage FD sans cache (par exemple, avec forçage zéro, ZF) dans des réseaux sans fil réalistes correspondant à des hypothèses réalistes de canal mmWave et aux exigences de la chaîne RF. De manière intéressante, nous verrons que dans le contexte de la multidiffusion MU assistée par ACC, et en présence d'un nombre suffisant d'utilisateurs, un émetteur à antenne unique correspond efficacement aux performances d'un système similaire avec un émetteur à plusieurs antennes.

Enfin, mais peut-être le plus important, après avoir supprimé la limitation de la chaîne RF (peut-être plus conforme aux hypothèses dans les bandes inférieures à 6 GHz), la thèse étudie les propriétés et les performances de la technique de cache à codage vectoriel, et révèle que cette technique peut atteindre, sous plusieurs hypothèses réalistes, une augmentation multiplicative du taux de somme par rapport à la contrepartie multi-antenne optimisée sans cache qui se concentre généralement sur la production d'un mélange approprié de gains de multiplexage et de formation de faisceaux. Dans cette

thèse, nous modifions la cache à codage vectoriel pour tenir compte des propriétés PHY et analysons — sous l’hypothèse de canaux d’évanouissement de Rayleigh symétriques — le taux de somme correspondant et les gains de cache à codage vectoriel efficaces (sur des systèmes multi-antennes optimisés) avec l’aide de la théorie des grandes matrices aléatoires. En particulier, pour un système MISO de liaison descendante déjà optimisé pour exploiter à la fois les gains de multiplexage et de formation de faisceaux, et pour un ensemble fixe d’antennes et de ressources SNR, notre analyse répond à une question simple : Quelle est l’augmentation de débit multiplicative obtenue en introduisant des caches côté récepteur de taille raisonnable qui peuvent pré-stocker le contenu des informations ? Les schémas sont très simples (on regroupe simplement les vecteurs de précodage en un seul vecteur), et les gains enregistrés sont notables. Par exemple, pour 32 antennes d’émission, un SNR reçu de 20 dB, une largeur de bande d’abstraction de cohérence de 300 kHz, une période de cohérence de 40 ms, et sous des contraintes réalistes de taille de fichier et de taille de cache, la cache à codage vectoriel est montrée ici pour offrir une augmentation de débit multiplicative d’environ 310% avec le précodage ZF (ou ZF régularisé) et une augmentation de 430% des performances des systèmes déjà optimisés basés sur le filtrage adapté (MF) (sans cache). C’est après comptabilisation des coûts de CSI. Nous étudions plus en détail les performances des systèmes MU-MIMO assistés par cache à codage vectoriel sous des considérations plus réalistes qui incluent les coûts CSI, mais également la perte de chemin variable, l’équité max-min et la présence de récepteurs multi-antennes, là encore les gains importants susmentionnés sont maintenus. en particulier dans les scénarios de micro-cellules.



# Acknowledgements

This thesis has been the effort of approximately three and a half years and has flourished only because of the many people that have contributed directly, by being part of these works, or indirectly, through their support during this process.

First and foremost, I would like to express my gratitude to my thesis advisor Prof. Petros Elia. His constant strive for excellence, his vision and his patience have helped me improve my skills and taught me to always look for the underlying meaning of things. I am deeply grateful to Antonio Bazco-Nogueras and Eleftherios Lampiris, who help my advisor to conduct my research. I really appreciate their profound technical knowledge, in conjunction with their human kindness and comprehension. I would also like to thank Prof. Christoph Studer for hosting me at Swiss Federal Institute of Technology in Zürich, Switzerland, as well as to the rest of his group for their warm hospitality and the enlightening technical discussions. I am also very grateful to Prof. Mohamed-Slim Alouini and Prof. Gaofeng Pan who were my supervisors during the master's and undergraduate student periods respectively. I really appreciate their mentorship during my early stage of research, which helped me make a solid foundation for my PhD studying.

Last but not least, I would like to express my heartfelt gratitude to my family, for their endless love and continuous far-reaching support, and to Shijian, for her support, patience, and love.



# Contents

Abstract . . . . .	i
Abrégé [Français] . . . . .	v
Acknowledgements . . . . .	ix
Contents . . . . .	xi
List of Figures . . . . .	xiv
List of Tables . . . . .	xix
Acronyms . . . . .	xxi
Notations . . . . .	xxiii
<b>1 Introduction</b>	<b>1</b>
1.1 Two Fundamental Bottlenecks in Coded Caching . . . . .	3
1.1.1 Subpacketization Bottleneck and the Need for Shared Caches . . . . .	3
1.1.2 Worst-User Bottleneck: Motivation, Nature of the Problem, and Prior Work . . . . .	4
1.2 Multi-Antenna Coded Caching . . . . .	5
1.2.1 Multi-Antenna Coded Multicasting . . . . .	6
1.2.2 Vector Coded Caching . . . . .	7
1.3 Thesis Outline and Main Contributions . . . . .	8
1.3.1 Main Contributions Toward Resolving Worst-User Bottleneck . . . . .	8
1.3.2 Main Contributions in Vector Coded Caching . . . . .	14
1.3.3 Main Contributions in Land Mobile Satellite Systems . . . . .	19
<b>2 Aggregated Coded Caching: Design and Analysis</b>	<b>23</b>
2.1 System Model and Problem Definition . . . . .	23
2.2 Aggregated Coded-Caching Scheme . . . . .	24
2.2.1 Aggregated Coded-Caching Design . . . . .	25
2.3 Average Rate Analysis . . . . .	28
2.3.1 Average Rate of the $\Lambda$ -MN and ACC Schemes . . . . .	28
2.3.2 Rate Approximations and Effective Gains at Low SNR . . . . .	31
2.3.3 Effective Gain in the Large- $B$ Region . . . . .	33
2.3.4 High-Fidelity Approximation of $\bar{R}^{\text{ACC}}$ for Any SNR Value . . . . .	34

2.3.5	Numerical Results for the ACC scheme . . . . .	35
2.4	Extension A: Delivery Time Analysis of the ACC scheme . . . . .	42
2.4.1	Delivery Time of the MN Scheme at Low SNR . . . . .	43
2.4.2	Delivery Time of the ACC Scheme at Low SNR . . . . .	46
2.4.3	Numerical Results . . . . .	48
2.5	Extension B: ACC Performance in Ergodic-Fading Scenario With Different Pathloss . . . . .	50
2.5.1	System Model . . . . .	52
2.5.2	MN Scheme in the Single-Cell Scenario . . . . .	53
2.5.3	ACC Scheme in the Single-Cell Scenario . . . . .	55
2.5.4	Numerical Results . . . . .	57
2.6	Extension C: Multi-Antenna ACC . . . . .	58
2.6.1	Introduction . . . . .	59
2.6.2	Multi-antenna Cache-aided Coded Multicasting (Multi-antenna MN) . . . . .	60
2.6.3	ACC-Aided Multicasting . . . . .	62
2.6.4	Cacheless Full-Digital Precoding . . . . .	63
2.6.5	Performance Metric . . . . .	64
2.6.6	Numerical Results . . . . .	65
2.7	Conclusions . . . . .	70
<b>3</b>	<b>Vector Coded Caching: Design and Analysis</b> . . . . .	<b>73</b>
3.1	System Model and Problem Description . . . . .	73
3.1.1	System Model . . . . .	73
3.1.2	Signal-Level Vector Coded Caching for Finite SNR . . . . .	74
3.1.3	Vector Coded Caching for the Physical Layer . . . . .	77
3.2	Analysis of the Average Rate and of the Effective Gain over MISO . . . . .	78
3.2.1	MF Precoding . . . . .	78
3.2.2	ZF Precoding . . . . .	79
3.2.3	RZF Precoding . . . . .	80
3.2.4	Accounting for the CSI Costs . . . . .	81
3.2.5	Effective Gains over Cacheless MISO Systems . . . . .	82
3.3	Optimizing Physical Layer Vector Coded Caching . . . . .	82
3.4	Numerical Results . . . . .	84
3.5	Conclusions . . . . .	88
<b>4</b>	<b>More Practical Considerations in Vector Coded Caching</b> . . . . .	<b>91</b>
4.1	System Model and Problem Description . . . . .	92
4.1.1	BD Precoding and MRC Combining . . . . .	93
4.1.2	Main Performance Metrics . . . . .	96

4.2	BD-MRC Analysis for Multi-Antenna Receivers . . . . .	96
4.2.1	Effective Sum-Rate and Effective Gain: the case of BD-MRC . . . . .	96
4.2.2	Special Case (i): Massive MIMO Regime Over Rayleigh Fading Channels . . . . .	98
4.2.3	Special Case (ii): Single-Antenna Receivers in BD . . . . .	99
4.3	ZF Precoding Analysis for Multi-Antenna Receivers . . . . .	99
4.4	Numerical Results . . . . .	102
4.5	Conclusions . . . . .	104
<b>5</b>	<b>ACC-aided Land Mobile Satellite System</b>	<b>109</b>
5.1	Introduction . . . . .	109
5.2	System Model . . . . .	110
5.2.1	LMS Channel Model . . . . .	111
5.2.2	Adopted Assumptions and Preliminaries . . . . .	112
5.3	Average Rate and Effective Gain of the MN Scheme . . . . .	113
5.4	Average Rate and Effective Gain of the ACC Scheme . . . . .	115
5.4.1	Exact Analytical Expression . . . . .	115
5.4.2	Large-User Approximation . . . . .	116
5.4.3	Extension to General Fading Channels . . . . .	118
5.5	Numerical Results . . . . .	121
5.6	Conclusions . . . . .	123
<b>6</b>	<b>Conclusions and Perspectives</b>	<b>129</b>
	<b>Appendices</b>	<b>133</b>
<b>A</b>	<b>Proofs in Chapter 2</b>	<b>135</b>
A.1	Capacity Region of Proposition 2.1 . . . . .	135
A.2	Proof of Lemma 2.1 . . . . .	135
A.3	Proofs for Section 2.3.2 and Section 2.3.3 . . . . .	137
A.3.1	Proof of Lemma 2.2 . . . . .	137
A.3.2	Proof of Proposition 2.2 . . . . .	137
A.3.3	Proof of Lemma 2.3 . . . . .	137
A.3.4	Proof of Corollary 2.1 . . . . .	138
A.3.5	Proof of Lemma 2.4 . . . . .	138
A.3.6	Proof of Lemma 2.5 . . . . .	139
A.4	Proof of Lemma 2.6 . . . . .	140
A.5	Proof of Lemma 2.9 . . . . .	141
A.6	Proof of Lemma 2.14 . . . . .	142
A.7	Proof of Lemma 2.15 . . . . .	143

<b>B</b>	<b>Proofs in Chapter 3</b>	<b>145</b>
B.1	Proof of Theorem 3.1 . . . . .	145
B.2	Proof of Theorem 3.3 . . . . .	146
B.2.1	Two Useful Lemmas . . . . .	147
B.2.2	Proof of Theorem 3.3 . . . . .	148
B.3	Proof of Theorem 3.4 . . . . .	150
B.4	Proof of Theorem 3.5 . . . . .	150
 <b>C</b>	 <b>Proofs in Chapter 4</b>	 <b>153</b>
C.1	Proof of Lemma 4.1 . . . . .	153
C.2	Proof of Lemma 4.2 . . . . .	154
C.3	Proof of Lemma 4.3 . . . . .	155
C.4	Proof of Proposition 4.1 . . . . .	156
 <b>D</b>	 <b>Proofs in Chapter 5</b>	 <b>159</b>
D.1	Proof of Proposition 5.1 . . . . .	159
D.2	Proof of Lemma 5.1 . . . . .	160
D.3	Proof of Lemma 5.2 . . . . .	160
D.4	Proof of Lemma 5.3 . . . . .	161
D.5	Proof of Lemma 5.4 . . . . .	162
D.6	Proof of Lemma 5.5 . . . . .	163

# List of Figures

1.1	Placement phase in MN Coded Caching . . . . .	2
1.2	Delivery phase in MN Coded Caching . . . . .	2
1.3	Ratio between the average rates of the MN scheme and TDM (i.e., the <i>effective coded-caching gain</i> ) over quasi-static Rayleigh fading channel for different values of $G = \Lambda\gamma + 1$ . . . . .	6
1.4	An example of ACC delivery. . . . .	9
2.1	Comparison of MN and ACC for a nominal coded-caching gain of 3. . . . .	26
2.2	The ACC improvement $\left(\frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}}\right)$ over the MN scheme in Corollary 2.1 for different $B$ and $G$ . . . . .	33
2.3	Effective gain versus $\rho$ for $G = 10$ . Right-side plot focuses on realistic SNR values. . . . .	36
2.4	Effective gain versus $\rho$ for $B = 6$ . Right-side plot focuses on realistic SNR values. . . . .	36
2.5	$\bar{R}^{\text{ACC}}$ versus $\rho$ for $G = 4$ . . . . .	37
2.6	$\bar{R}^{\text{ACC}}$ versus $\rho$ for $B = 3$ . . . . .	38
2.7	$\bar{R}^{\text{ACC}}$ versus $B$ for $\rho = 0$ dB. . . . .	39
2.8	$\bar{R}^{\text{ACC}}$ versus $B$ for $G = 10$ . . . . .	40
2.9	$\bar{R}^{\text{ACC}}/\bar{R}^{\text{MN}}$ versus $\rho$ for $V = 7$ in GHQ. . . . .	40
2.10	$\bar{R}^{\text{ACC}}/\bar{R}^{\text{MN}}$ versus $\rho$ for $G = 4$ . . . . .	41
2.11	Comparison of Delivery Time versus $\rho$ for $m = 2$ in the decentralized-placement scenario. . . . .	42
2.12	Average delivery time of the MN (left) and ACC (right) schemes versus $\rho$ for $K = 300$ , $B = 5$ , $\gamma = 5\%$ and $V = 7$ in the GHQ and GLQ. . . . .	49
2.13	Effective coded caching gains of the ACC and MN schemes for $K = 300$ , $B = 5$ , $\gamma = 5\%$ , and $m = 3, 4$ . . . . .	50
2.14	Delay Ratio of $\hat{T}_{\text{MN}}$ over $\hat{T}_{\text{ACC}}$ versus $\rho$ for $m = 3$ , $\gamma = \frac{1}{12}$ , and $G = 6$ . . . . .	51
2.15	Effective coded caching gains of MN (left) and ACC (right) schemes for $B = 10$ and $\gamma = 10\%$ as $\rho \rightarrow 0$ . . . . .	51
2.16	Effective coded caching gains of MN (left) and ACC (right) schemes for $G = 5$ and $\gamma = 5\%$ as $\rho \rightarrow 0$ . . . . .	52
2.17	Performance in a dense urban Micro-Cell, for the case of $K = 800$ , $\gamma = 5\%$ , and $\Lambda = 80$ (edge SNR: $15 \rightarrow 35\text{dB}$ ). . . . .	57

2.18	Performance in a dense urban Macro-Cell for the case of $K = 2000$ , $\gamma = 5\%$ , and $\Lambda = 80$ (edge SNR: $7 \rightarrow 35$ dB). . . . .	58
2.19	Performance in the Macro-Cell setting with $\gamma = 10\%$ and $P_t = 40$ dBm. $\mathcal{H}_G$ is computed through GHQ with 10 terms. . . . .	59
2.20	Multicasting using a single-RF-chain analog (AG) beamformer (left), and multicasting using an $L$ -RF-chain full digital (FD) beamformer (right). . . . .	61
2.21	Delivery performance over symmetric Rayleigh fading channels. . . . .	66
2.22	Delivery performance over symmetric Rayleigh fading channels. . . . .	67
2.23	Delivery performance over non-symmetric Rayleigh fading channels with $D_1 = 35$ , $D_2 = 500$ , $L = 128$ , $G = 6$ , $\eta_0 = 3.76$ , and $l_0 = 10^{-3.53}$ . . . . .	67
2.24	Delivery performance over non-symmetric Rayleigh fading channels with $P_t = 40$ dBm, $D_1 = 35$ , $D_2 = 500$ , $L = 128$ , $G = 6$ , $\eta_0 = 3.76$ , and $l_0 = 10^{-3.53}$ . . . . .	68
2.25	Delivery performance over non-symmetric Rayleigh fading channels with $D_1 = 35$ , $D_2 = 500$ , $\eta_0 = 3.76$ , and $l_0 = 10^{-3.53}$ . . . . .	68
2.26	Delivery performance over non-symmetric Rayleigh fading channels with $D_1 = 35$ , $D_2 = 500$ , $\eta_0 = 3.76$ , and $l_0 = 10^{-3.53}$ . . . . .	69
2.27	Delivery performance over non-symmetric Rayleigh fading channels with $D_1 = 35$ , $D_2 = 500$ , $L = 128$ , $G = 6$ , $\eta_0 = 3.76$ , and $l_0 = 10^{-3.53}$ . . . . .	69
2.28	Delivery performance over non-symmetric Rayleigh fading channels with $D_1 = 10$ , $D_2 = 100$ , $L = 32$ , $G = 6$ , $\eta_0 = 3$ , and $l_0 = 10^{-3.7}$ . . . . .	71
2.29	Delivery performance over mmWave channels with $L = 32$ , $G = 6$ , $B = 5$ , $B_w = 200$ MHz, $N_p = 2$ and $F_{\text{sub}} = 50$ bytes. . . . .	71
3.1	An example of vector coded caching with $G = 3$ , $\Lambda = 40$ , $B = 4$ and $Q = 2$ . . . . .	76
3.2	Effective rate $\bar{\mathcal{R}}$ and optimized effective rate for $P_t = 10$ dB and $G = 5$ . . . . .	85
3.3	Effective gain $\mathcal{G}^*$ over optimized cacheless system for $L \in \{32, 64\}$ and $G = 6$ . . . . .	85
3.4	Hardening-constrained effective gain over a constrained classical downlink system. $Q$ is fixed for both systems at $Q = 8$ , while $G = 6$ . . . . .	86
3.5	Effective gain versus $\Gamma$ in medium SNR (10 dB) and high SNR (30 dB). . . . .	86
4.1	Effective rate versus $P_{\text{tot}}$ in the Macro-cell setting for $J = M$ with equal-power allocation. . . . .	103
4.2	Effective rate versus $P_{\text{tot}}$ in the Macro-cell setting under MMF. . . . .	104
4.3	Effective rate versus $P_{\text{tot}}$ in a Macro-cell under MMF. . . . .	105
4.4	Effective rate versus $P_{\text{tot}}$ in a Macro-cell under MMF. . . . .	106
4.5	Optimal effective gain versus $P_{\text{tot}}$ in a Macro-cell under MMF. . . . .	106
4.6	Optimal effective gain versus $P_{\text{tot}}$ in a Macro-cell under MMF. . . . .	107
4.7	Delivery Performance of vector coded caching in a Micro-cell . . . . .	107
5.1	Effective gain of MN coded caching versus $G$ in low-SNR limit. . . . .	115
5.2	Effective coded caching gain versus $\rho$ for $G = 5$ in the average shadowing case. . . . .	122
5.3	Effective coded caching gain versus $\rho$ for $G = 6$ and $B = 20$ . . . . .	123

*List of Figures*

---

5.4	Average rate of the MN scheme versus $\rho$ for $G = 4$ . . . . .	124
5.5	Average rate of the uncoded TDM scheme versus $\rho$ . . . . .	125
5.6	Average rate of the ACC scheme versus $\rho$ for $G = 6$ and $B = 20$ . . . . .	125
5.7	Average rate of the ACC versus $\rho$ for $G = 5$ in the average shadowing case.	126
5.8	Average rate of the ACC versus $\rho$ for $B = 20$ in the average shadowing case.	126
5.9	Effective gain of the ACC versus $\rho$ for $G = 6$ in the average shadowing case.	127



# List of Tables

2.1	Exact Value of $\mathcal{H}_G$ for $G \leq 5$ . . . . .	35
3.1	Derived Theorems (Thms.) and Corollaries (Cors.) in Chapter 3 . . . . .	84
4.1	User Percentage and Receiver SNR Range in a Macro Cell with $D_1 = 35$ , $D_2 = 500$ , $\eta = 3.76$ , and $l_0 = 10^{-3.53}$ . . . . .	104
5.1	Common Fading Models in Wireless Channels . . . . .	119
5.2	Parameters of Fading Models Listed in Table 5.1 . . . . .	119
5.3	Parameters in MG Distribution . . . . .	120
5.4	Typical Fading Scenarios in LMS Channel . . . . .	121



# Acronyms and Abbreviations

The acronyms and abbreviations used throughout the manuscript are specified in the following. They are presented here in their singular form, and their plural forms are constructed by adding and *s*, e.g. TX (transmitter) and TXs (transmitters). The meaning of an acronym is also indicated the first time that it is used.

ACC	Aggregated Coded Caching.
AG	analog.
<i>a.s.</i>	Almost Sure Convergence.
AWGN	Additive White Gaussian Noise.
BC	Broadcast Channel.
BD	Block-diagonalization.
BPSK	Binary Phase-Shift Keying.
BS	Base Station.
CDF	Cumulative Distribution Function.
CLT	Central Limit Theorem.
CSI	Channel State Information.
CF	Characteristic Function.
<i>d.</i>	Convergence in Distribution.
DPC	Dirty Paper Coding.
DoF	Degrees of Freedom.
FD	Full Digital.
GHQ	Gauss-Hermite Quadrature.
GLQ	Gauss-Laguerre Quadrature.
i.i.d.	independent and identically distributed.
LMS	Land Mobile Satellite.
LOS	Line of Sight.
MMF	Max-Min Fairness.
mMTC	Machine-Type Communication.
MF	Matched Filtering
MG	Mixture Gamma.
MN	Maddah-Ali and Niesen.
MRC	Maximal Ratio Combining.
MIMO	Multiple Input Multiple Output.

MISO	Multiple Input Single Output.
MU	Multi-User.
MU-MIMO	Multi-User Multiple Input Multiple Output.
MU-MISO	Multi-User Multiple Input Single Output.
NL	Nakagami-Lognormal.
PDF	Probability Density Function.
PZF	Phased Zero-Forcing.
QAM	Quadrature Amplitude Modulation.
RF	Radio Frequency.
RHS	Right hand side.
RZF	Regularized Zero-Forcing.
SIMO	Single Input Multiple Output.
SINR	Signal to Noise and Interference Ratio.
SNR	Signal-to-Noise Ratio.
SISO	Single-Input Single-Output.
s.t.	such that.
TDD	Time Division Duplex.
TDM	Time Division Multiplexing.
TDMA	Time Division Multiple Access.
ULA	Uniform Linear Array
ZF	Zero Forcing.

# Notations

The next list describes an overview on the notation used throughout this manuscript. We use boldface uppercase letters ( $\mathbf{A}$ ) for matrices, boldface lowercase letters for vectors ( $\mathbf{a}$ ), and regular lowercase letters for scalars ( $a$ ).

## Algebraic Notation

$(a)^+$	Maximum value between $a$ and zero.
$[a]$	Set of first $a$ natural numbers $\{1, \dots, a\}$ for positive integer $a$ .
$\oplus$	Bit-wise XOR operator.
$ a $	Magnitude of the complex number $a$ .
$\ \mathbf{a}\ $	Norm-2 of vector $\mathbf{a}$ .
$ \Omega $	Cardinality of set $\Omega$ .
$\mathbb{E}\{\cdot\}$	Expectation operator.
$\mathbb{E}_h\{\cdot\}$	Expectation over channel states.
$\mathbb{E}_r\{\cdot\}$	Expectation over random locations.
$\mathbb{E}_{h,r}\{\cdot\}$	Expectation over both channel states and random locations.
$\mathbb{I}\{\cdot\}$	Indicator function.
$\max\{\cdot, \dots, \cdot\}$	Maximum of the included elements.
$\min\{\cdot, \dots, \cdot\}$	Minimum of the included elements.
$\Omega \setminus \Phi$	Difference set between set $\Omega$ and set $\Phi$ .
$F_X(\cdot)$	Cumulative distribution function of random variable $X$ .
$f_X(\cdot)$	Probability density function of random variable $X$ .
$\mathbf{A} \circ \mathbf{B}$	Hadamard product of matrix $\mathbf{A}$ and matrix $\mathbf{B}$ .
$\mathbf{A}^*$	Conjugate of matrix $\mathbf{A}$ .
$\mathbf{A}^H$	Conjugate transpose of matrix $\mathbf{A}$ .
$\mathbf{A}^T$	Non-conjugate transpose of matrix $\mathbf{A}$ .
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$ .
$\binom{\cdot}{\cdot, \dots, \cdot}$	Multinomial coefficient.
$\binom{\cdot}{\cdot}$	Binomial coefficient.
$\text{Im}\{a\}$	Imaginary part of complex number $a$ .
$\text{Pr}\{\cdot\}$	Probability of the included event.
$\text{Rank}\{\mathbf{A}\}$	Rank of matrix $\mathbf{A}$ .
$\text{Tr}\{\mathbf{A}\}$	Trace of matrix $\mathbf{A}$ .

---

$\text{Var}\{\cdot\}$	Variance operator.
<b>System Symbols</b>	
$\aleph$	Average of the LOS amplitude power in Shadowed-Rician fading.
$\alpha$	Regulation factor of the regularized zero-forcing precoder.
$\beta$	Pathloss.
$\beta_{\text{tot}}$	Number of resources per user's antenna and per block used for pilot transmission.
$\eta_0$	Pathloss exponential factor.
$\gamma$	Cache size normalized to the library.
$j$	Imaginary unit which equals $\sqrt{-1}$ .
$\Lambda$	Number of dedicated cache states.
$\mathbb{C}$	Set of complex numbers.
$\mathcal{F}$	Library.
$\mathcal{G}$	Effective coded caching gain.
$\mathcal{H}_G$	Expectation of the maximum of $G$ i.i.d. standard normal random variables.
$\Psi$	Selected user-group set for service.
$\rho$	Average SNR.
$B$	Number of users caching the same content.
$b_0$	Half average value of the scatter amplitude power in Shadowed-Rician fading.
$B_w$	Spectrum bandwidth.
$C$	Euler–Mascheroni constant.
$c$	Ratio of the multiplexing gain to the number of transmit antennas.
$D_1$	Inner radius of a cell in meters.
$D_2$	Outer radius of a cell in meters.
$e$	Natural constant.
$F$	Total bits in each library file.
$f_{\text{GHz}}$	Carrier frequency in GHz.
$G$	Nominal gain in the high-SNR limit.
$K$	Total number of served users.
$L$	Number of transmit antennas.
$l_0$	Pathloss regulation factor.
$M$	Number of receive antennas.
$m$	Parameter $m$ in Nakagami- $m$ distribution.
$m_0$	Reflecting the (average) obstruction of the LOS component in Shadowed-Rician fading.
$N$	Total number of library files.
$N_0$	AWGN power.
$P_t$	Transmit power.

$P_{\text{tot}}$	Maximum allowable transmit power.
$Q$	Multiplexing gain.
$r$	Delivery distance in meters.
$T_c$	Coherence block time (in symbols).
$V$	Number of summation terms in Gaussian quadrature.
$W_c$	Coherence bandwidth.
$W_n$	$n$ -th library file.
$W_n^{\mathcal{T}}$	Subfile labelled by $\mathcal{T}$ of the $n$ -th library file.
$x_v, \omega_v$	$v$ -th sample point and the corresponding weight in Gauss-Hermite quadrature respectively.
$y_v, \chi_v$	$v$ -th sample point and the corresponding weight in Gauss-Laguerre quadrature respectively.
$\mathbf{0}_L$	Vector with all $L$ elements equaling zero.
$\mathbf{I}_L$	Identity matrix with dimensions $L \times L$ .
<b>Special Functions</b>	
$\Gamma(\cdot)$	Gamma function.
$\Gamma(\cdot, \cdot)$	Upper incomplete Gamma function.
$\mathcal{Q}(\cdot)$	Tail distribution function of the standard normal distribution.
$\mathcal{U}(\cdot, \cdot, \cdot)$	Confluent hypergeometric function of the second kind.
$\text{Ei}(\cdot)$	Exponential integral function.
$\text{E}_n(\cdot)$	Exponential integral function of order $n$ .
$G_{\cdot, \cdot, \cdot, \cdot}^{\cdot, \cdot, \cdot, \cdot}(\cdot, \cdot)$	Extended generalized bivariate Meijer's G-function (EGBMGF).
$G_{\cdot, \cdot}^{\cdot, \cdot}(\cdot)$	Meijer's G-function.
$\text{K}(\cdot)$	Modified Bessel function of the second kind.
${}_2F_1(\cdot; \cdot; \cdot)$	Generalized hypergeometric function.



# Chapter 1

## Introduction

Cache-aided communication is a promising approach toward reducing congestion in modern communication networks [1, 2]. The promise of this approach has been recently fostered by the seminal paper of Maddah-Ali and Niesen [1], who proposed *coded caching* as a means to speed up content delivery by exploiting receiver-side stored content to remove interference.

The work in [1] considers the error-free (or equivalently, the high-SNR) shared-link Broadcast Channel (BC), where a transmitter with access to a library of  $N$  content files serves  $K$  cache-aided users. Each such user enjoys a local (cache) memory of size equal to a fraction  $\gamma \in [0, 1]$  of the library size. The so-called *MN scheme* in [1] involves a cache placement phase and a subsequent delivery phase. During the first phase, each file is typically split into a very large number of subfiles, which are selectively placed in various different caches. During the second phase, the communication process is split into a generally large number of *transmission stages*, and at each such stage, a different subset of  $K\gamma + 1$  users is simultaneously served via an XOR multicast transmission, thus allowing for a theoretical speed-up factor (again in the infinite SNR regime) of  $K\gamma + 1$  as compared to the uncoded (effectively cacheless) case. This speed-up factor of  $K\gamma + 1$  is also referred to as the *coded caching gain* or the *degrees-of-freedom (DoF)* achieved by this scheme. In the following, let us elaborate on the placement phase and the delivery phase of this MN coded caching scheme.

**Placement Phase:** During the placement phase, each file  $W_n$  is partitioned into  $\binom{K}{K\gamma}$  non-overlapping and equal-sized segments (subfiles), such that

$$W_n \rightarrow \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [K], |\mathcal{T}| = K\gamma\},$$

where  $[K] \triangleq \{1, 2, \dots, K\}$ , and where  $|\mathcal{T}|$  denotes the cardinality of the set  $\mathcal{T}$ . User  $k$  stores all the segments  $W_n^{\mathcal{T}}$  such that  $k \in \mathcal{T}$ , for any  $n \in [N]$ . The content cached at user  $k$ , denoted by  $\mathcal{Z}_k$ , is hence given by

$$\mathcal{Z}_k = \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [K], |\mathcal{T}| = K\gamma, \mathcal{T} \ni k, \forall n \in [N]\}.$$

This placement phase is also illustrated in Fig. 1.1. It follows that the total content cached at each user amounts to  $\gamma NF$  bits, which satisfies the local cache size constraint.

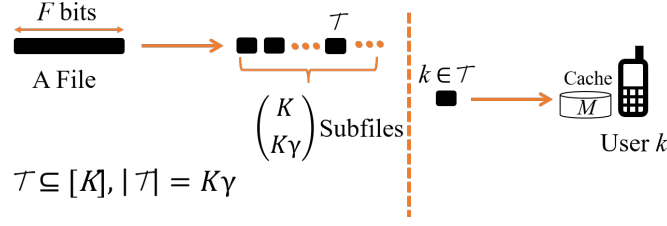


Figure 1.1: Placement phase in MN Coded Caching

**Delivery Phase:** At the beginning of this phase (cf. Fig. 1.2), each user requests a different file  $W_{d_k}$  from the library, where  $d_k \in [N]$  denotes the index of the file demanded by user  $k \in [K]$ . The transmission is divided in different transmission stages. At each transmission stage, the transmitter simultaneously serves a unique set of  $K\gamma + 1$  users. Since there are  $\binom{K}{K\gamma+1}$  different subsets of  $K\gamma + 1$  users in  $[K]$ , the delivery phase consists of  $\binom{K}{K\gamma+1}$  transmission stages, and at each stage the transmitter serves a different subset of users  $\Psi \subseteq [K]$  of  $|\Psi| = K\gamma + 1$  users. Specifically, for the transmission stage intended for a particular set of users  $\Psi$ , the transmitted signal is designed as

$$X_{\Psi} = \bigoplus_{k \in \Psi} W_{d_k}^{\Psi \setminus \{k\}},$$

where  $\bigoplus$  stands for the bit-wise XOR operator, and the superscript  $\Psi \setminus \{k\}$  implies that the segment transmitted to user  $k$  is the one stored at all other users in  $\Psi$ . In the physical layer,  $X_{\Psi}$  is mapped into a common multicast message which is then sent to the users in  $\Psi$  via a BC.

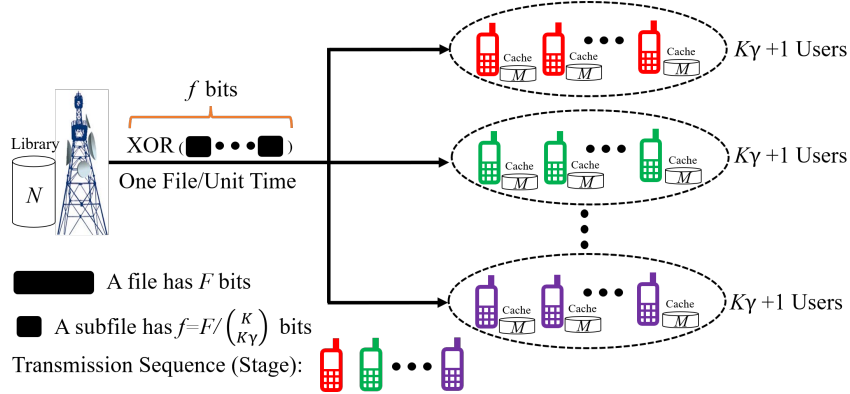


Figure 1.2: Delivery phase in MN Coded Caching

After successfully receiving  $X_{\Psi}$  — according to the MN approach — user  $k$  can “cache out” the undesired messages in  $X_{\Psi}$  by using its locally-cached content, and thus it obtains the desired subfile  $W_{d_k}^{\Psi \setminus \{k\}}$ . This is possible because all subfiles  $\{W_n^{\Psi \setminus \{k'\}}\}_{n=1}^N$ , for  $k' \in \Psi$  and  $k' \neq k$ , have been stored in the cache of user  $k \in \Psi$ . After  $\binom{K}{K\gamma+1}$  transmission stages, all the users obtain their demanded files. Note that even when  $K\gamma$  is not an

integer, the MN scheme can still achieve the coded caching gain  $K\gamma + 1$  by considering the memory-sharing strategy. We refer to [1] for more details about the MN scheme.

The above algorithm was originally developed for the scenario where the channel is error-free and the capacity of the channel to each user is identical. In recent years, a variety of works have investigated coded caching under more realistic wireless settings, considering for example uneven channel qualities [3–6], the role of Channel State Information (CSI) availability [7–9], statistically diverse channels [10,11], and a variety of other scenarios [12–17].

## 1.1 Two Fundamental Bottlenecks in Coded Caching

Unfortunately, it is the case that coded caching suffers from two major constraints. The first is often referred to as the “file-size constraint” of coded caching, which, as we will clarify later, effectively forces different users to fill up their caches with identical content [18–20]. This constraint, which has been extensively analyzed in the literature [12, 19–22], essentially foregoes the freedom to endow users with their own dedicated caches, and rather forces these users to share a very limited number of cache states<sup>1</sup> that is considerably smaller than  $K$ . This will result in reduced gains, as we will note later on. On the other hand, there is a seemingly unrelated constraint which stems from the fact that the XOR multicast transmissions are fundamentally and inevitably limited by the rate of the worst user that they address [23]. This constraint, often referred to as the “worst-user bottleneck” of coded caching, arises when users experience different channel strengths, and it is a constraint that is severely exacerbated as the SNR becomes smaller. This bottleneck also substantially diminishes the real cache-induced gains.

Both these realities, of bounded file sizes and limited SNR, are naturally inherent to any practical wireless content-delivery system. Let us look at these bottlenecks in greater detail.

### 1.1.1 Subpacketization Bottleneck and the Need for Shared Caches

Our work builds on the premise that almost any realistic single-stream coded caching scenario will involve the use of shared, rather than dedicated, cache states. As we will see right below, this has to do with the simple fact that, under realistic assumptions on  $\gamma$  and  $K$ , the file sizes required by coded caching schemes dwarf any realistic file sizes that we encounter in wireless downlink applications. The evidence for this is overwhelming, and to date, under realistic assumptions, any high-performance coded caching scheme requires files to be split into a number of sub-files that grows exponentially or near-exponentially with  $K$  (unless  $K$  itself becomes astronomical, which is a scenario that is of no interest in this thesis). For example, the MN algorithm requires files to be split into at least  $\binom{K}{K\gamma}$  subfiles, and it is known from [21, Thm. 3] that this same subpacketization is indeed necessary for any algorithm to achieve this same gain under some basic symmetry conditions. Similarly, it was shown in [18] that decentralized schemes (cf. [24]) require

---

<sup>1</sup>Hereinafter, *cache state* refers to the content stored at the cache of a certain user. Thus, two users storing the exact same content in their local cache are said to have the same cache state.

exponential (in  $K$ ) subpacketization in order to achieve linear caching gains, and, along similar lines, [25, Thm. 12] proved that, under basic assumptions, there exists no coded caching scheme that enjoys both linear caching gains and linear subpacketization.

From these previous results, we are in a position to say that such schemes will inevitably require many users to store the same cache content. While there is not a fundamental limitation that forces users to cache the same content, an extensive literature overview indicates that there are two possible solutions to keep the subpacketization low while maintaining the gains: either to repeat the same cache state at several users, or to massively increase  $K$  [21, 22, 25]. Let us consider for instance the original MN scheme. Under the constraint that the subpacketization (number of subfiles) cannot exceed a realistic value  $S_{max}$ , we know that the best course of action is to encode over a limited number of  $\Lambda < K$  users at a time [18], creating  $\Lambda$  different cache states. This  $\Lambda$  is indeed limited by the file size constraint that asks that  $\binom{\Lambda}{\Lambda\gamma} \leq S_{max}$ . This approach naturally limits the aforementioned (error-free) gain to  $\Lambda\gamma + 1$  [26], and it entails cache replication simply because now there are only  $\Lambda$  cache states to be shared<sup>2</sup> or replicated among the  $K$  users. As this thesis will show later on, this forced replication can be exploited to circumvent the other major problem: the worst-user bottleneck.

### 1.1.2 Worst-User Bottleneck: Motivation, Nature of the Problem, and Prior Work

As we have suggested, the worst-user limitation induced by the nature of the multicast transmission [23] is exacerbated when the SNR becomes smaller and when the channel strengths for each of the served users are different. Consequently, this dependence on multicasting can severely affect the applicability of coded caching in many wireless scenarios that possess such characteristics. These scenarios prevail in cellular or satellite communications settings [28] that suffer from heavy path-loss and/or shadowing, as well as in other massive Machine-Type Communication (mMTC) settings [29]. Similarly, we know that in 4G LTE networks, the range of users' signal-to-interference-plus-noise ratio (SINR) is typically 0–20 dB [30], while the SINR of cell-edge users can be closer to 0–5 dB. The worst-user bottleneck is also exacerbated when considering the well-established setting of quasi-static fading that we will consider in the following, and which generally comes about in the presence of longer coherence periods and shorter latency constraints. This quasi-static setting applies to low-mobility scenarios, which nicely capture coded-caching use-cases where pedestrians or static users are consuming video streaming.

This bottleneck has sparked considerable research interest that resulted in a variety of notable results [3, 4, 31–33]. For example, the work in [31] shows that, in a single transmit-antenna setting with finite power and quasi-static fading, the effective gain does not scale as  $K$  becomes larger *even in the absence of a file-size constraint*; moreover, the power must scale linearly with  $K$  in order to preclude the collapse of the multicast rate

<sup>2</sup>It is worth noting that the shared cache setting not only captures the effect of the file-size constraint, but also reflects promising heterogeneous scenarios where a main station serves users with the help of smaller cache-endowed helper nodes [26, 27].

(cf. [31, Table I]). Taking a different approach, the work in [32] employs superposition coding for opportunistic scheduling. Another notable work can be found in [33], which groups together users that experience similar SNR and delivers to each group in a separate way after neglecting users with the weakest channels.

However, to date, no scheme is known to overcome the worst-user bottleneck without user selection for the single transmit-antenna setting. In this context, we analyze in this work the worst-user bottleneck when no user selection techniques are applied, in order to expressly show that these techniques are not needed to overcome the worst-user bottleneck. This is an interesting outcome because user selection increases the complexity of the transmission in several aspects. First, because of the CSI required to decide which users to serve in each transmission. This CSI requirement entails a trade-off: In order to better exploit the benefits of user-selection techniques, the transmitter would require CSI from many users (ideally, all) at every time, which in turn may consume a lot of resources. Second, the transmitter would need to add an extra step to select the suitable user subset. Both CSI acquisition and selection algorithm can become challenging when the number of users become large, as it is sometimes assumed in our analysis.

In all these previous scenarios, this bottleneck substantially diminishes the aforementioned coded caching gain<sup>3</sup>. Had the SNR been infinite, or the instantaneous link strengths identical, this hypothetical gain would have taken the form  $\Lambda\gamma + 1$  for any allowable  $\Lambda$  up to  $K$  (where, we recall, this allowable  $\Lambda$  is generally much less than  $K$  due to the bounded file sizes). Yet, as the SNR decreases, the effect of the worst-user bottleneck becomes more accentuated<sup>4</sup>, and the effective gain eventually collapses. This collapse will be rigorously described in Proposition 2.2, and it is illustrated in Fig. 1.3.

## 1.2 Multi-Antenna Coded Caching

At the same time, it became apparent that for coded caching to develop into an impactful ingredient in wireless systems, it would have to work in conjunction with multi-antenna arrays which are rightfully recognized as the most valuable resource in modern networks. This realization brought to the fore notable research in the area of *multi-antenna coded caching* [34, 35]. In recent years, several related works explored various aspects of the problem, with substantial emphasis on physical-layer (PHY) considerations. One of the first such works can be found in [13] which designed PHY adaptations of various multi-antenna coded caching schemes. Another interesting approach can be found in [36] which presented a multi-antenna coded-caching scheme for lower SNR regimes when the placement exploits prior information on the users' locations. Furthermore, the work of [31] considered the use of transmit antennas for achieving rate scalability in the limit of large  $K$ .

---

<sup>3</sup>We remind the reader that the gain describes the cache-aided speed-up factor over the approach which employs the basic Time Division Multiplexing (TDM) method that serves one user at a time.

<sup>4</sup>To see this, simply recall that for smaller values of SNR and for  $z < 1$ , it follows that  $\ln(1 + z\text{SNR}) \approx z \ln(1 + \text{SNR})$ .

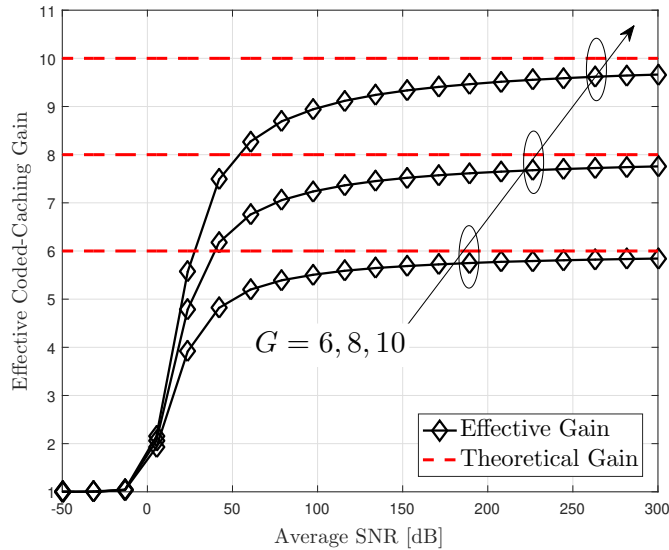


Figure 1.3: Ratio between the average rates of the MN scheme and TDM (i.e., the *effective coded-caching gain*) over quasi-static Rayleigh fading channel for different values of  $G = \Lambda\gamma + 1$ .

### 1.2.1 Multi-Antenna Coded Multicasting

Multi-antenna base stations (BSs) with full-digital (FD)/hybrid precoding allow us to simultaneously serve many users at a time, thus yielding higher spectral efficiencies. This comes though at the cost of needing a large number of radio frequency (RF) chains [37], which in turn requires larger energy and hardware costs compared to analog beamforming which uses, for example, a single-RF chain. Indeed the energy consumption on an RF chain is several dozens of times larger than that of a phase shifter [38]. Typically, employing a single RF chain analog beamforming forces TDMA delivery across the users requesting different content. As one can imagine, traditional coded caching bypasses this limitation by employing *coded multicasting* [34, 35], which considers the same model as the aforementioned cache-aided BC, except that now the server (the BS) is endowed with multiple transmit antennas. This approach though, even in its multi-antenna implementation, is still fundamentally constrained by the aforementioned worst-user bottleneck, which becomes particularly detrimental in realistic SNR implementations. As one can imagine, adding a large number of BS antennas to mitigate this bottleneck, will bring about higher costs in hardware/software and higher energy costs. Furthermore, again as one would expect, the overall throughput of this coded multicasting approach, even with a FD beamformer, would be much lower than that of its FD/hybrid precoding counterpart (for example, zero-forcing). Thus, under constraints on the number of RF chains, it seems that multi-antenna coded multicasting is meaningful in terms of energy savings and hardware/software costs, but runs the risk of yielding much lower overall throughputs. Therefore, how to efficiently boost coded multicasting rates in practical

cache-aided systems is an open question of practical interest.

### 1.2.2 Vector Coded Caching

Assuming though that we have some freedom on the number of RF chains<sup>5</sup>, one could employ more advanced techniques that nicely consider *the fusion of multi-antenna multicast beamforming and coded caching* toward improved interference management, e.g., [17, 39]. Interesting works can also be found in [7, 11, 15, 40–46] and in a variety of other publications. It is the case though that for most of the above schemes, the corresponding DoF impact of file-size constrained coded caching was merely additive to the multiplexing gain (denoted here by  $Q$ ), in the sense that, in most of the above scenarios, the DoF performance stagnated at around  $Q + \Lambda\gamma$  for very modest values of  $\Lambda\gamma$  that rarely exceed 6 under realistic assumptions. In essence, *due to the severity of the file-size constraint, the impact of caching was dwarfed by the existing and available multiplexing gains* which have been extensively demonstrated in various field trials [48].

This imbalance in the impact of caching on multi-antenna systems was reversed with the introduction in [12] of vector coded caching. This reversal is owed in part to the fact that this new approach could dramatically ameliorate the subpacketization problem previously associated to XOR-based schemes. While previous multi-antenna coded caching techniques essentially focused on using multiple antennas ( $L$  transmit antennas) to efficiently deliver the aforementioned sequence of XORs of the original MN scheme, the novel method in [12] applied a decomposition-based approach that employed a clique structure *on vectors* rather than on scalars. Vector coded caching need not entail the transmission of XORs. Building on the idea of employing  $\Lambda$  shared caches ( $\Lambda$  cache states) and linear precoding, the algorithm in [12] was able to offer unprecedented DoF performance as well as a dramatically reduced subpacketization. To be precise, for some  $Q \leq L$  representing the aforementioned multiplexing gain of choice, the algorithm in [12] reduced subpacketization from being exponential in  $K$  to being exponential in  $K/Q$ , all while being able to serve up to  $Q(1 + \Lambda\gamma)$  users at a time. This implied a theoretical multiplicative boost over the DoF of multiplexing-gain systems by a factor of  $1 + \Lambda\gamma$ , with the new DoF of  $Q(1 + \Lambda\gamma)$  far exceeding the additive impact (see DoF of  $Q + \Lambda\gamma$ ) of previous XOR-based multi-antenna coded caching approaches.

It is the case though that the work in [12] focused on the error-free, asymptotically high-SNR regime, without considering any practical aspects such as power dissemination across signals, realistic SNR values, the effects of beamforming gain, or the costs of gathering channel state information (CSI). To date, we know very little about the practical performance of vector coded caching in wireless systems. While this new approach was shown to be useful in an information-theoretic (DoF) sense, the real impact that this approach has on optimized downlink systems, has remained an open question.

---

<sup>5</sup>Such freedom may be acceptable in the conventional sub-6GHz systems, where high spectrum efficiency is often the prime focus.

## 1.3 Thesis Outline and Main Contributions

This thesis focuses on designing coded caching delivery schemes adapted to realistic wireless networks, as well as focuses on analyzing the corresponding performance. The main results will be presented in the next four chapters. In Chapter 2, we develop a novel coded caching delivery scheme — referred to here as the *Aggregated Coded Caching (ACC)* scheme — which is tailored for wireless settings. We also analyze the achievable rates and effective gains of this new scheme, and do so in various settings. In Chapter 3, we slightly modify the vector coded caching scheme originally developed in [12] to tailor it to realistic wireless networks, as well as proceed to rigorously analyze the corresponding performance. We further investigate the impact of more practical considerations on the delivery performance of vector coded caching in Chapter 4. In Chapter 5, we apply the ACC in cache-aided satellite systems and proceed to analyze the corresponding achievable rates and effective gains. Chapter 6 concludes the thesis. In the following, we give an extended summary of the main contributions found in Chapters 2–5.

### 1.3.1 Main Contributions Toward Resolving Worst-User Bottleneck

#### Aggregated Coded Caching (ACC) Design

In Chapter 2, focusing on the file-size constrained scenario (which corresponds to having a limited number  $\Lambda$  of different cache states, and which is effectively forced upon us), we present the novel ACC transmission scheme, which substantially improves the effective gain, and which manages to recover — without user selection or any additional overhead — the entire nominal coded caching gain  $\Lambda\gamma + 1$  in the presence of sufficiently many users. This means that the proposed ACC enables us to asymptotically remove the aforementioned worst-user bottleneck which was thought to diminish any coded caching gains in realistic wireless networks. The proposed ACC scheme builds on the practical inevitability of having users with identical cache content, and it employs multi-rate encoding that avoids XOR transmissions, thus allowing each user to receive at a rate that matches its single-link capacity. We note that the transmission rate for a served user achieving its single-link capacity is possible because this user knows (and has access to) the messages intended for other simultaneously served users (cf. [49]), which here holds due to the shared-cache placement policy. Fig. 1.4 offers a small illustration of the main difference between ACC and XOR-based (MN) coded caching, where in the latter case we see that one has to wait for the worst user to decode its desired subfile before starting a new transmission round, whereas with ACC, when a served user decodes its desired subfile successfully, another user with the same cache state can replace the served user, immediately, without interrupting the subfile decoding, and without generally reducing the numbers of simultaneously served users. As it turns out, the ACC introduces a time (or space) diversity effect, where having  $B$  users per cache state is — in terms of effective channel statistics — akin to enjoying the time diversity benefits of encoding across  $B$  coherence times, or the space diversity benefits of degree  $B$  (see Fig. 2.3 which is also shown at the top of next page). To be clear, neither do we encode over coherence times, nor over space. We refer to Section 2.2 for more details on the ACC design which will be

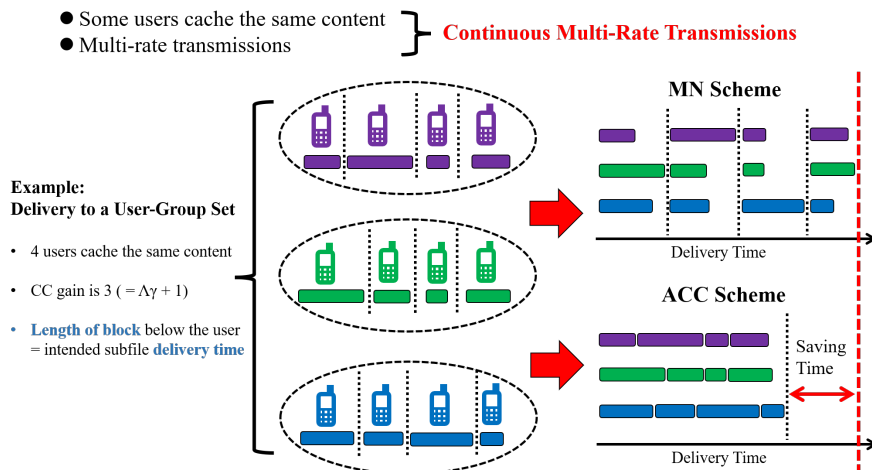
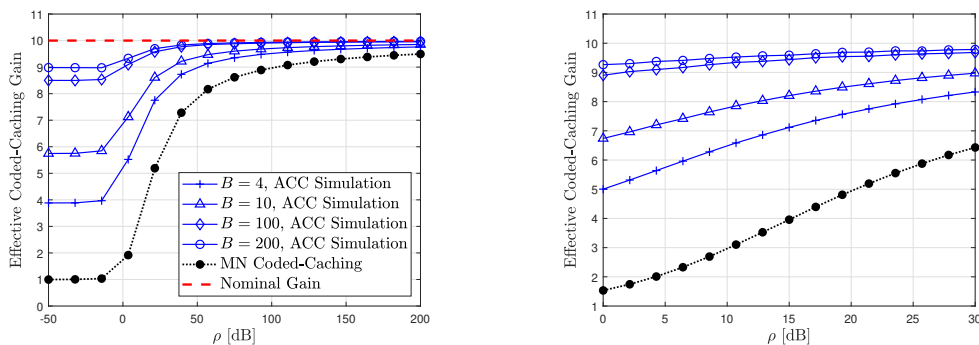


Figure 1.4: An example of ACC delivery.


 Fig. 2.3: Effective gain versus  $\rho$  (SNR) for  $G = 10$  over symmetric Rayleigh fading channels. Right-side plot focuses on realistic SNR values.

based on the dual idea of combining cache replication and multi-rate transmission. This same idea can be readily applied to a variety of different coded caching algorithms in order to ameliorate the worst-user effect. This ability is highlighted in Fig. 2.11, where we also apply our ideas in the context of decentralized coded caching.

### Average Rate Analysis of ACC

We analyze the delivery performance of ACC in terms of average achievable rate. With the help of the inverse transform of characteristic functions (CF), we are able to derive an analytical expression for the average rate  $\bar{R}^{\text{ACC}}$  of ACC over quasi-static Rayleigh fading channels. To simplify the expression and to provide insight, we also consider the practical large- $B$  (large- $K$ ) scenario. Using the Central Limit Theorem (CLT) and after defining  $\mathcal{H}_G$  as the expectation of the maximum of  $G$  i.i.d. standard normal random variables, we derive in Lemma 2.6 a simple rate approximation in the form

$$\bar{R}^{\text{ACC}} \approx \frac{G}{\ln 2} \left( \varrho - \frac{\sigma}{\sqrt{B}} \times \mathcal{H}_G \right) \text{ bits/s/Hz}, \quad (1.1)$$

where  $G$  denotes the file-size constrained nominal gain, and where  $\frac{\rho}{\ln 2}$  is the average rate of time-division multiplexing (TDM) transmissions (corresponding to uncoded caching delivery, or effectively corresponding to the cacheless case). In (1.1),  $\sigma$  reflects the average channel fluctuation resulting from random fading. The above shows that when  $B$  is sufficiently large, the impact of channel fluctuation on  $\bar{R}^{\text{ACC}}$  is negligible, and it also shows that  $\bar{R}^{\text{ACC}}$  converges to  $\frac{G}{\ln 2}\rho$ , which mathematically explains why ACC can fully recover the nominal gain  $G$  at any SNR. Numerical results then also show that this simple approximation is very tight even for small values of  $B$  (cf. Fig. 2.7). Moreover, based on the second-order Taylor expansion, we also develop a simple approximation (without any use of special/advanced functions) for the average rate  $\bar{R}^{\text{MN}}$  of the original (MN) coded caching in the low SNR region (SNR is also often denoted by  $\rho$ ), which is shown in Lemma 2.2 to take the form

$$\bar{R}^{\text{MN}} \approx \frac{G}{\ln 2} \left( \ln \left( 1 + \frac{\rho}{G} \right) - \frac{\rho^2}{2G^2 (1 + \rho/G)^2} \right) \text{ bits/s/Hz}, \quad (1.2)$$

which is later numerically shown to be robust even in the medium-SNR region (cf. Fig. 2.5), or even in the higher SNR regions.

### Delivery Time Analysis of the ACC Scheme

In order to analyze the impact of the novel ACC on the delivery time, we consider Nakagami- $m$  fading and analyze the delivery time for both the MN and the ACC schemes. We do so in the low SNR region. Our considering of Nakagami- $m$  fading — in addition to allowing for convergence of the delivery time — also allows us to capture a broad spectrum of practical wireless scenarios, such as land-mobile and indoor-mobile scenario with multi-path propagation, the scenario with scintillating ionospheric radio links [50], or (when  $m$  is a positive integer), the scenario of having  $m$ -antenna receivers over symmetric Rayleigh fading after applying maximal-ratio combining (MRC).

Similar to the average rate analysis, we first employ the CF inverse transform to derive an analytical expression for the average delivery time of ACC in the low SNR region. To further simplify the derived double-integral expression, we consider the large- $B$  case and use the CLT to obtain a simple expression for the effective coded caching gain that takes the form

$$\mathcal{G}_{\text{ACC}} \approx \frac{G}{1 + \mathcal{H}_G / \sqrt{B(m-2)}}, \text{ for } m > 2, \quad (1.3)$$

which reveals how the fading parameter  $m$  and the ratio  $\frac{K}{\Lambda} = B$  tend to impact in a similar manner the performance of ACC, since they both offer an equivalent diversity effect. The above reveals that we can recover the full nominal gain of coded caching even at low SNR, provided that there are enough users or a sufficiently large  $m$ . Numerical results (see for example Fig. 2.16 which is also shown at the top of next page) also validate this observation. In addition, we also utilize the Strong Law of Large Numbers to simplify the analysis on the effective gain of MN coded caching that is here shown to take the form

$$\mathcal{G}_{\text{MN}} \xrightarrow{a.s.} G = \Lambda\gamma + 1 \text{ as } m \rightarrow \infty, \quad (1.4)$$

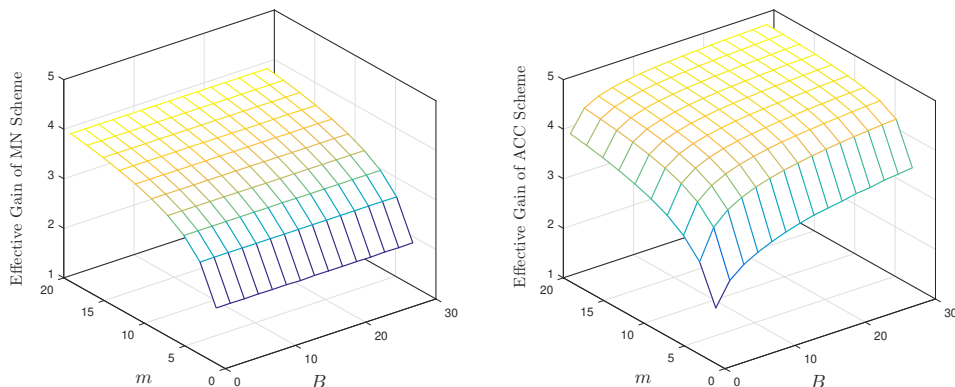


Fig. 2.16: Effective gains of MN (left) and ACC (right) schemes for  $G = 5$  and  $\gamma = 5\%$  as  $\rho \rightarrow 0$ .

which then implies that the worst-user bottleneck in MN coded caching can be asymptotically resolved for  $m \gg 1$  for any SNR (see also Fig. 2.16). This naturally comes in direct contrast to the previous conclusion that when considering Rayleigh fading and a single receive antenna, the effective gain of MN completely vanishes in the low SNR limit (see also Fig. 1.3).

### ACC Performance Over Ergodic Fading Channels

The worst-user effect in coded caching will become more detrimental when the users are distributed across different locations at different distances from the BS. This will bring about the well-known near-far problem, and it is in this setting that we will show major ACC advantages, which will be rigorously described by analyzing its performance in the ergodic single-cell cache-aided setting. For a scenario with an inner cell radius  $D_1$  and an outer cell radius  $D_2$ , we provide simple and accurate expressions, comparing the performance of the MN and ACC schemes, under various realistic scenarios. Deviating from the previous low-SNR assumptions, we now take a high-SNR approach that paradoxically manages to tightly quantify the near-far bottleneck which is generally associated to low or moderate SNR values. Specifically, employing Jensen's inequality and a very basic and robust high SNR channel capacity approximation, we parameterize (cf. Corollary 2.4) the MN effective gain to take the form

$$\mathcal{G}_{\text{MN}} \approx \frac{G \ln \rho - G \left[ \ln \left( \frac{2G(D_2^{\eta_0+2} - D_1^{\eta_0+2})}{(\eta_0+2)(D_2^2 - D_1^2)} \right) + C \right]}{\ln \rho - C + \frac{\eta_0}{2} - \eta_0 \frac{D_2^2 \ln D_2 - D_1^2 \ln D_1}{D_2^2 - D_1^2}}, \quad (1.5)$$

where  $\eta_0$  denotes the pathloss exponent, and where  $C = 0.5772\dots$  denotes the Euler–Mascheroni constant. The above quantifies the impact of system parameters  $\eta_0$ ,  $D_1$  and  $D_2$  on the MN effective gain, revealing a bound on the transmit power required to achieve a certain effective gain. Fig. 2.19 (also shown at the top of next page) validates the tightness of (1.5), comparing to the actual performance, and does so under 3GPP guidelines for an urban cell.

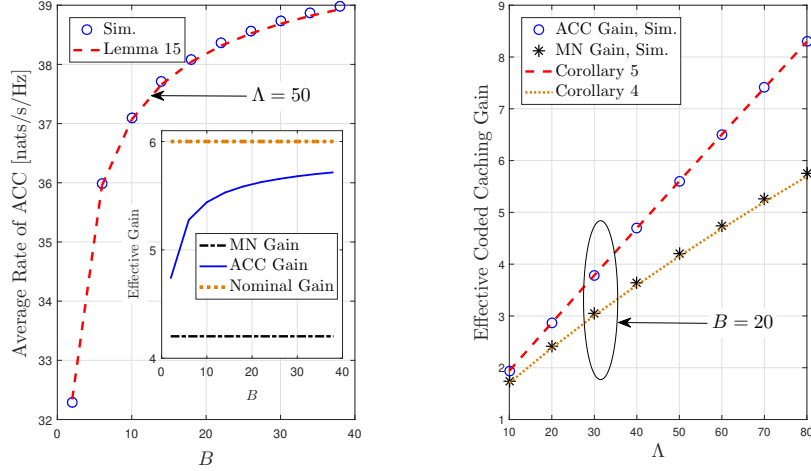


Fig. 2.19: Performance in the Macro-Cell setting with  $\gamma = 10\%$  and  $P_t = 40$  dBm.  $\mathcal{H}_G$  is computed through GHQ with 10 terms.

Then we further analyze the ACC scheme by employing a large- $B$  (large  $K$ ) assumption to simplify (2.56). Using the methodology corresponding to (1.1), we derive (cf. Corollary 2.5) a very simple expression for the high-SNR ACC effective gain, that takes the form

$$\mathcal{G}_{\text{ACC}} \approx \frac{G}{\varrho_s} \left( \varrho_s - \mathcal{H}_G \sqrt{\sigma_s^2/B} \right), \text{ for } B \rightarrow \infty, \quad (1.6)$$

where  $\varrho_s$  denotes the rate of TDM averaged over channel fading and random user locations, and where  $\sigma_s$  reflects the average channel fluctuation due to the small-scale fading and the random user locations. As  $\varrho_s$  and  $\sigma_s$  are independent of  $B$ , we can conclude from (1.6) that our ACC scheme recovers most of the nominal gain  $G$  for reasonable values of  $B$ . This again reveals that the ACC scheme offers the equivalent of a spatial-averaging effect, which helps overcome the near-far bottleneck. Fig. 2.19 — which also demonstrates the tightness and accuracy of (1.6) — also clearly shows that ACC recovers over 93% of the nominal gain for reasonable values of  $B$ , especially in urban settings.

### Multi-Antenna ACC

We here combine the ACC idea with the multi-antenna MU multicasting idea (of using multiple antennas to steer a coded caching stream), to show that the two work together in a synergistic manner. To see this, let us recall that in the previous case of employing a single-antenna transmitter, when we complete the delivery to all the users that share the same cache state, then the degree of multicasting is reduced, and the overall delivery rate generally decreases. What we now see though is that we can take advantage of this ‘freeing up’ of a cache state: Now, in the multi-antenna (rather than single-antenna) ACC scenario, upon completion of the delivery to all the users that share the same cache state, the multi-antenna transmitter can form narrower beams dedicated to a now smaller number of cache state groups; this accelerates the delivery to those groups, thus avoiding

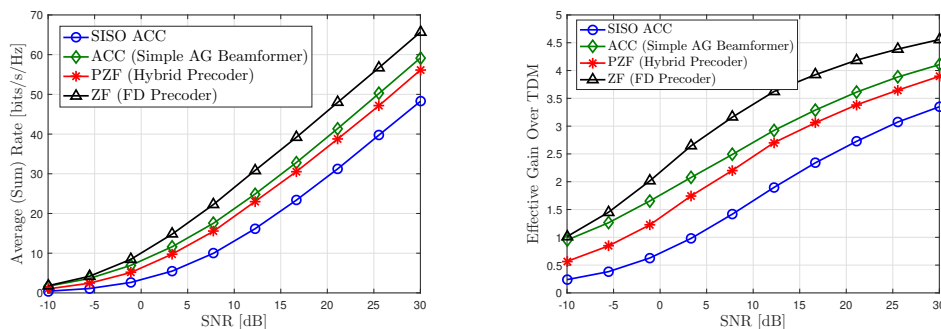


Fig. 2.29: Delivery performance over mmWave channels with  $L = 32$ ,  $G = 6$ ,  $B = 5$ ,  $B_w = 200$  MHz,  $N_p = 2$  and  $F_{\text{sub}} = 50$  bytes.

the aforementioned dilution of the effect of removing cache-states from delivery. While the multicasting gain may progressively decrease, the beamforming gain progressively increases. As we will see, the above ACC-aided multicasting method with a simple AG beamformer (one RF chain), not only significantly outperforms the optimized FD multicasting beamformer in the conventional MU multicasting, but also comes close to matching the performance of FD precoding (e.g., ZF) operating with  $L$  RF-chains and does so under realistic scenarios such as the mmWave setting (see Fig. 2.29 above). This performance is also illustrated in Fig. 2.23. In the extreme case of a large number of users sharing the same cache content, and for a fixed theoretic DoF (nominal gain) of  $\Lambda\gamma + 1$ , we can use a single transmit antenna to match (or even exceed) the performance of multiple transmit antennas in the conventional MU multicasting. We note that this simple AG beamformer has only a single RF chain and requires much lower hardware/software costs, as well as lower power consumption, than the FD precoder. *As we will see later, fully exploiting multiplexing and beamforming gains (by employing PHY-optimized vector coded caching) would yield a multiplicatively larger theoretic DoF than the DoF of the above coded MU multicasting, but we recall that this performance boost would require a higher resource consumption for operating the RF chains.*

Some related publications include:

[51] Hui Zhao, Antonio Bazco-Nogueras, and Petros Elia, “Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5450–5466, Jul. 2022.

[52] Hui Zhao, Antonio Bazco-Nogueras, and Petros Elia, “Resolving the worst-user bottleneck of coded caching: Exploiting finite file sizes,” in *Proc. IEEE Information Theory Workshop (ITW)*, Apr. 2021, pp. 1–5

[53] Hui Zhao, Antonio Bazco-Nogueras, and Petros Elia, “Wireless coded caching with shared caches can overcome the near-far bottleneck,” in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jul. 2021, pp. 350–355.

[54] Hui Zhao, Antonio Bazco-Nogueras, and Petros Elia, “Coded caching gains at low

*SNR over Nakagami fading channels,” in Proc. 55th Asilomar Conference on Signals, Systems, and Computers (ACSSC), Nov. 2021, pp. 1–7. (Best Student Paper Finalists)*

### 1.3.2 Main Contributions in Vector Coded Caching

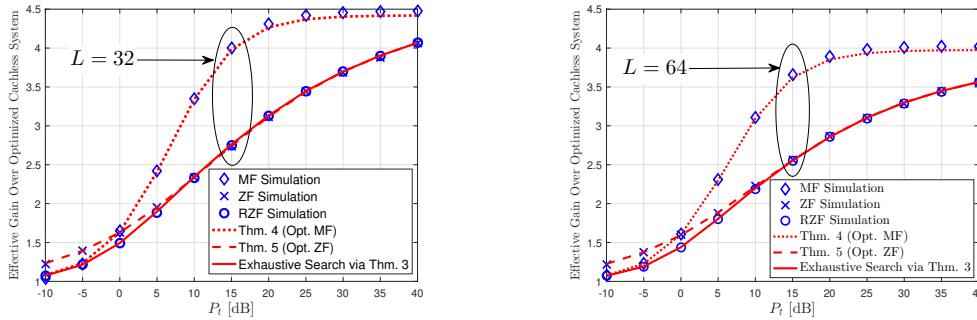
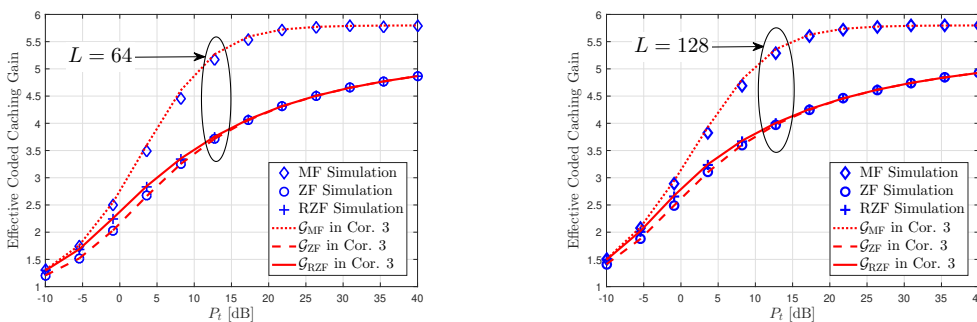
The recent introduction of vector coded caching in [12] has revealed that multi-rank transmissions in the presence of receiver-side cache content can dramatically ameliorate the file-size bottleneck of coded caching and substantially boost performance in error-free wire-like multi-rank channels. Any attempt to establish the real impact of vector coded caching — over realistic wireless channels — must answer a simple question: Under a fixed set of antenna and SNR resources, what is the multiplicative throughput boost obtained from being able to add receiver-side caches to downlink systems that would have otherwise been able to enjoy an optimized exploitation of multiplexing and beamforming gains. Indeed, spatial multiplexing and beamforming in multi-antenna downlink systems, and its well-studied application in the large-antenna regime or *massive* MIMO [55–58], is a key technology in current and future wireless networks that significantly enhances spectral efficiency. Such enhancements have been recently proven in the aforementioned field trials [48] which convincingly demonstrate that a sizeable fraction of the promising theoretic gains brought about by spatial multiplexing approaches (naturally without caches), can indeed be attained under practical constraints.

In the context of vector coded caching, which will enjoy caching gains and spatial-multiplexing gains, our focus will be on the work-horses of spatial-multiplexing precoding which are the optimized versions of linear precoding techniques such as Zero-Forcing (ZF), Regularized ZF (RZF), and Matched Filtering (MF). These techniques maintain low complexity and can provide very high spectral efficiency that often comes close to the optimal performance of the non-linear Dirty-Paper Coding, especially when the number of transmit antennas  $L$  is large [55]. Our focus will also incorporate power allocation aspects (where we recall that the more symbols one transmits, the more power splitting one encounters), as well as, as one would expect, aspects regarding the acquisition of CSI which is another ingredient of crucial importance, even in the presence of Time Division Duplexing (TDD) that partially reduces the CSI overhead as the dimensionality of the problem becomes larger [59]. This same CSI overhead will naturally bring to the fore the issue of channel hardening, which arises as the number of antennas increases, and which partially alleviates the stringent CSI requirements [60].

#### Single-Antenna Receivers

In Chapter 3, we employ large-matrix analysis to explore the effect of vector coded caching in realistic wireless multi-antenna downlink systems. For a given downlink MU-MISO system already optimized to exploit both multiplexing and beamforming gains, and for a fixed set of antenna and SNR resources, our analysis answers the aforementioned simple question of what is the multiplicative increase in the throughput that we can obtain when we can add simple and relatively small caches at the receivers?

The employed scheme that we optimize and analyze is indeed very simple (we simply collapse precoding vectors into a single vector), and the recorded gains are notable. For


 Fig. 3.3: Effective gain  $\mathcal{G}^*$  over optimized cacheless system for  $L \in \{32, 64\}$  and  $G = 6$ .

 Fig. 3.4: Hardening-constrained effective gain over a constrained classical downlink system.  $Q$  is fixed for both systems at  $Q = 8$ , while  $G = 6$ .

example, as we witness in the above Fig. 3.3, for the case of 32 transmit antennas, a received SNR of 20 dB, a coherence bandwidth of 300 kHz, a coherence period of 40 ms, and under realistic file-size and cache-size constraints, vector coded caching is shown to offer a multiplicative throughput boost of about 310% with ZF/RZF precoding and a 430% boost in the performance of already optimized MF-based (cacheless) systems. Interestingly, as we will clarify later on, vector coded caching also accelerates channel hardening (which it self yields benefits with regards to feedback acquisition) because it allows us — roughly speaking — to have a larger ratio between the number of antennas and the number of interfering symbols that must be resolved by precoding. In this setting, we see vector coded caching often surpassing 540% gains over traditional hardening-constrained cacheless downlink systems (see above Fig. 3.4). To better understand the presented gains in practice, we refer to Example 3 in Chapter 3 for more information.

The derived closed-form expressions capture the various aforementioned linear precoders, capture a variety of practical considerations such as power dissemination across signals, realistic SNR values, as well as capture feedback costs. For example, given a multiplexing gain  $Q'$  in the cacheless counterpart, and a transmit power  $P_t$  normalized to AWGN noise, with the help of large random matrix theory, we derive the effective gains (these are gains over the cacheless MU-MISO equivalent: see Definition 3) of vector coded caching — for MF, ZF and RZF precoding respectively — and show them to take

the form

$$\mathcal{G}_{\text{MF}} \triangleq \frac{\bar{\mathcal{R}}^{\text{MF}}(G, Q)}{\bar{\mathcal{R}}^{\text{MF}}(1, Q')} = \xi_{\text{csi}} \frac{GQ \ln \left( 1 + \frac{L}{Q} \frac{P_t}{P_t + G} \right)}{Q' \ln \left( 1 + \frac{L}{Q'} \frac{P_t}{P_t + 1} \right)}, \quad (1.7)$$

$$\mathcal{G}_{\text{ZF}} \triangleq \frac{\bar{\mathcal{R}}^{\text{ZF}}(G, Q)}{\bar{\mathcal{R}}^{\text{ZF}}(1, Q')} = \xi_{\text{csi}} \frac{GQ \ln \left( 1 + \frac{P_t}{G} \left( \frac{L}{Q} - 1 \right) \right)}{Q' \ln \left( 1 + P_t \left( \frac{L}{Q'} - 1 \right) \right)}, \quad (1.8)$$

$$\mathcal{G}_{\text{RZF}} \triangleq \frac{\bar{\mathcal{R}}^{\text{RZF}}(G, Q)}{\bar{\mathcal{R}}^{\text{RZF}}(1, Q')} \xrightarrow{a.s.} \xi_{\text{csi}} \frac{\hat{R}^{\text{RZF}}(G, cL)}{\hat{R}^{\text{RZF}}(1, c'L)}, \quad (1.9)$$

where  $\xi_{\text{csi}}$  accounts for the effective gain loss factor due to CSI costs, and  $\hat{R}^{\text{RZF}}$  is the asymptotic deterministic equivalence of the sum-rate in RZF as  $L \rightarrow \infty$ . We refer to Corollary 3.3 for more details. We then employ the derived expressions with ZF and MF precoding to optimize the number of simultaneously served users ( $Q^*$ ) in order to maximize the overall throughput as a function of the transmit power  $P_t$ , the CSI cost  $\zeta_{G,Q}$ , the nominal gain  $G$  and the optimal ratio  $c^* = Q^*/L$ . For some  $\Omega \triangleq \frac{P_t}{P_t + G}$ , we can see that the optimal ratio  $c^* = Q^*/L$  under MF and ZF precoding schemes can be found via numerically solving the identities: (cf. Theorems 3.4, 3.5 respectively)

$$(1 - 2\zeta_{G,Q}c^*) \ln \left( 1 + \frac{\Omega}{c^*} \right) - \frac{\Omega(1 - \zeta_{G,Q}c^*)}{\Omega + c^*} = 0, \quad (1.10)$$

$$(1 - 2\zeta_{G,Q}c^*) \ln \left( 1 + \frac{P_t}{G} \left( \frac{1}{c^*} - 1 \right) \right) - \frac{(1 - \zeta_{G,Q}c^*)P_t/G}{(1 - P_t/G)c^* + P_t/G} = 0, \quad (1.11)$$

which are utilized to generate the numerical results in Fig. 3.3 to show the effective gains over the optimized cacheless counterparts. As mentioned, these gains exceed 300% and 400%, depending on the precoder. It is worth noting a novel proof in this thesis, which rigorously provides the asymptotic sum-rate of the conventional (cacheless) MF-based MIMO BC (cf. Corollary 3.1) which takes the form

$$\bar{R}^{\text{MF}} = Q' \ln \left( 1 + \frac{L}{Q'} \frac{P_t}{P_t + 1} \right) + o(1), \text{ as } Q', L \rightarrow \infty. \quad (1.12)$$

It is worth noting that while there have been various works (cf. [61–64]) analyzing the MF sum-rate in traditional massive MIMO systems, the result derived in this work here entails less assumptions. For example, focusing on the large- $L$  regime, the result in [61] directly assumes a tight Jensen’s bound, while the result in [63] is under a so-called “near deterministic” assumption in low/high SNRs. On the other hand, our method here draws from the uplink analysis in [65], and only employs a large- $L$  assumption to derive the exact asymptotic optimality for any value of SNR.

The work presented in this chapter has resulted in the following publications:

[66] Hui Zhao, Antonio Bazco-Nogueras, and Petros Elia, “Vector coded caching multiplicatively increases the throughput of realistic downlink systems,” *IEEE Transactions on*

*Wireless Communications, accepted for publication, doi: 10.1109/TWC.2022.3213475.*

[67] Hui Zhao, Antonio Bazco-Nogueras, and Petros Elia, “Vector coded caching greatly enhances massive MIMO,” in *Proc. IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC), Jul. 2022, pp. 1–5.*

### Multi-Antenna Receivers in Various Scenarios with Additional Practical Considerations

In Chapter 4, we continue to investigate the impact that vector coded caching has in various MU-MIMO scenarios<sup>6</sup>. In this particular part of our work, we consider again various practical aspects such as channel fading, CSI costs, linear precoding, multi-antenna receiver combining, and variable path-loss, as well as we consider max-min fairness (MMF) where the minimum transmission rate among the simultaneously served users is maximized via power allocation.

In addition to us considering multi-antenna receivers, we also consider block diagonalization (BD) precoding at the BS which allows the BS to send multiple data streams to each served multi-antenna user. We also consider MRC receivers, again in the context of vector coded caching. To elaborate, let the  $q$ -th data symbol to user  $U_{\psi,k}$  be denoted by  $s_{\psi,k,q}$ , and let the number of data streams simultaneously sent to  $U_{\psi,k}$  be denoted by  $J_{\psi,k}$ . Furthermore, let  $\mathbf{H}_{\psi,k}$  be the channel matrix from the BS to  $U_{\psi,k}$ , and let  $\mathbf{T}_{\psi,-k}$  be the null-space projection matrix for interference cancellation. Using the properties of the projection matrix (e.g., Hermitian, positive semidefinite, and idempotent), we are able to analytically derive the optimal precoding vector  $\mathbf{v}_{\psi,k,q}^*$  for the data symbol  $s_{\psi,k,q}$  as

$$\mathbf{v}_{\psi,k,q}^* = \frac{\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}}{\|\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}\|}, \quad (1.13)$$

where  $\mathbf{t}_{\psi,k,q}$  is the eigenvector associated to the  $q$ -th largest (non-zero) eigenvalue  $\lambda_{\psi,k,q}$  of  $\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^*$ . We note that the dimension size of  $\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^*$  depends on the number of receive antennas at user  $U_{\psi,k}$  which is in practice much lower than the number of transmit antennas. This considerably reduces the implementation complexity of finding  $\mathbf{t}_{\psi,k,q}$  and  $\lambda_{\psi,k,q}$ . Given the optimized BD precoder in (1.13) and of the MRC receiver, the resulting SINR for decoding  $s_{\psi,k,q}$  at  $U_{\psi,k}$  is shown to be

$$\text{SINR}_{\psi,k,q}^{\text{BD-MRC}} = \frac{P_{\psi,k,q}}{N_0} \lambda_{\psi,k,q}, \quad (1.14)$$

where  $P_{\psi,k,q}$  is the transmit power allocated to  $s_{\psi,k,q}$ . We refer to Lemma 4.1 for more information.

Based on the SINR expression in (1.14) and recalling that  $\zeta_{G,Q}$  accounts for the CSI cost, we formulate the power allocation problem for MMF to maximize the minimum effective rate among the simultaneously served users in vector coded caching as

$$\max_{\mathbf{P}_{\Psi}} \min_{\psi \in \Psi} \min_{k \in [Q]} \zeta_{G,Q} \sum_{q=1}^{J_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q}}{N_0} \lambda_{\psi,k,q} \right)$$

<sup>6</sup>The work presented in this chapter will constitute a journal paper, which is currently under preparation.

$$\text{s.t. } P_t = \text{Tr}\{\mathbf{P}_\Psi\} = \sum_{\psi \in \Psi} \sum_{k \in [Q]} P_{\psi,k} \leq P_{\text{tot}}, \quad (1.15)$$

where  $P_{\text{tot}}$  denotes the allowable maximum transmit power at the BS. We present Theorem 4.1 to solve this power allocation problem for MMF, as well as provide simplified solutions in some special cases. Lemma 4.2 provides some further simplified results. When the served users have the same number of receive antennas  $M$ , in the massive MIMO regime, we can explicitly and tightly approximate the optimal sum-rate  $R_{\text{BD-MRC}}^*$  by solving (1.15) over Rayleigh fading channels, to get

$$R_{\text{BD-MRC}}^* \approx \zeta_{G,Q} GQM \ln \left( 1 + \frac{P_{\text{tot}}(L - (Q - 1)M)}{N_0 M \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right), \quad (1.16)$$

where  $\beta_{\psi,k}$  accounts for the pathloss of user  $U_{\psi,k}$ . We can see from (1.16) that the sum-rate under this optimized BD-MRC scheme depends on the cumulative effect of the inverse pathloss in all served users<sup>7</sup>. The numerical results in Fig. 4.2 also validate the tightness of (1.16). Then approximately parameterize the effective gain over the optimized cacheless counterpart as

$$\mathcal{G}_{\text{BD-MRC}}^* \approx \frac{\max_{Q \in [Q_{\text{max}}]} R_{\text{BD-MRC}}^*(G, Q)}{\max_{Q' \in [Q'_{\text{max}}]} R_{\text{BD-MRC}}^*(1, Q')}. \quad (1.17)$$

In addition, in the same chapter, we will design a simple ZF precoder, and we will formulate a similar MMF power allocation problem. For the case where the served users have the same number of receive antennas  $M$ , we will derive (cf. Theorem 4.2) a tight lower-bound  $\tilde{R}_{\text{ZF}}^*$  and a tight upper-bound  $\hat{R}_{\text{ZF}}^*$  for the optimal sum-rate, that will take the form

$$\tilde{R}_{\text{ZF}}^* = \zeta_{G,Q} GQM \ln \left( 1 + \frac{P_{\text{tot}}(L - QM)}{MN_0 \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right), \quad (1.18)$$

$$\hat{R}_{\text{ZF}}^* = \zeta_{G,Q} GQM \ln \left( 1 + \frac{P_{\text{tot}}(L - QM + 1)}{MN_0 \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right). \quad (1.19)$$

After a quick comparison between (1.16) and (1.19), we will conclude that  $R_{\text{BD-MRC}}^*$  is always greater than  $\hat{R}_{\text{ZF}}^*$  (since  $QM - M \leq QM - 1$ ), which is also validated via numerical results in Fig. 4.2. We will also note that the performance gap between BD and ZF precoders is very small for modest values of  $M$ , e.g.,  $M \leq 4$ . By using the derived lower and upper bounds, the effective gain in ZF precoding over the optimized cacheless counterpart will be tightly lower bounded by

$$\tilde{\mathcal{G}}_{\text{ZF}}^* \geq \frac{\max_{Q \in [Q_{\text{max}}]} \tilde{R}_{\text{ZF}}^*(G, Q)}{\max_{Q' \in [Q'_{\text{max}}]} \hat{R}_{\text{ZF}}^*(1, Q')}. \quad (1.20)$$

<sup>7</sup>That is, if some users were to change their locations while keeping the inverse pathloss summation intact, this would not affect  $R_{\text{BD-MRC}}^*$ . This can have practical considerations in scenarios with many users.

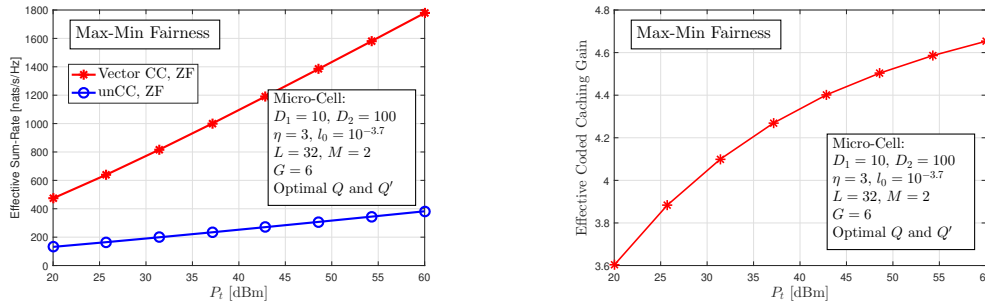


Fig 4.7: Delivery Performance of vector coded caching in a Micro-cell

In Fig. 4.6, for the case of having 64 transmit antennas, 4 receive antennas at each user, a transmit power of 50 dBm, a coherence bandwidth of 300 kHz, a coherence period of 40 ms, under realistic file-size and cache-size constraints and in realistic pathloss settings, vector coded caching will be here shown to offer a multiplicative throughput boost of about 250% with ZF/BD precoding over the already optimized cacheless system in a Macro-cell with an inner radius of  $D_1 = 35$  m and an outer radius of  $D_2 = 500$  m. Moreover, this effective gain will be elevated to 450% in a Micro-cell with an inner radius of  $D_1 = 10$  m and an outer radius of  $D_2 = 100$  m, even for a smaller antenna arrays (32 transmit antennas and 2 receive antenna each user), as shown in the above Fig 4.7.

Our preliminary research work for this topic is shown below, while other works are in preparation.

[68] Hui Zhao, Eleftherios Lampiris, Giuseppe Caire, and Petros Elia, “Multi-antenna coded caching analysis in finite SNR and finite subpacketization,” in *Proc. 25th International ITG Workshop on Smart Antennas (WSA)*, Nov. 2021, pp. 433–438.

[69] Hui Zhao, and Petros Elia, “Vector coded caching substantially boosts MU-MIMO: Pathloss, CSI and power-allocation considerations,” in *Proc. 26th ITG International Workshop on Smart Antennas (WSA) and 13th Conference on Systems, Communications, and Coding (SCC)*, Feb. 2023.

### 1.3.3 Main Contributions in Land Mobile Satellite Systems

Motivated by the upcoming satellite integration in beyond-5G networks, our work in [70] has explored land mobile satellite systems (LMSs) and the role coded caching can play in such systems. In particular, our work has shown that even a basic MN implementation of coded caching, can double the goodput of a basic LMS — generally operating at very low SNR — despite the previously discussed worst-user drawback of such coded caching approaches<sup>8</sup>. Motivated by this finding, in Chapter 5, in the downlink setting in an LMS with a single-antenna transmitting satellite, we consider the use of coded caching and

<sup>8</sup>In the context of our previous observation that — in the presence of Rayleigh fading — the low-SNR gains entirely vanish, we note that here the gain does not fully vanish because of the existence of line-of-sight (LOS) components over the satellite-terrestrial channel.

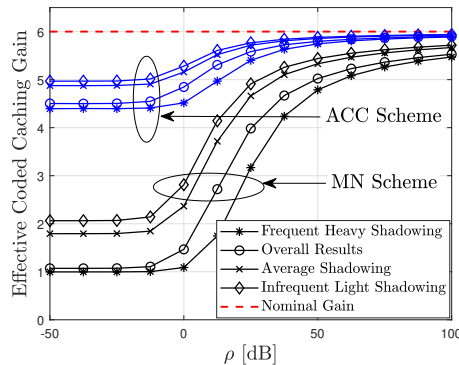


Fig. 5.3: Effective coded caching gain versus  $\rho$  for  $G = 6$  and  $B = 20$ .

analyze the cache-aided delivery performance in terms of the achievable rate and the effective coded caching gain.

As before, and motivated by the fact that the worst-user bottleneck may be particularly detrimental because of the low SNR induced by the large distances between the satellite and the terrestrial users, we here again consider the ACC scheme in order to overcome this bottleneck. In particular, the work presented in this chapter and in the corresponding paper under preparation, considers Rician-shadowed fading, where the lowpass-equivalent complex signal envelope is composed<sup>9</sup> by the scatter and the LOS components [71]. In this LMS setting, we first derive the analytical expressions for the average achievable rate of the ACC scheme over Rician-Shadowed fading channels. To then insightfully simplify the derived double-integral analytical expression, we provide simple approximations in the regimes of low SNR and/or large  $K$ . Specifically, a simple but tight approximation for the effective coded caching gain in the low SNR and large  $B$  regime, will take the form

$$\mathcal{G}_{\text{ACC}} \approx G - G \frac{\mathcal{H}_G \sqrt{(4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0})/B}}{2b_0 + \aleph}, \quad (1.21)$$

which is increasing in  $m_0$ . Moreover, when  $B$  is sufficiently large, we have that  $\mathcal{H}_G \sqrt{(4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0})/B} < 2b_0 + \aleph$ , showing how  $\mathcal{G}_{\text{ACC}}$  is increasing in the nominal gain  $G$ . In the above Fig. 5.3 for a realistic  $G = 6$ , we observe that a) even in the ‘optimistic’ scenario of infrequent light shadowing, the low-SNR MN gain is approximately 2, and b) that for approximately  $B = 20$ , ACC boosts MN by a factor of approximately 2.5. We will see that both the large- $B$  approximation as well as the low-SNR approximation are all tight and are numerically validated to be very close to the real performance. We consider Rayleigh/Rician-Shadowed fading channels, and we further consider the mixture Gamma (MG) distribution as a very general fading model [72], which enables us to provide a general performance analysis on the ACC-aided delivery over many common

<sup>9</sup>The setting is calibrated by the parameter  $2b_0$  for the average power of the scatter component, the parameter  $\aleph$  for the average power of the LOS component, and by the parameter  $m_0$  that reflects the (average) obstruction of the LOS component (i.e., the blockage of the LOS by buildings, trees, hills, etc.) (here  $m_0 = 0$  stands for complete obstruction, whereas  $m_0 \rightarrow \infty$  corresponds to no obstruction [71]).

practical fading channels such as Nakagami- $q$ , Nakagami- $n$ , and Nakagami-Lognormal (NL) composite fading channels. With the help of the CLT, the extended generalized bivariate Meijer's G-function, and the low-SNR robust approximation, we are able to parameterize the average rate and the effective gain of ACC respectively over this MG channel in the large- $B$  and/or the low-SNR case. We refer to Lemmas 5.5, 5.6 for more information.

Our preliminary research work for this topic is shown below.

[70] Hui Zhao, Antonio Bazco-Nogueras, and Petros Elia, "Coded caching in land mobile satellite systems," in *Proc. IEEE International Conference on Communications (ICC)*, May 2022, pp. 4571–4576.



## Chapter 2

# Aggregated Coded Caching: Design and Analysis

In this chapter, we first develop a novel delivery scheme, which we will refer to as the ACC scheme. This scheme is tailored for wireless coded caching, and what it does is that it dramatically ameliorates (and asymptotically removes) the aforementioned worst-user bottleneck that main appears in wireless channels (cf. Section 2.2). Then, in Section 2.3, we rigorously analyze the ACC average rate and effective coded caching gain in the presence of the quasi-static Rayleigh fading channels. To simplify the derived expressions and get some valuable insights, we also consider some special but practical cases in Section 2.3. We subsequently, in Section 2.4, analyze the delivery time of ACC and validate the delivery boost effect of ACC over ergodic fading channels with different pathloss in Section 2.5. We then, in Section 2.6, proceed to exploit the main idea behind the ACC scheme, in the presence of a multi-antenna BS and in the coded multicasting scenario. Finally, Section 2.7 concludes this chapter.

### 2.1 System Model and Problem Definition

We consider the quasi-static Rayleigh fading BC in which a single-antenna transmitter serves a set of  $K$  users. As mentioned before, each user requests a file from a library  $\{W_n\}_{n=1}^N$  of  $N$  files, and each user is assisted by a cache of normalized size  $\gamma \in [0, 1]$ . We consider an arbitrary number  $\Lambda$  of allowable cache states, and we assume that  $K$  is an integer multiple of  $\Lambda$ .

The received signal at user  $k \in [K]$  is given by  $y_k = h_k X_{\text{ts}} + z_k$ , where  $h_k$  denotes the channel coefficient for user  $k$ ,  $X_{\text{ts}}$  denotes the transmit signal satisfying an average power constraint  $\mathbb{E}[|X_{\text{ts}}|^2] = P_t$ , and  $z_k$  denotes the zero-mean, unit-power, additive white Gaussian noise (AWGN) at user  $k$ . Each user  $k$  experiences an instantaneous SNR of  $\text{SNR}_k = P_t |h_k|^2$ , and an average SNR of  $\rho \triangleq \mathbb{E}_h \{\text{SNR}_k\}$ . As is common in the coded caching literature (cf. [31]), we will assume that  $h_k$  remains fixed during a transmission stage, but may change between different transmission stages. We will further assume that the users experience statistically symmetric Rayleigh fading.

As it is common in works that study coded caching under quasi-static fading [3, 31], we adopt the *average rate*<sup>1</sup> as the metric of interest. Toward this, we define the instantaneous rate  $R$  as the maximal sum-rate that can be transmitted to the simultaneously served users for a instantaneous channel realization. Then, the *average rate*  $\mathbb{E}_h\{R\}$  is defined as the average (over fading statistics) of the above instantaneous rate. It is important to not confuse this long-term average  $\mathbb{E}_h\{R\}$  with the ergodic rate, since the latter implies an ability to encode over several fading realizations (cf. [31]). Henceforth, all the values for rate (bits/s) and time (s) are normalized to one Hz of bandwidth.

In this context, a coded caching scheme seeks to provide an *effective coded-caching gain*, which represents the true (multiplicative) speed up factor, at finite SNR, that the said scheme offers over the average rate obtained by TDM. This effective gain is contrasted to the (ideal, or high-SNR) *nominal coded-caching gain*, which is the gain  $G = \Lambda\gamma + 1$  provided by file-size constrained coded caching in the error-free scenario with fixed and identical link capacities.

The proposed scheme and the analysis are motivated by the fact that the effective gain of the MN scheme collapses at low SNR, which will be proven in Section 2.3. This collapse is irrespective of  $\Lambda$  and  $K$ , i.e., it happens even in the absence of file-size constraints.

## 2.2 Aggregated Coded-Caching Scheme

We now introduce a novel scheme, coined as the *Aggregated Coded Caching* ACC scheme, which will be shown to overcome the previous collapse of the effective gains. The idea behind the ACC scheme is to combine multi-rate transmission (in place of a basic XOR-based transmission) together with cache replication (a necessity in our opinion, for any realistic coded caching scenario). While the placement part of our scheme is presented here for the so-called “the  $\Lambda$ -MN” scheme in [18, Section V-A] (which is simply the MN placement but with cache replication), the ACC scheme can in fact be applied to various coded caching schemes

In particular, the ACC scheme clusters the users into  $\Lambda$  *groups* of  $B = K/\Lambda$  users per group, such that every member of the same group is assigned identical cache content (i.e., they share the same cache state). As we have seen, this is essentially inevitable under realistic file-size constraints. The scheme also follows a standard clique-based approach [1], such that the transmission is divided into *transmission stages* that experience a clique-side information pattern. As in [1], this implies that any desired subfile of some served user can be found in the cache of every other user involved in that same transmission stage. Thus, this approach matches a side-information structure that was addressed in the following well-known result from [49].

**Proposition 2.1** ([49, Thm. 6]). *The capacity region of a  $t$ -user Gaussian BC, where each user  $i \in [t]$  is endowed with SNR equal to  $\text{SNR}_i$  and requests message  $W_i'$  while having access to side information  $\mathcal{W}_i = \{W_j'\}_{j \neq i, j \in [t]}$ , is given by*

$$\mathcal{C} = \{(R_1, \dots, R_t) : 0 \leq R_i \leq \log_2(1 + \text{SNR}_i), i \in [t]\}.$$

<sup>1</sup>We recall that, for quasi-static Rayleigh fading, the typical metric of the worst-case *delivery time* does not have an expectation [54].

*Proof.* Proposition 2.1 is known as a special case of [49, Thm. 6], and this particular form has been considered in [73, 74]. More details on this, as well as on the association to our setting, are described in Appendix A.1.  $\square$

Proposition 2.1 implies that, under this particular configuration of side information, each user can achieve its own point-to-point capacity, as if no other user was being served at the same time. There are various optimal *multi-rate transmission* schemes for this setting [73, 74], and the proposed ACC scheme can remain oblivious to the encoding choice.<sup>2</sup>

**Remark 2.1.** *We state in advance that the aforementioned multi-rate transmission must indeed be combined with the method of shared caches in order to yield the desired gains. While multi-rate transmission performs better than MN-based XORs, this rate improvement appears only when we focus our attention on a single isolated delivery stage that serves some fixed set of users  $\Psi$ . However, when considering the entire delivery problem over all sets  $\Psi$ , we would see no gain, because the MN placement and multicast group generation without shared caches would not allow for an additional subfile to be sent to a potentially ‘fast’ user in  $\Psi$  without generating interference to the remaining (slower) users. This latter point, which is that the MN placement does not allow exploitation of fast users, is presented below in the original context of XORs.*

**Example 1.** *Consider the delivery of XOR  $\mathcal{A}_{2,3} \oplus \mathcal{B}_{1,3} \oplus \mathcal{C}_{1,2}$  meant for users  $\Psi = \{1, 2, 3\}$  who respectively ask for files  $W_1 = \mathcal{A}, W_2 = \mathcal{B}, W_3 = \mathcal{C}$ . Even if user 1 decodes  $\mathcal{A}_{2,3}$  very quickly, she must wait for  $\mathcal{B}_{1,3}$  and  $\mathcal{C}_{1,2}$  to be decoded, because (by definition of the MN placement) there exists only one subfile that is desired by user 1 and can be decoded by users 2 and 3. An illustrative example is shown in Fig. 2.1a.*

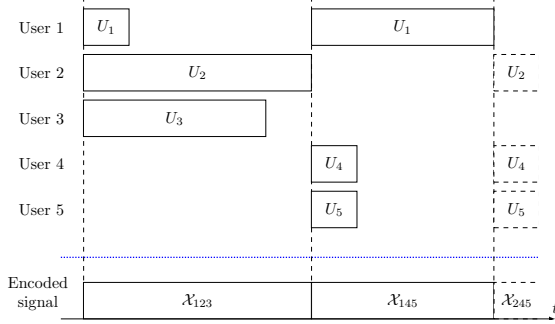
### 2.2.1 Aggregated Coded-Caching Design

We proceed with the description of the placement and delivery phases of the ACC scheme. At the end, we will also present a small clarifying example.

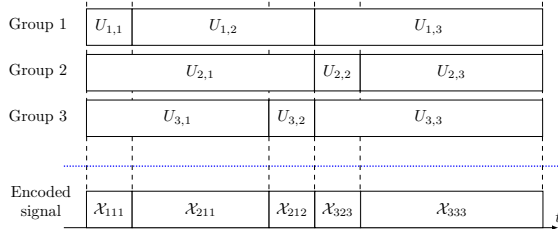
#### Placement Phase

This phase begins by arbitrarily splitting the  $K$  users into  $\Lambda$  ordered groups of  $B = \frac{K}{\Lambda}$  users each. Placement is exactly as in [18, 26], and thus it simply applies the MN placement of the  $\Lambda$ -user problem, and then each user of the same group stores the same cache content. In particular, each file  $W_n$ ,  $n \in [N]$ , is partitioned into  $\binom{\Lambda}{\Lambda\gamma}$  segments as  $W_n \rightarrow \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\}$ , and then each user in group  $g \in [\Lambda]$  stores all the subfiles belonging to the set  $\mathcal{Z}_g \triangleq \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma, \mathcal{T} \ni g, \forall n \in [N]\}$ .

<sup>2</sup>In terms of practicality, we know that very simplified schemes, such as nesting BPSK into M-QAM constellations (cf. [3]), come extremely close to achieving the above capacity region, and in fact achieve the single-user capacity when we restrict ourselves to QAM modulations [75]. Such practical codes can be directly applied in our cache-aided setting with minor performance losses.



(a) Dedicated caches: Delay depends on the worst-user capacity at each transmission stage.  $\mathcal{X}_{abc}$  denotes the signal encoded for users  $a$ ,  $b$ , and  $c$ .



(b) ACC scheme: Delay depends on the worst group sum rate.  $\mathcal{X}_{abc}$  denotes the encoded signal for users  $a$ ,  $b$ , and  $c$  of groups 1, 2, and 3, respectively.

Figure 2.1: Comparison of MN and ACC for a nominal coded-caching gain of 3.

### Delivery Phase

The delivery phase is split into  $\binom{\Lambda}{\Lambda\gamma+1}$  transmission stages, where each stage involves a set  $\Psi \subseteq [\Lambda]$  of  $|\Psi| = G = \Lambda\gamma + 1$  groups. During each stage, the transmitter *simultaneously* delivers to as many as  $\Lambda\gamma + 1$  users, each from a different group in set  $\Psi$ . The users within each group are served one after the other in a round-robin manner. For a given set  $\Psi$ , the transmitter employs a multi-rate code that achieves the capacity in Proposition 2.1, which implies that the channel state information should be available at the transmitter *only* for the  $|\Psi|$  served users. We emphasize that the multi-rate transmission in the ACC delivery does not require power-splitting and guarantees the successful information decoding in each served user due to the rate adaptation.

Let  $\Psi(i)$  denote the  $i$ -th group in  $\Psi$ ,  $i \in [|\Psi|]$  (recall the group-set  $\Psi$  is ordered). We represent the set of users that are being served at a particular time by the vector  $\mathbf{v} \in \mathbb{Z}^{|\Psi|}$ .<sup>3</sup> Consistently,  $\mathbf{v}(i) \in [B]$  tells us which user of the  $i$ -th group in  $\Psi$  is currently being served, and  $d_{\mathbf{v}(i)} \in [N]$  denotes the file index requested by user  $\mathbf{v}(i)$ . Hence, the transmitter serves the users  $\mathbf{v}$  of the groups in  $\Psi$  by transmitting

$$X_{\Psi, \mathbf{v}} = \mathcal{X}\left(\left\{W_{d_{\mathbf{v}(i)}}^{\Psi \setminus \{\Psi(i)\}}\right\}_{i \in [|\Psi|]}\right), \quad (2.1)$$

<sup>3</sup>Please note here that the dependence of  $\mathbf{v}$  on the time index and on  $\Psi$  is assumed but omitted for simplicity.

where, for any set of messages  $\mathcal{M}$ ,  $\mathcal{X}(\mathcal{M})$  denotes the transmitted signal obtained from encoding the messages in  $\mathcal{M}$  with a coding scheme achieving the capacity region in Proposition 2.1. We recall that the ACC scheme is oblivious to the selected coding scheme, as long as it achieves the capacity region. As usual in coded caching schemes,  $W_{d_{\mathbf{v}(i)}}^{\Psi \setminus \{\Psi(i)\}}$  represents the subfile intended by user  $\mathbf{v}(i)$  that is stored in the cache of all groups in  $\Psi$  except group  $\Psi(i)$ .

Algorithm 1 presents the transmission for a specific group set  $\Psi$ . Every time the user of some group  $\Psi(i')$  obtains its subfile,  $\mathbf{v}(i')$  is updated<sup>4</sup> as  $\mathbf{v}(i') \leftarrow \mathbf{v}(i') + 1$ . This process is repeated until all users in all groups in  $\Psi$  are served. If every user of a group has obtained its subfile, the transmission is composed only of the remaining groups. Algorithm 1 is iterated over all possible  $\binom{\Lambda}{\Lambda\gamma+1}$  sets  $\Psi$ . After this, the  $K$  users obtain their requested files. We reemphasize that the ACC scheme does not apply user selection. Let us proceed with a simple clarifying example.

**Example 2.** Consider a transmission stage serving groups  $\{1, 2, 3\} = \Psi$ , where each group is composed of  $B = 3$  users. To simplify the explanation of this example, let us denote the  $b$ -th user of group  $g$  as  $U_{g,b}$  and the subfile intended for this user as  $W'_{g,b}$ . Let us further assume that the normalized capacity of each user (expressed in transmitted subfiles per time slot) is as follows:

	User 1	User 2	User 3
Group 1	1	0.25	0.2
Group 2	0.2	1	0.25
Group 3	0.25	1	0.2

which simply implies that the point-to-point capacity of users  $U_{1,1}$ ,  $U_{2,2}$ , and  $U_{3,2}$  is four times the capacity of users  $U_{1,2}$ ,  $U_{2,3}$ , and  $U_{3,1}$ , and five times the capacity of  $U_{1,3}$ ,  $U_{2,1}$ , and  $U_{3,3}$ . The encoded signal for this example is illustrated in Fig. 2.1b. Initially, the first user of each group is selected to be served, and the transmitter sends  $\mathcal{X}(W'_{1,1}, W'_{2,1}, W'_{3,1})$ . Following the result of Proposition 2.1, each user can decode its own subfile at a rate matching its single-user capacity ( $\log_2(1 + \text{SNR}_{g,b})$ ) because each user knows the subfiles of the other two served users.

After the first slot, user  $U_{1,1}$  has successfully decoded its subfile. Hence,  $U_{1,1}$  is substituted by  $U_{1,2}$ , and the transmitter sends  $\mathcal{X}(W'_{1,2}, W'_{2,1}, W'_{3,1})$ . The key is that we can serve any of the users storing the same cache state because all of them can cache out the subfiles intended by the users of the other groups in  $\Psi$ , and vice versa. Thus, every time a user obtains its subfile, a new member of the same group substitutes this user, while the other served users can continue decoding their subfile. In the same way,  $U_{3,1}$  obtains its subfile after the fourth time slot, it is replaced by  $U_{3,2}$ , and the transmitter then sends  $\mathcal{X}(W'_{1,2}, W'_{2,1}, W'_{3,2})$ . After the fifth slot, the three users obtain their subfile and the transmitter starts sending  $\mathcal{X}(W'_{1,3}, W'_{2,2}, W'_{3,3})$ , and so on.

<sup>4</sup>We are actually incurring an abuse of notation in (2.1) and Algorithm 1. Specifically, when a group updates its served user, the transmitter continues encoding the partially-decoded subfiles taking into account that there remains only a part of such subfiles to be transmitted. This is intuitive from Fig. 2.1b.

---

**Algorithm 1:** Transmission stage for a set of groups  $\Psi$ 


---

```

1 Initialize  $\mathbf{v} \in \mathbb{Z}^{|\Psi|}$  as  $\mathbf{v}(i) \leftarrow 1$  for any  $i \in [|\Psi|]$ 
2 Initialize Number of finished groups  $\leftarrow 0$ 
3 while Number of finished groups  $\neq |\Psi|$  do
4     Transmit
5     |  $X_{\Psi, \mathbf{v}} \leftarrow \mathcal{X}\left(\left\{W_{d_{\mathbf{v}(i)}}^{\Psi \setminus \{\Psi(i)\}} \mid i \in [|\Psi|] \text{ and } \mathbf{v}(i) \leq B\right\}\right)$ 
6     until  $A$  served user  $\mathbf{v}(i)$ ,  $i \in [|\Psi|]$ , fully obtains its subfile
7     Set  $i^*$  as the index of the group  $\Psi(i^*)$  whose user has decoded its subfile
8     if  $\mathbf{v}(i^*) = B$  then
9     | Number of finished groups  $\leftarrow$  Number of finished groups + 1
10     $\mathbf{v}(i^*) \leftarrow \mathbf{v}(i^*) + 1$ 

```

---

## 2.3 Average Rate Analysis

In this section, we analyze the long-term average rate of the  $\Lambda$ -MN and ACC schemes. First, we will derive the exact expression of the average rate for both schemes. Afterward, we will approximate this rate at low SNR, and we will also derive the limit in the regime of many users. It will turn out, as we will see in the following, that these two approximations are very robust in realistic scenarios. Furthermore, we obtain the effective gain of this scheme with respect to TDM as well as its improvement with respect to the  $\Lambda$ -MN scheme, and we show that while the effective gain of the  $\Lambda$ -MN scheme vanishes at low SNR, the ACC scheme recovers — at any SNR value — the nominal (high-SNR) gain as the number of users per cache increases.

We recall that, under Rayleigh fading, the SNR follows an exponential distribution. Hence, for user  $k \in [K]$ , the probability density function (PDF) and cumulative distribution function (CDF) of  $\text{SNR}_k$  are given respectively by  $f_{\text{SNR}_k}(x) = \frac{1}{\rho} \exp\left(-\frac{x}{\rho}\right)$  and  $F_{\text{SNR}_k}(x) = 1 - \exp\left(-\frac{x}{\rho}\right)$ , for any  $x \geq 0$ , where  $\rho = \mathbb{E}_h\{\text{SNR}_k\}$  denotes the average SNR with respect to channel states. We recall that the user channels are statistically symmetric. As for the ACC scheme, we will use  $\text{SNR}_{g,b}$ ,  $f_{\text{SNR}_{g,b}}(x)$ , and  $F_{\text{SNR}_{g,b}}(x)$  to refer to the SNR, PDF, and CDF corresponding to the  $b$ -th user of the group  $g$ , where  $b \in [B]$  and  $g \in [\Lambda]$ .

### 2.3.1 Average Rate of the $\Lambda$ -MN and ACC Schemes

#### Average Rate of the $\Lambda$ -MN Scheme

We first note that the  $\Lambda$ -MN is an adaptation for finite-file sizes settings from [18] of the standard MN scheme of [1]. Placement is analogous to the one of the ACC scheme, and the transmission consists of repeating  $B$  times the transmission of the dedicated caches setting. Consequently, the  $\Lambda$ -MN scheme consists of  $B \binom{\Lambda}{\Lambda\gamma+1}$  transmission stages, each of them employed to deliver an XOR to a group of users of size  $|\Psi| = G = \Lambda\gamma + 1$ .

Consider the delivery to a particular set  $\Psi$  of  $\Lambda\gamma + 1$  users. We know from the multicast capacity theorem in [23] that the maximum instantaneous rate for any user  $i \in \Psi$  takes the form

$$R_{i,\Psi}^{\text{MN}} = \log_2 \left( 1 + \min_{k \in \Psi} \text{SNR}_k \right) \quad \text{bits/s}, \quad (2.2)$$

where the minimum operator guarantees the successful information decoding at all the users in  $\Psi$ . Note that the delay (or delivery time) required to transmit one sub-file to every user in  $\Psi$  at this transmission stage is given by

$$T_{\text{MN},\Psi} = \frac{F}{\binom{\Lambda}{\Lambda\gamma}} \left[ \log_2 \left( 1 + \min_{k \in \Psi} \text{SNR}_k \right) \right]^{-1} \quad \text{s}, \quad (2.3)$$

where  $F$  is the total information bits of a file, and  $F/\binom{\Lambda}{\Lambda\gamma}$  is the size of the subfiles generated from subpacketization. Since  $\min_{k \in \Psi} \text{SNR}_k$  follows an exponential distribution with rate  $|\Psi|/\rho$ , the expectation of  $T_{\text{MN},\Psi}$  diverges. For this reason, we consider the average rate as a main metric of interest, which crisply reflects the worst-user effect.

The instantaneous sum rate is given by  $\sum_{i \in \Psi} R_{i,\Psi}^{\text{MN}}$ , since we are simultaneously serving all the  $|\Psi|$  users. Consequently, the average (sum) rate for that specific set  $\Psi$  takes the form

$$\bar{R}_{\Psi}^{\text{MN}} \triangleq \mathbb{E}_h \left\{ \sum_{i \in \Psi} R_{i,\Psi}^{\text{MN}} \right\} = \frac{G}{\ln 2} \mathbb{E}_h \left\{ \ln(1 + \min_{k \in \Psi} \text{SNR}_k) \right\} \quad (2.4)$$

which follows because the users are statistically equivalent, which in turn also implies that the average sum rate  $\bar{R}^{\text{MN}}$  remains the same for any set  $\Psi$ , i.e., it implies that  $\bar{R}^{\text{MN}} = \bar{R}_{\Psi'}^{\text{MN}} \forall \Psi' \subseteq [\Lambda], |\Psi'| = \Lambda\gamma + 1$ .

Naturally, the average rate under the TDM scheme, which we denote as  $\bar{R}^{\text{TDM}}$ , is a special case of  $\bar{R}^{\text{MN}}$  obtained by setting  $G = 1$ . The variable  $\min_{k \in \Psi} \{\text{SNR}_k\}$  is the minimum of  $G$  i.i.d. exponential variables of rate  $\frac{1}{\rho}$  (i.e., mean  $\rho$ ), and, consequently, it follows an exponential distribution with rate  $G/\rho$  (or mean  $\rho/G$ ). Thus, it follows from [50, Eq. (15.26)] that

$$\bar{R}^{\text{MN}} = -\frac{G}{\ln 2} \exp\left(\frac{G}{\rho}\right) \cdot \text{Ei}\left(-\frac{G}{\rho}\right), \quad (2.5)$$

where  $\text{Ei}(\cdot)$  represents the exponential integral function [76]. Note that  $G = 1$  in (2.5) yields the closed-form expression for  $\bar{R}^{\text{TDM}}$ .

### Average Rate of the ACC Scheme

Due to the symmetry of the ACC scheme and the statistical symmetry of the channel, we now focus on a particular set  $\Psi$  of  $|\Psi| = G = \Lambda\gamma + 1$  user groups, where we recall that each group is composed of  $B$  users.

As explained in Section 2.2, the ACC scheme allows us to serve some user  $b$  of group  $g$  at its own point-to-point capacity, and it allows us to immediately start serving another

user of the same group as soon as the said user  $b$  has completed the decoding of its subfile. Furthermore, in the ACC scheme, the delivery to a group-set  $\Psi$  is completed when every user belonging to one of these groups has obtained its subfile. Consequently, the resulted delay (or delivery time) to serve all user-groups in  $\Psi$  (which include  $|\Psi| \cdot B$  users) is given by

$$T_{\text{ACC},\Psi} = \frac{F}{(\Lambda^\gamma)} \max_{g \in \Psi} \sum_{b=1}^B \left[ \log_2(1 + \text{SNR}_{g,b}) \right]^{-1} \text{ s}, \quad (2.6)$$

which will become (2.3) for  $B = 1$ . As for (2.3), the expectation of (2.6) diverges. Then, as explained for the  $\Lambda$ -MN scheme, we consider the average rate. The (per-user average) rate with which any group  $j$  in the set  $\Psi$  is served is here captured by

$$R_{j,\Psi}^{\text{ACC}} = \min_{g \in \Psi} \frac{1}{B} \sum_{b=1}^B \log_2(1 + \text{SNR}_{g,b}) \text{ bits/s} \quad (2.7)$$

for all  $j \in \Psi$ . By applying the same reasoning as in (2.2)–(2.4), we obtain that the average rate with which the transmitter delivers data across the users is given by

$$\bar{R}^{\text{ACC}} = \frac{G}{\ln 2} \mathbb{E}_h \left\{ \min_{g \in \Psi} \frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right\} \text{ bits/s}. \quad (2.8)$$

We quickly note that, by comparing (2.8) with (2.4), we can see how the worst-user effect is essentially averaged out into a cumulative “worst-group” effect. By considering dedicated caches (i.e.,  $B = 1$ ), we obtain the same average rate as that of the  $\Lambda$ -MN scheme in (2.4) despite having a different (not XOR-based) coding scheme, which is consistent with Remark 2.1.

In the following,  $j \triangleq \sqrt{-1}$  denotes the imaginary unit,  $\text{Im}\{\cdot\}$  denotes the imaginary part of a complex number, and  $\text{E}_{-jt}(\cdot)$  denotes the exponential integral function of the  $(-jt)$ -th order [76]. Next, we present our first main result.

**Lemma 2.1.** *The exact average rate of the ACC scheme over symmetric quasi-static Rayleigh fading can be derived in a double-integral form, which takes the form*

$$\bar{R}^{\text{ACC}} = \frac{G}{B \ln 2} \int_0^\infty \left( \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\text{Im} \left\{ \exp(-jxt) \frac{\exp(B/\rho)}{\rho^B} \text{E}_{-jt}^B \left( \frac{1}{\rho} \right) \right\}}{t} dt \right)^G dy.$$

*Proof.* The proof is relegated to Appendix A.2. □

The numerical implementation of the above expression is very complex and it provides little insight. In the following, we obtain the effective gains in both the low-SNR limit and the large- $B$  limit for the  $\Lambda$ -MN scheme and the ACC scheme, and we derive approximations of their rates, from which some meaningful insights can be easily drawn.

### 2.3.2 Rate Approximations and Effective Gains at Low SNR

#### $\Lambda$ -MN Scheme analysis

First, we present a low-SNR approximation for the average rate of the  $\Lambda$ -MN scheme, which is in fact a special case of the ACC scheme with  $B = 1$ . Although the exact form has been derived in (2.5), we can provide a simple but tight approximation which allows us to remove the special function  $\text{Ei}(\cdot)$  from the expression.

**Lemma 2.2.** *In the low-SNR region, the average rate of the  $\Lambda$ -MN scheme can be approximated by*

$$\bar{R}^{\text{MN}} \approx \frac{G}{\ln 2} \left( \ln \left( 1 + \frac{\rho}{G} \right) - \frac{\rho^2}{2G^2 (1 + \rho/G)^2} \right). \quad (2.9)$$

*Proof.* See Appendix A.3.1. □

In the numerical evaluation section (cf. Fig. 2.5), it will be shown that this computationally efficient second-order approximation can in fact provide us with an extremely reliable estimation of the performance even in the medium-SNR region.

Let us now consider the exact effective gain of the  $\Lambda$ -MN scheme, which, directly from (2.5), takes the form

$$\frac{\bar{R}^{\text{MN}}}{\bar{R}^{\text{TDM}}} = \frac{G \exp\left(\frac{G}{\rho}\right) \cdot \text{Ei}\left(-\frac{G}{\rho}\right)}{\exp\left(\frac{1}{\rho}\right) \cdot \text{Ei}\left(-\frac{1}{\rho}\right)}. \quad (2.10)$$

As expected, the effective gain converges to the nominal gain  $G$  at high SNR, since the limit of (2.10) as  $\rho \rightarrow \infty$  is  $G$ . On the other hand, in the low-SNR region, this effective gain entirely vanishes, as stated in the following proposition.

**Proposition 2.2.** *For any value of  $K$  and  $\Lambda$ , the effective gain of the  $\Lambda$ -MN scheme converges to*

$$\lim_{\rho \rightarrow 0} \frac{\bar{R}^{\text{MN}}}{\bar{R}^{\text{TDM}}} = 1 \quad (2.11)$$

*meaning that this effective coded-caching gain entirely vanishes at low SNR.*

*Proof.* See Appendix A.3.2. □

As noted before, Proposition 2.2 holds for any scheme which requires decoding of single XORs.

### ACC Scheme Analysis

After presenting the previous result for the  $\Lambda$ -MN scheme, let us now consider the ACC scheme. In the following, for any integer vector  $\mathbf{b} \triangleq [b_1, b_2, \dots, b_B] \in \mathbb{Z}^B$  composed of  $B$  non-negative elements, we will use

$$\binom{n}{\mathbf{b}} \triangleq \frac{n!}{b_1! b_2! \dots b_B!} \quad (2.12)$$

to denote the multinomial coefficient. We can now state our following result, which presents an expression of the rate of the ACC scheme for the low-SNR regime.

**Lemma 2.3.** *In the low-SNR region, the average rate of the ACC scheme can be approximated by  $\bar{R}^{\text{ACC}} \approx \frac{\rho G}{B \ln 2} \mathcal{L}_G$ , since it holds that*

$$\bar{R}^{\text{ACC}} = \frac{\rho G}{B \ln 2} \mathcal{L}_G + o(\rho), \quad (2.13)$$

where  $\mathcal{L}_G$  is defined as

$$\mathcal{L}_G \triangleq \sum_{\|\mathbf{b}\|_1=G} \binom{G}{\mathbf{b}} \frac{G^{-1-\sum_{t=1}^B (t-1)b_t}}{\prod_{t=1}^B ((t-1)!)^{b_t}} \left( \sum_{t=1}^B (t-1)b_t \right)!,$$

and where the sum is over all the vectors composed of  $B$  non-negative integer elements and whose norm-1 equals  $G$ .

*Proof.* The proof is relegated to Appendix A.3.3.  $\square$

From Lemma 2.3 and Proposition 2.2, we obtain a corollary on the gain of the ACC scheme over the  $\Lambda$ -MN scheme.

**Corollary 2.1.** *In the limit of low SNR, the ratio of  $\bar{R}^{\text{ACC}}$  over  $\bar{R}^{\text{MN}}$  converges to the constant*

$$\lim_{\rho \rightarrow 0} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} = \frac{G}{B} \mathcal{L}_G \quad (2.14)$$

where we recall that  $G = \frac{K}{B} \gamma + 1$ .

*Proof.* The proof is relegated to Appendix A.3.4.  $\square$

The expression in Corollary 2.1 is illustrated in Fig. 2.2 for different values of  $B$  and  $G$ .

**Remark 2.2.** *In Fig. 2.2, we can see that  $\frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}}$  is concave with respect to  $B$ , and that this concavity increases with  $G$ . This signals that, for large  $G$ , most of the gain from having  $B > 1$  is obtained quickly, at relatively small values of  $B$ . For example, when  $G = 100$  (which is unrealistic), we see that the ACC rate for  $B = 2$  is up to 20 times higher than the  $\Lambda$ -MN rate ( $B = 1$ ).*

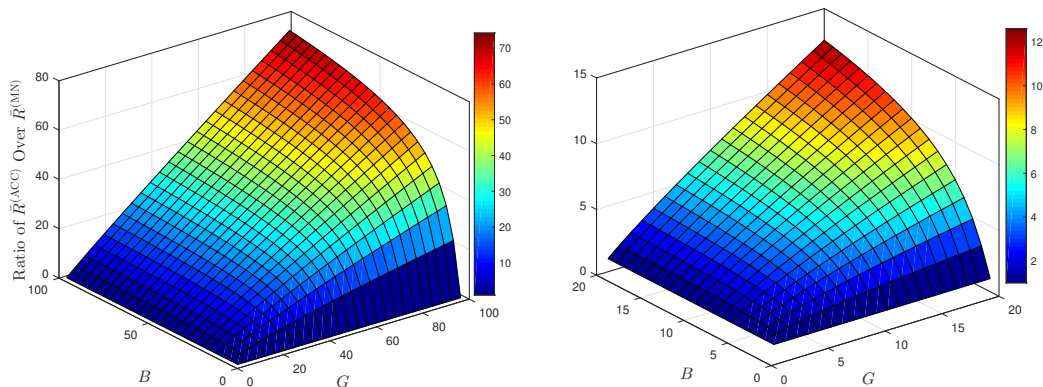


Figure 2.2: The ACC improvement  $\left(\frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}}\right)$  over the MN scheme in Corollary 2.1 for different  $B$  and  $G$ .

### 2.3.3 Effective Gain in the Large- $B$ Region

We now move away from the low-SNR regime, and we consider instead the limit of many users. This regime is nicely motivated by the ever increasing density of users in wireless networks. Therefore, we consider that  $\Lambda$  remains fixed and  $K$  can grow unboundedly, which also implies that  $B \rightarrow \infty$  since  $B = K/\Lambda$ . The following shows that, in the limit of many users, the effective gain of the ACC scheme matches — for any SNR value — the nominal gain.

**Lemma 2.4.** *For any average SNR  $\rho$ , the ACC scheme guarantees*

$$\lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{TDM}}} = \Lambda\gamma + 1, \quad (2.15)$$

and, thus, its effective gain matches the nominal gain for any value of SNR.

*Proof.* The proof is relegated to Appendix A.3.5. □

We now proceed to compare the ACC scheme with the  $\Lambda$ -MN scheme, again in the limit of large  $B$ . We will also obtain the low-SNR approximation of this comparison, which nicely captures scenarios such as cell-free or satellite networks, where the majority of the users is distributed in the edge area and/or suffers from heavy path-loss or heavy shadowing.

**Lemma 2.5.** *In a setting with  $\Lambda$  cache states and  $K = \Lambda B$  users, and for any average SNR  $\rho$ , the ratio  $\frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}}$  satisfies*

$$\lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} = \exp\left(\frac{1-G}{\rho}\right) \frac{\text{Ei}\left(-\frac{1}{\rho}\right)}{\text{Ei}\left(-\frac{G}{\rho}\right)}. \quad (2.16)$$

Furthermore, it holds that

$$\lim_{\rho \rightarrow 0} \lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} = \Lambda\gamma + 1. \quad (2.17)$$

*Proof.* The proof is relegated to Appendix A.3.6. Note that (2.17) follows from (2.16), but the same conclusion can be seen directly by combining Proposition 2.2 and Lemma 2.4.  $\square$

**Remark 2.3.** *The key for recovering the nominal gain is that a larger  $B$  implies a smaller fluctuation around the average transmission rate within a user group, which inherently reduces the impact of the worst-user (or worst-group) bottleneck.*

**Remark 2.4.** *As previously mentioned, the preservation of the nominal gains in Lemma 2.5 would also hold for other coded caching schemes. Indeed, works that focus on the finite file-size constraint are normally based on XOR transmissions, and thus they are not robust to the worst-user effect. Consequently, we can improve the low-SNR performance in scenarios such as those found in [20, 21, 26] by incorporating our approach of multi-rate transmission and cache replication into these schemes. This is shown in Fig. 2.11 and the corresponding text for the setting of [20].*

### 2.3.4 High-Fidelity Approximation of $\bar{R}^{\text{ACC}}$ for Any SNR Value

The previous subsections offered crisp and insightful approximations of the performance of the ACC scheme. We now take a step back and seek to provide high-accuracy approximations that can be evaluated very easily.

Indeed, both the exact value of  $\bar{R}^{\text{ACC}}$  in Lemma 2.1 and the approximation at low SNR in Lemma 2.3 have time-consuming implementations when  $B$  is large. To counter this, we now provide a simple but very precise large- $B$  approximation of  $\bar{R}^{\text{ACC}}$ , which accurately approximates the average rate even if  $B$  is relatively small. This expression involves the well-known Q-function  $\mathcal{Q}(\cdot)$ , i.e., the tail distribution function of the standard normal distribution, and the Meijer's G-function  $G_{\cdot}(\cdot)$  defined in [76, Eq. (9.301)].

Before presenting the new approximation, let us denote the expectation of the maximum of  $G$  i.i.d. standard normal random variables by  $\mathcal{H}_G$ . Consequently, the expectation of the minimum of such set of variables is given by  $-\mathcal{H}_G$ . We can now present our next result.

**Lemma 2.6.** *In the large- $B$  regime, the average rate of the ACC scheme can be approximated by*

$$\bar{R}^{\text{ACC}} \approx \frac{G}{\ln 2} \left( \varrho - \frac{\sigma}{\sqrt{B}} \times \mathcal{H}_G \right), \quad (2.18)$$

where  $\varrho$  and  $\sigma$  respectively represent the average and the standard deviation of  $\ln(1 + \text{SNR}_{g,b})$  for  $g \in [A]$  and  $b \in [B]$ , which are given by

$$\varrho = -\exp\left(\frac{1}{\rho}\right) \cdot \text{Ei}\left(-\frac{1}{\rho}\right), \quad (2.19)$$

$$\sigma = \sqrt{2 \exp\left(\frac{1}{\rho}\right) G_{2,3}^{3,0}\left(\frac{1}{\rho} \middle| \begin{matrix} 1,1 \\ 0,0,0 \end{matrix}\right) - \rho^2}. \quad (2.20)$$

*Proof.* See Appendix A.4.  $\square$

The term  $\mathcal{H}_G$  is given by the following integral form,

$$\mathcal{H}_G = \frac{-G}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y \left( \mathcal{Q}(y) \right)^{G-1} \exp\left(-\frac{y^2}{2}\right) dy, \quad (2.21)$$

and the proof of (2.21) is relegated to Appendix A.4.

At this point, we note that the value of  $\mathcal{H}_G$  for  $G = 1, 2, 3, 4, 5$  is known (cf. [77, Sec. 5.16]) and it is given in the following table.

Table 2.1: Exact Value of  $\mathcal{H}_G$  for  $G \leq 5$

$G$	1	2	3	4	5
$\mathcal{H}_G$	0	$\pi^{-1/2}$	$\frac{3}{2}\pi^{-1/2}$	$3\pi^{-3/2} \cos^{-1}\left(-\frac{1}{3}\right)$	$\frac{5}{2}\pi^{-3/2} \cos^{-1}\left(-\frac{23}{27}\right)$

For larger values of  $G$ , there are not known closed-form expressions, but it is known (cf. [78]) that one can have a simple approximation by substituting  $\mathcal{H}_G$  by  $\sqrt{2 \ln(G)}$ . This approximation is based on the fact that  $\mathcal{H}_G$  is bounded as  $\frac{1}{\sqrt{\pi \ln 2}} \sqrt{\ln(G)} \leq \mathcal{H}_G \leq \sqrt{2 \ln(G)}$ , and the fact that  $\lim_{G \rightarrow \infty} \frac{\mathcal{H}_G}{\sqrt{\ln(G)}} = \sqrt{2}$  (cf. [78]).

In order to obtain a better approximation of  $\mathcal{H}_G$  than  $\sqrt{2 \ln(G)}$ , which is simple but only accurate for large values of  $G$ , a very interesting approximation is to adopt the Gauss-Hermite quadrature (GHQ) [79, Ch. 9], which nicely balances high accuracy and low complexity. Applying this method to the specific integral form in (2.21) yields

$$\mathcal{H}_G \approx \frac{-\sqrt{2}G}{\sqrt{\pi}} \sum_{v=1}^V \omega_v x_v \left( \mathcal{Q}(\sqrt{2}x_v) \right)^{G-1}, \quad (2.22)$$

where  $V$ ,  $x_v$ , and  $\omega_v$  are the summation terms, sample points and weights in the GHQ, respectively. Generally speaking, we can get an approximate result with high accuracy by summing up several terms in the GHQ.

### 2.3.5 Numerical Results for the ACC scheme

In this subsection, we illustrate through numerical analysis both the exact results and the previously obtained approximations<sup>5</sup>. The derived approximations on the average rate are computationally efficient, can handle large-dimensional problems, and, as we will show via Monte-Carlo simulations, tightly approximate the true performance of the algorithms. In the following, we characterize the different considered scenarios in the simulations by

<sup>5</sup>For the convenience of annotation in the simulation figures, we omit the chapter labels of theorems, lemmas, corollaries, propositions and equations. As the numerical results are independent across different chapters, this kind of omission does not bring about any confusion.

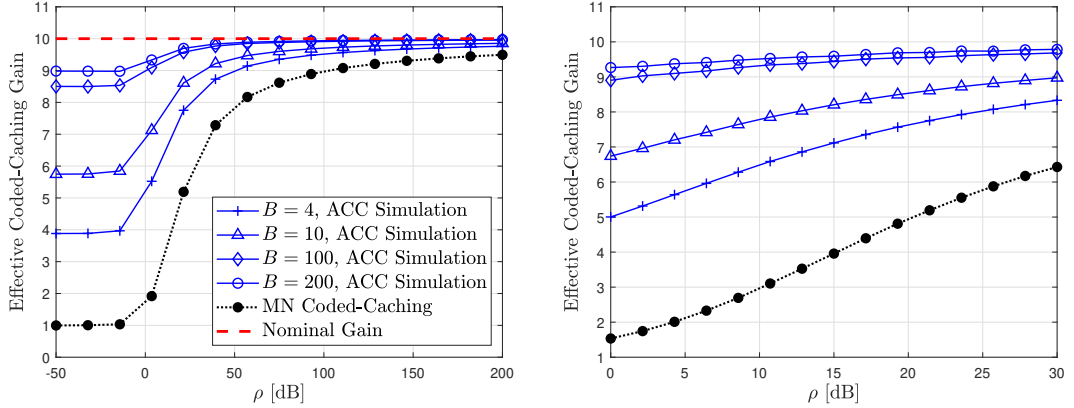


Figure 2.3: Effective gain versus  $\rho$  for  $G = 10$ . Right-side plot focuses on realistic SNR values.

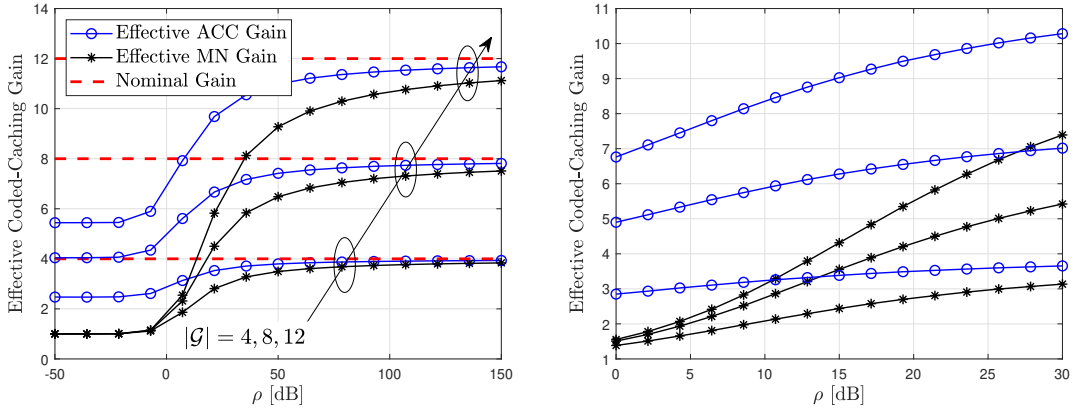


Figure 2.4: Effective gain versus  $\rho$  for  $B = 6$ . Right-side plot focuses on realistic SNR values.

the parameters  $B$  and  $G$ . Note that the use of these two parameters can apply to various  $K, \Lambda, \gamma$  scenarios, where the relation follows from the fact that  $G = \Lambda\gamma + 1 = \frac{K}{B}\gamma + 1$ .

To motivate the values of  $B$  that we use, let us consider a scenario with  $\gamma = 10\%$  and a realistic subpacketization limit of  $10^5$ . For a file size of  $10^8$  bytes, this implies an atomic sub-file size of about 1000 bytes. This gives  $\Lambda = \arg \max_{x \in \mathbb{Z}} \left\{ \binom{x}{0.1x} < 10^5 \right\} \approx 40$ , which means that having  $K = 800$  users reasonably allows for  $B$  up to 20. Such (or even higher) values of  $K$  are motivated by several different scenarios [80, 81]. In order to obtain the simulation results with high accuracy,  $10^6$  channel states are generated and averaged over Rayleigh fading.

### ACC and MN Effective Gains With Respect to TDM

In Figs. 2.3–2.4, we present the effective coded-caching gains of the ACC and MN schemes versus  $\rho$ , for different values of  $B$  and different nominal gains ( $G$ ). As expected, the effective gains of both the ACC scheme and the MN scheme converge to the nominal gain as  $\rho$  increases. However, the convergence of the ACC scheme is much faster than that of the MN scheme and, furthermore, the convergence of the ACC scheme becomes faster as  $B$  grows.

From the same figures, it is also worth noting that, when  $\rho$  is relatively small, the effective coded-caching gains of both schemes arrive to a flat lower bound. The lower bound for the ACC scheme is notably greater and improves as either  $B$  or  $G$  become bigger. However, this behavior does not extend to the MN scheme, which is consistent with the result of Proposition 2.2 stating that the effective gain of the MN scheme collapses at low SNR regardless of the value of the high-SNR caching gain  $G$ . Moreover, in Fig. 2.4, we can see that for the MN scheme the worst-user effect is amplified as  $G$  increases. Therefore, Figs. 2.3–2.4 show that the advantages of the ACC scheme in terms of average rate are still significant even for a small group size ( $B = 4, 6$ ).

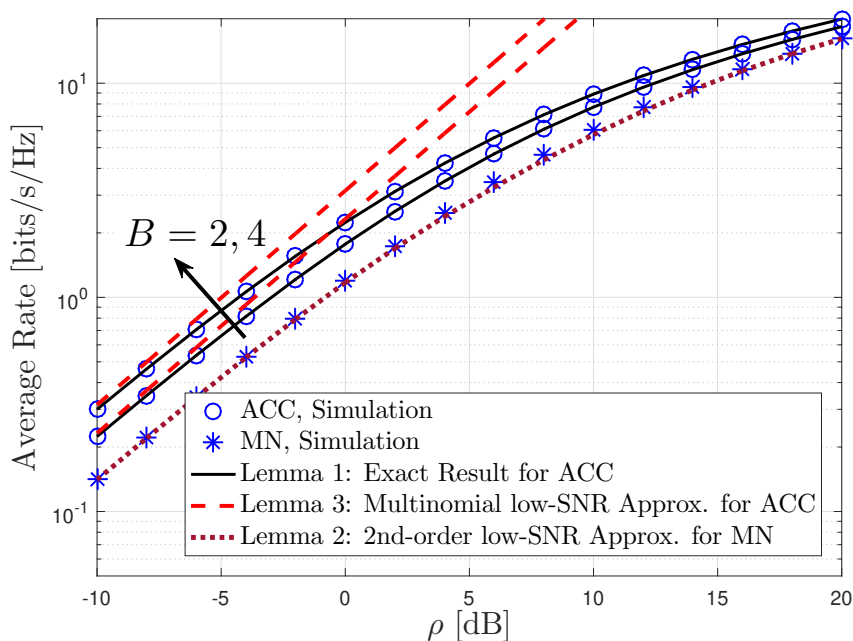
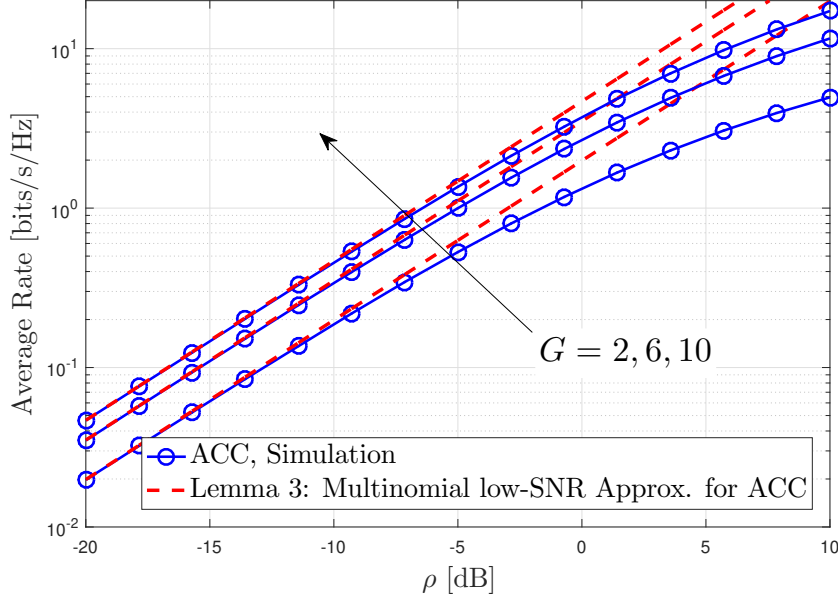


Figure 2.5:  $\bar{R}^{\text{ACC}}$  versus  $\rho$  for  $G = 4$ .

### Approximations on the Average Rate $\bar{R}^{\text{ACC}}$

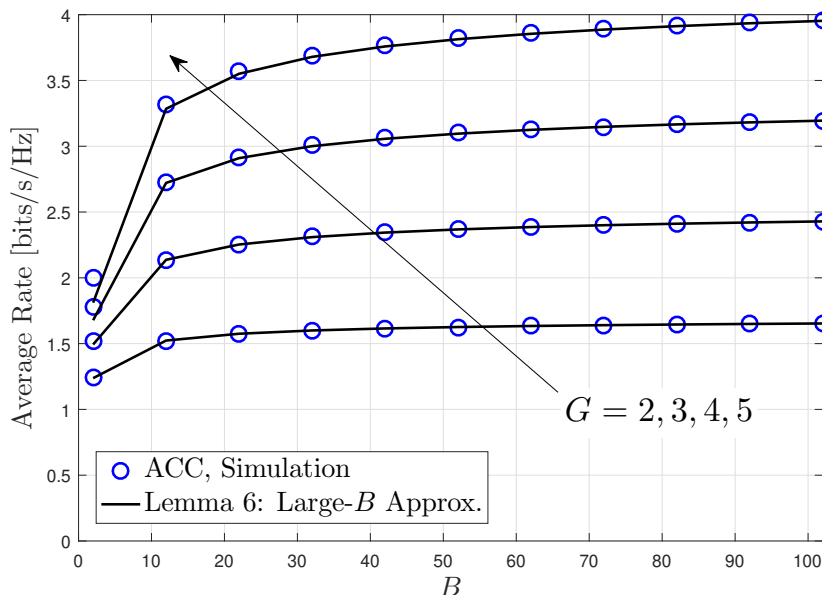
In Figs. 2.5–2.8, we validate the derived analytical approximations and highlight some interesting trends and comparisons. First, Fig. 2.5 shows the average rate  $\bar{R}^{\text{ACC}}$  versus  $\rho$  for different values of  $B$ . Note that, for  $B = 1$ ,  $\bar{R}^{\text{ACC}} = \bar{R}^{\text{MN}}$ . For comparison, Fig. 2.5


 Figure 2.6:  $\bar{R}^{\text{ACC}}$  versus  $\rho$  for  $B = 3$ .

displays the simulated result (circle and asterisk symbols), the exact derived average rate  $\bar{R}^{\text{ACC}}$  in Lemma 2.1 (solid line), the low-SNR multinomial approximation in Lemma 2.3 (dashed line), and the low-SNR second-order approximation for  $\bar{R}^{\text{MN}}$  in Lemma 2.2 (dotted line). The rate enhancement due to the ACC scheme is exhibited by comparing the results of Lemma 2.1 and Lemma 2.2 (solid and dotted lines, respectively). Fig. 2.5 shows that the accuracy of the approximation for  $\bar{R}^{\text{MN}}$  in Lemma 2.2 is better than the approximation for  $\bar{R}^{\text{ACC}}$  in Lemma 2.3, mainly because Lemma 2.3 considers a first-order approximation. Fig. 2.6 reveals that the approximation derived in Lemma 2.3 becomes more accurate as  $G$  increases, which indicates that the value of  $\rho$  at which the nonlinear part of the average rate becomes significant increases as  $G$  increases.

The large- $B$  approximation of  $\bar{R}^{\text{ACC}}$  from Lemma 2.6 is validated in Fig. 2.7, where the average rate is plotted for different  $G$ . The values of  $\mathcal{H}_G$  are taken from Table 2.1. This large- $B$  approximation tightly approximates the simulation results, even for a small  $B$ . In fact, this approximation is extremely tight for any value of  $B$  bigger than 1. To further demonstrate the accuracy of Lemma 2.6, we show in Fig. 2.8 the results derived by using *i*) the integral calculation in (2.21), *ii*) the GHQ method in (2.22), and *iii*) the  $\sqrt{2 \ln(G)}$  approximation of  $\mathcal{H}_G$  for  $G > 5$ .

After verifying the high accuracy of the approximation in Lemma 2.6, we exploit it to present some interesting comparisons between the ACC scheme and the MN scheme in Figs. 2.9–2.10. In Fig. 2.9, we can see through the ratio  $\frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}}$  that  $\bar{R}^{\text{ACC}}$  provides significant boost for realistic SNR values. In order to illustrate the extent to which this ratio approaches the theoretical gain in the low-SNR regime, we show in Fig. 2.10 the different ratios/improvements achieved by varying  $B$ .

Figure 2.7:  $\bar{R}^{\text{ACC}}$  versus  $B$  for  $\rho = 0$  dB.

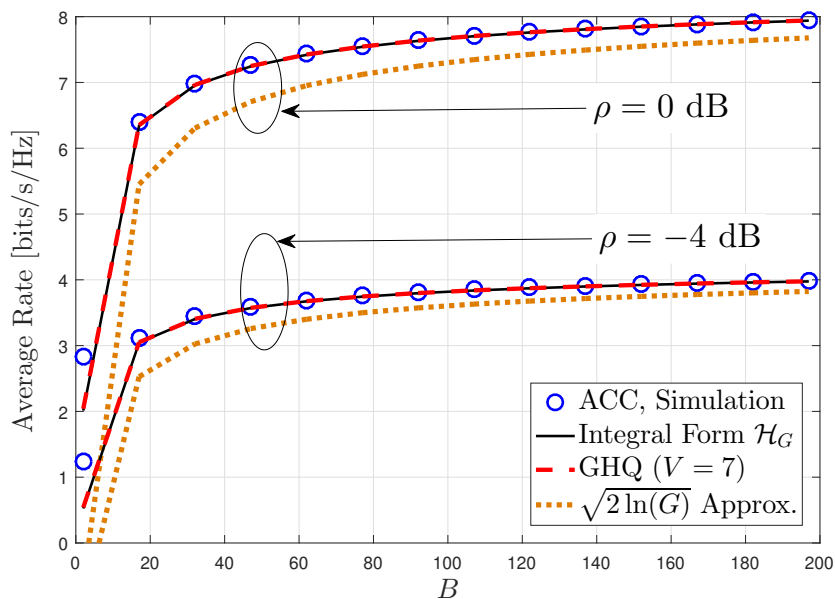
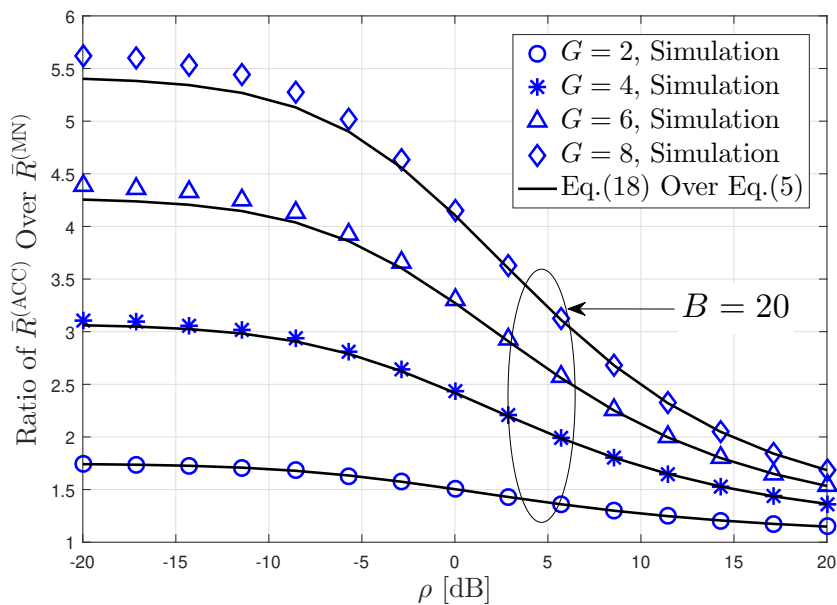
### Analysis of Delivery Time with Decentralized-Placement

In order to show the generality of the key ideas underlying the ACC scheme, we provide an example of its application in a decentralized coded caching setting with finite file-size constraints. We then compare our new decentralized scheme with the state-of-the-art scheme from [20].

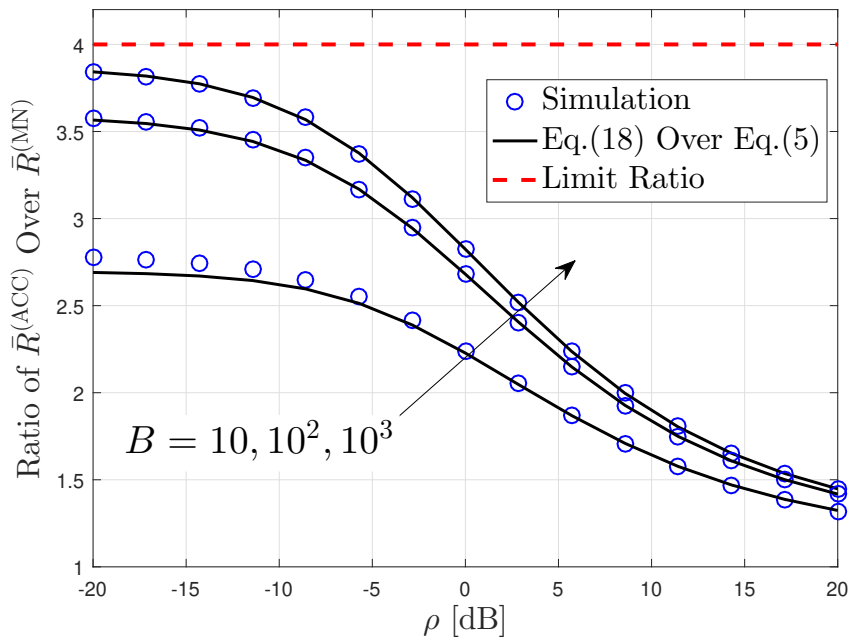
In the decentralized scenario of [20], the subpacketization constraint induces a certain number  $\Lambda$  of cache states. The main difference with our previous setting is that, during placement, each user *independently* selects one of the  $\Lambda$  cache states *uniformly at random* and stores it in its cache, such that each of the  $\Lambda$  cache states will be stored at a different number of users. Let  $B_g$  denote the number of users storing the  $g$ -th cache state, and note that  $\sum_{g=1}^{\Lambda} B_g = K$ .

During the delivery phase, the scheme in [20] first serves one user from each cache state by implementing sequential XOR transmissions as if we applied the standard MN scheme for  $\Lambda$  users. After that, the transmission procedure is repeated for the next user of each cache state for the cache states still including some not-served users. We refer to [20] for more details.

We recall that, in contrast to the scheme from [20], the proposed ACC scheme sequentially serves all users in a set of cache states  $\Psi$  of  $G = \Lambda\gamma + 1$  cache states (i.e.,  $\sum_{g \in \Psi} B_g$  users). Once all these users have received their subfile, the transmitter starts to serve another cache-state set  $\Psi'$ . As mentioned in Remark 2.4, the ACC scheme can be directly applied to the case in which each cache state is stored at a different number of users.


 Figure 2.8:  $\bar{R}^{\text{ACC}}$  versus  $B$  for  $G = 10$ .

 Figure 2.9:  $\bar{R}^{\text{ACC}}/\bar{R}^{\text{MN}}$  versus  $\rho$  for  $V = 7$  in GHQ.

Let us now analyze the benefits of using the ideas from the ACC scheme in this setting. Since now the number of users per cache state may (and probably will) differ,


 Figure 2.10:  $\bar{R}^{\text{ACC}}/\bar{R}^{\text{MN}}$  versus  $\rho$  for  $G = 4$ .

we need to consider the average delivery time instead of the average rate. Note, however, that the delivery time over Rayleigh fading channels does not converge, as previously mentioned. Hence, for comparative purposes, we consider Nakagami- $m$  fading to model the wireless propagation [50]. The delivery time of the centralized ACC scheme over Nakagami- $m$  fading channels has been recently analyzed in [54].

We can obtain from (2.6) the total delivery time of the decentralized ACC scheme as

$$T_{\text{ACC}} = \frac{F}{\binom{\Lambda}{\Lambda\gamma}} \sum_{\substack{\Psi \subseteq [\Lambda] \\ |\Psi| = \Lambda\gamma + 1}} \max_{g \in \Psi} \left\{ \sum_{b=1}^{B_g} [\ln(1 + \text{SNR}_{g,b})]^{-1} \right\} \text{ s.}$$

Upon defining  $B_{\max} \triangleq \max_{g \in [\Lambda]} \{B_g\}$  and considering the same assumption of quasi-static fading as for the ACC scheme, the total delivery time in the coded caching scheme of [20] is

$$T_{\text{Dec}} = \frac{F}{\binom{\Lambda}{\Lambda\gamma}} \sum_{b=1}^{B_{\max}} \sum_{\substack{\Psi \subseteq [\Lambda] \\ |\Psi| = \Lambda\gamma + 1}} \max_{\substack{g \in \Psi \\ B_g \geq b}} \left\{ [\ln(1 + \text{SNR}_{g,b})]^{-1} \right\} \text{ s,}$$

where  $\text{SNR}_{g,b}$  is the SNR of the  $b$ -th ( $b \in [B_g]$ ) user of the  $g$ -th cache state.

For comparison, we also consider the performance of uncoded caching. When users request different files, the total delivery time is  $T_{\text{unCC}} = \sum_{k=1}^K \frac{(1-\gamma)F}{\ln(1+\text{SNR}_k)}$  s.

Next, we numerically evaluate the ratios  $T_{\text{unCC}}/T_{\text{ACC}}$  and  $T_{\text{unCC}}/T_{\text{Dec}}$ , averaged over channel states and cache-state allocations, to compare the delivery time boost of the proposed approach over Nakagami- $m$  fading channels. We consider that the distribution of users in cache states follows a Multinomial distribution with  $\Lambda$  equally probable outcomes (cf. [20]).

We can observe in Fig. 2.11 how the decentralized ACC scheme considerably improves the performance of coded caching, and this enhancement is more acute in the low-to-moderate SNR region. As previously pointed out, the main reason for this improvement is the amelioration of the worst-user bottleneck, where this amelioration is again the result of using shared caches as a leverage to reduce delay variability. The fact that this reduction of the worst-user bottleneck is improved as the total number of users  $K$  increases (cf. Lemma 2.4) is exemplified by comparing the  $K = 600$  case in Fig. 2.11 with the  $K = 300$  case.

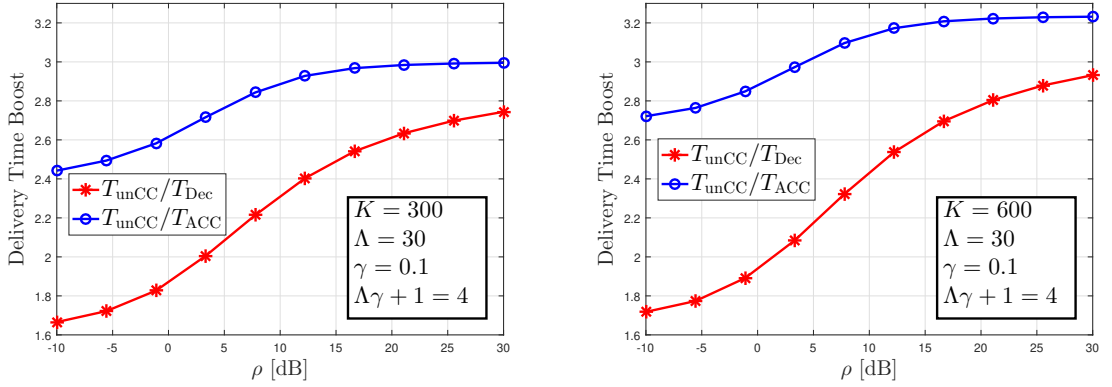


Figure 2.11: Comparison of Delivery Time versus  $\rho$  for  $m = 2$  in the decentralized-placement scenario.

## 2.4 Extension A: Delivery Time Analysis of the ACC scheme

In this section, we analyze the MN and ACC delivery time performance at finite SNR values. Motivated by the fact that the delivery time under Rayleigh fading does not have an expectation, Nakagami- $m$  fading is assumed to model the PHY channel. We derive the analytical expression of the delivery time at lower SNR, and we also consider the regime of large number of users in order to draw some valuable insights.

Let us recall that in ACC, the delivery time for serving a specific set  $\Psi$  of user groups, indeed depends on the worst (slowest) group, and not on the worst user. Let us now present our metric of interest.

**Delivery time** Consider the  $\ell$ -th group set  $\Psi_\ell$ ,  $\ell \in [(\frac{\Lambda}{\Lambda\gamma+1})]$ . From the previous description of the ACC scheme, the delivery time to  $\Psi_\ell$  is given by (cf. [51])

$$T_\ell = \frac{F}{B_w(\frac{\Lambda}{\Lambda\gamma})} \max_{g \in \Psi_\ell} \left\{ \sum_{b=1}^B \frac{1}{\log_2(1 + \text{SNR}_{g,b})} \right\} = \rho^{-1} \frac{F \ln 2}{B_w(\frac{\Lambda}{\Lambda\gamma})} \max_{g \in \Psi_\ell} \left\{ \sum_{b=1}^B \frac{1}{|h_{g,b}|^2} \right\} + o(1), \quad (2.23)$$

where  $\lim_{\rho \rightarrow 0} o(1) = 0$ , where  $h_{g,b}$  denotes the channel coefficient between the transmitter and user  $U_{g,b}$ , where  $\text{SNR}_{g,b} = \rho|h_{g,b}|^2$  denotes the instantaneous SNR at user  $U_{g,b}$ , where  $B_w$  refers to the transmission bandwidth, and where  $\rho$  denotes the normalized transmit power. Since we assume Nakagami- $m$  fading,  $|h_{g,b}|^2$  follows a Gamma distribution with shape parameter  $m$  and unitary scale parameter (i.e.,  $|h_{g,b}|^2 \sim \text{Gamma}(m, 1)$ ). Since there are  $(\frac{\Lambda}{\Lambda\gamma+1})$  transmission stages, the total delay, averaged over the channel state, is

$$\mathbb{E}_h\{T_{\text{ACC}}\} = \mathbb{E}_h \left\{ \sum_{\ell=1}^{(\frac{\Lambda}{\Lambda\gamma+1})} T_\ell \right\} = \sum_{\ell=1}^{(\frac{\Lambda}{\Lambda\gamma+1})} \mathbb{E}_h\{T_\ell\} = \rho^{-1} \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)} \mathbb{E}_h\{T_\Psi\} + o(1), \quad (2.24)$$

where  $T_\Psi \triangleq \max_{g \in \Psi} \left\{ \sum_{b=1}^B \frac{1}{|h_{g,b}|^2} \right\}$  focuses on a specific user-group set  $\Psi$  during the ACC delivery phase. In this section, we focus on the delivery time at low SNR, and thus we define the following approximation

$$\mathring{T}_{\text{ACC}} \triangleq \rho^{-1} \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)} \mathbb{E}_h\{T_\Psi\} \quad (2.25)$$

by omitting the term  $o(1)$  in (2.24).

To show the extent to which the delivery performance is enhanced at reasonable values of the SNR, we define the *effective coded caching gain*, now in terms of delivery time<sup>6</sup>. First, we will consider the performance of simple Time-Division Multiplexing (TDM) transmission without coded caching, in the sense that the local memory at the users is only useful to provide *local caching gains* that stem from the fact that  $\gamma F$  bits of the requested file are already stored at the user. This scheme allows us to characterize the net gain generated by coded caching. Henceforth, we will refer to this scheme as the *uncoded TDM scheme*.

**Definition 1** (Effective coded caching gain). *The effective coded caching gain of a particular coded caching scheme is the ratio of the average delivery time of the uncoded TDM scheme over the average delivery time of the said particular scheme (here, of the MN or ACC schemes), where the average is with respect to the quasi-static channel fading state.*

### 2.4.1 Delivery Time of the MN Scheme at Low SNR

In this subsection, we analyze the average delivery time of the MN scheme in order to present some insights about its performance at practical SNR values and to provide a benchmark for the ACC scheme.

<sup>6</sup>The effective gain of a coded caching scheme is actually a generic metric to reflect the real performance boost over the uncoded TDM scheme.

In order to serve the  $K$  users by means of the MN scheme in the scenario where the subpacketization constraint induces a maximum number of cache states  $\Lambda < K$ , we need to consider the MN transmission only over  $\Lambda$  users (one from each group), and then repeat the same process  $B = K/\Lambda$  times, in order to serve all the  $K$  users (cf. [51], [18]). In this setting, the delivery time required by the MN scheme to serve  $\Lambda$  users, matches the ACC delivery time when  $B = 1$  [51] and thus this MN average delivery time at low SNR follows from (2.25) as

$$\mathring{T}_{\text{MN}} \triangleq \rho^{-1} B \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)} \mathbb{E}\{T_\Psi\}, \quad (2.26)$$

where  $T_\Psi \triangleq \max_{g \in \Psi} \left\{ \frac{1}{|h_g|^2} \right\}$ , and  $h_g$  denotes the channel coefficient between the transmitter and user  $U_g$  (we omit the subscript  $b$  because here  $B = 1$ ). In (2.26), the introduction of the factor  $B$  reflects the fact that we repeat the MN scheme  $B = K/\Lambda$  times in order to serve all  $K$  users. In the next lemma, we provide the analytical expression of  $\mathring{T}_{\text{MN}}$ . We recall that  $G = 1 + \Lambda\gamma$ .

**Lemma 2.7.** *The approximate average delivery time of the MN scheme at low SNR over quasi-static Nakagami- $m$  fading is given by*

$$\mathring{T}_{\text{MN}} = \frac{K(1-\gamma)F \ln 2}{\rho B_w G} \int_0^\infty 1 - \left( \frac{\Gamma(m, 1/x)}{\Gamma(m)} \right)^G dx, \quad (2.27)$$

where  $\Gamma(\cdot)$  and  $\Gamma(\cdot, \cdot)$  denote the Gamma function and the upper incomplete Gamma function [76], respectively.

*Proof.* Under the assumption of Nakagami- $m$  fading, it follows that  $|h_g|^2 \sim \text{Gamma}(m, 1)$ . Then, we have that  $\frac{1}{|h_g|^2}$  follows an inverse Gamma distribution with shape parameter  $m$  and scale parameter equal to 1. Therefore, the CDF of  $\frac{1}{|h_g|^2}$  is (cf. [82])

$$F_{1/|h_g|^2}(x) = \frac{1}{\Gamma(m)} \Gamma(m, 1/x), \quad x \geq 0. \quad (2.28)$$

Hence, the CDF of  $T_\Psi \triangleq \max_{g \in \Psi} \left\{ \frac{1}{|h_g|^2} \right\}$  in (2.26) writes as

$$F_{T_\Psi}(x) = \left( F_{1/|h_g|^2}(x) \right)^G = \left( \frac{1}{\Gamma(m)} \Gamma(m, 1/x) \right)^G. \quad (2.29)$$

Since  $T_\Psi$  has a non-negative support,  $\mathbb{E}_h\{T_\Psi\}$  can be obtained by integrating  $1 - F_{T_\Psi}(x)$  from 0 to infinity. Substituting  $\mathbb{E}_h\{T_\Psi\}$  by this integral in (2.26) yields (2.27).  $\square$

The integral in Lemma 2.7 does not have a closed-form solution. Yet, we can obtain a closed-form expression for  $\mathring{T}_{\text{MN}}$  when  $m$  is a positive integer bigger than 1, as will be stated further down, in the subsequent corollary. Before presenting this result though, let us introduce some useful notations; Let  $\mathbf{k} \triangleq [k_1, k_2, \dots, k_m]$  be a non-negative integer vector and  $\binom{G-1}{\mathbf{k}} \triangleq \frac{(G-1)!}{k_1! k_2! \dots k_m!}$  be the multinomial coefficient, and let us use  $\|\mathbf{k}\|_1$  to denote the norm-1 operator of any vector  $\mathbf{k}$ . We can present now the following result.

**Corollary 2.2.** *For any positive integer  $m \geq 2$ , the approximate delivery time of the MN scheme becomes*

$$\begin{aligned} \mathring{T}_{\text{MN}} = & \rho^{-1} \frac{K(1-\gamma)F \ln 2}{B_w(m-1)!} \sum_{\|\mathbf{k}\|_1=G-1} \binom{G-1}{\mathbf{k}} \\ & \times \frac{G^{1-m-\sum_{j=1}^m(j-1)k_j}}{\prod_{j=1}^m ((j-1)!)^{k_j}} \left( m-2 + \sum_{j=1}^m (j-1)k_j \right)!. \end{aligned} \quad (2.30)$$

*Proof.* We first obtain the PDF of  $T_\Psi$  by differentiating the CDF of  $T_\Psi$  given in (2.29).

$$f_{T_\Psi}(x) = \frac{Ge^{-1/x}}{(\Gamma(m))^{Gx^{m+1}}} (\Gamma(m, 1/x))^{G-1}. \quad (2.31)$$

By applying [76, Eq. (8.352.2)], we can rewrite  $f_{T_\Psi}(x)$  for any positive integer  $m$  as

$$\begin{aligned} f_{T_\Psi}(x) &= \frac{Ge^{-1/x}}{(\Gamma(m))^{Gx^{m+1}}} \left( \Gamma(m) e^{-1/x} \sum_{j=0}^{m-1} \frac{x^{-j}}{j!} \right)^{G-1} \\ &= \frac{Ge^{-G/x}}{\Gamma(m)} \sum_{\|\mathbf{k}\|_1=G-1} \binom{G-1}{\mathbf{k}} \frac{x^{-m-1-\sum_{j=1}^m(j-1)k_j}}{\prod_{j=1}^m ((j-1)!)^{k_j}}. \end{aligned} \quad (2.32)$$

Since  $\mathbb{E}_h\{T_\Psi\} = \int_0^\infty x f_{T_\Psi}(x) dx$ , we obtain  $\mathbb{E}_h\{T_\Psi\}$  by considering the expression of  $f_{T_\Psi}(x)$  in (2.32). Finally, substituting  $\mathbb{E}_h\{T_\Psi\}$  in (2.26) yields (2.30).  $\square$

Next, we also consider the performance of simple TDM without coded caching.

**Corollary 2.3.** *The low SNR approximate delivery time with uncoded TDM is*

$$\mathring{T}_{\text{TDM}} = \rho^{-1} \frac{K(1-\gamma)F \ln 2}{B_w(m-1)} \quad \forall m > 1. \quad (2.33)$$

*Proof.* This result follows directly from Lemma 2.7 after fixing the number of simultaneously served users to correspond to  $G = 1$ . Then, it follows that  $\mathbb{E}_h\{T_\Psi\} = \mathbb{E}_h\{\frac{1}{|h_i|^2}\}$  for  $i \in K$ , which yields (2.33) after substituting  $\mathbb{E}_h\{T_\Psi\} = 1/(m-1)$  in (2.26). The factor  $1-\gamma$  in (2.33) is naturally due to the fact that every user has stored  $\gamma F$  bits of each file in the library  $\mathcal{F}$ , and the server only needs to deliver the remaining  $(1-\gamma)F$  bits of each requested file.  $\square$

**Remark 2.5.** *The expectation of the delivery time does not exist for  $m \leq 1$ , where  $m = 1$  corresponds to Rayleigh fading. This can also be inferred from Corollaries 2.2 and 2.3. Accordingly, hereon we only consider  $m > 1$  unless otherwise stated.*

From the previous results, we can analyze the effective coded caching gain of the MN scheme in the considered scenario, which leads to the following lemma.

**Lemma 2.8.** *The effective coded caching gain of the MN scheme in the low-SNR limit of  $\rho \rightarrow 0$ , takes the form*

$$\mathcal{G}_{\text{MN}} = \frac{G}{(m-1)} \left( \int_0^\infty 1 - \left( \frac{\Gamma(m, 1/x)}{\Gamma(m)} \right)^G dx \right)^{-1}.$$

*Proof.* This result is directly obtained by applying Definition 1 of the gain as the ratio of  $\mathring{T}_{\text{TDM}}$  in (2.33) over  $\mathring{T}_{\text{MN}}$  in (2.27).  $\square$

When  $m$  is a positive integer, we can treat  $|h_g|^2$  (which is Gamma distributed) as the summation over  $m$  i.i.d. exponential random variables with unit-mean. As  $m \rightarrow \infty$ , the Strong Law of Large Numbers implies that  $\frac{1}{m}|h_g|^2 \xrightarrow{a.s.} 1$ , where  $\xrightarrow{a.s.}$  stands for almost sure convergence. Substituting  $\frac{1}{m}|h_g|^2 \xrightarrow{a.s.} 1$  into (2.26) yields

$$\mathring{T}_{\text{MN}} \xrightarrow{a.s.} \rho^{-1} B \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)m}, \text{ as } m \rightarrow \infty, \quad (2.34)$$

and thus the effective gain at low SNR satisfies that

$$\mathcal{G}_{\text{MN}} \xrightarrow{a.s.} G = \Lambda\gamma + 1 \text{ as } m \rightarrow \infty, \quad (2.35)$$

which shows how the worst-user effect, even in the case of applying the MN scheme, can be mitigated as  $m$  increases.

## 2.4.2 Delivery Time of the ACC Scheme at Low SNR

Next, we analyze the delivery time of the ACC scheme in the low-SNR regime, again considering quasi-static Nakagami- $m$  fading. First, we present the exact expression of  $\mathring{T}_{\text{ACC}}$ , which is obtained through the Characteristic Function (CF) method. To simplify the derived complex expressions, we will then consider the large  $K$  regime to derive approximations that are robust even when  $K$  is modestly values.

### Low SNR Characterization

Before presenting our next result, let us introduce some useful notation. In the following,  $j \triangleq \sqrt{-1}$  denotes the imaginary unit,  $\text{Im}\{\cdot\}$  refers to the imaginary part of a complex number, and  $K_n(\cdot)$  denotes the modified Bessel function of the second kind [76].

We present now the exact expression of the approximated delivery time  $\mathring{T}_{\text{ACC}}$ . We recall that  $\mathring{T}_{\text{ACC}}$  is the delivery time obtained after applying the very basic low-SNR capacity approximation corresponding to  $\ln(1+x) = x + o(1) \approx x$  when  $x \rightarrow 0$ .

**Lemma 2.9.** *Under quasi-static Nakagami- $m$  fading (and  $m > 1$ ), the low SNR approximated ACC delivery time takes the form*

$$\begin{aligned} \mathring{T}_{\text{ACC}} = & \rho^{-1} \frac{\Lambda(1-\gamma)F \ln 2}{GB_w} \\ & \times \int_0^\infty 1 - \left( \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \text{Im} \left\{ \frac{e^{-jtz}}{t} \left( \frac{2(-jt)^{\frac{m}{2}}}{\Gamma(m)} K_m(\sqrt{-4jt}) \right)^B \right\} dt \right)^G dz. \end{aligned} \quad (2.36)$$

*Proof.* See Appendix A.5.  $\square$

To further simplify the above, we now consider the large  $K$  regime where  $\Lambda$  remains fixed but where  $B \rightarrow \infty$ .

### Large $K$ Performance in the Low SNR Regime

In the following, we consider a large number  $B$  of users per group (i.e., that  $B \rightarrow \infty$ ) while we consider a fixed number  $\Lambda = \frac{K}{B}$  of cache states. This in turn implies a nominal (high-SNR) coded-caching gain (under of course the subpacketization constraint that forced our  $\Lambda$ ) to be fixed at  $G = 1 + \Lambda\gamma$ .

Let us first introduce some useful notations. We use  $\varrho_t$  and  $\sigma_t^2$  to denote the expectation and variance of  $1/\text{SNR}_{g,b}$ , respectively. We also use  $\mathcal{H}_G$  to denote the expectation of the maximum of  $G$  i.i.d. Gaussian random variables with zero-mean and unit-variance.

With these notations in place, let us present, in the following lemma, the approximation of  $\hat{T}_{\text{ACC}}$  in the regime of large  $B$ .

**Lemma 2.10.** *For a sufficiently large number of users, the expectation of the delivery time over quasi-static Nakagami- $m$  fading channels with  $m > 2$ , can be tightly approximated in the low SNR regime, as*

$$\hat{T}_{\text{ACC}} \approx \frac{\Lambda(1-\gamma)F \ln 2}{B_w G} \left( B\varrho_t + \mathcal{H}_G \sqrt{B\sigma_t^2} \right), \quad (2.37)$$

where  $\varrho_t$  and  $\sigma_t^2$  are given respectively by

$$\varrho_t = \frac{1}{\rho(m-1)}, \quad \sigma_t^2 = \frac{1}{\rho^2(m-1)^2(m-2)}. \quad (2.38)$$

*Proof.* We first note that the condition  $m > 2$  allows us to conclude that  $1/\text{SNR}_{g,b}$  has finite variance, which in turn will allow us to employ the Central Limit Theorem (CLT).

As  $1/\text{SNR}_{g,b} = \rho^{-1}|h_{g,b}|^{-2}$  is drawn from an inverse Gamma distribution of mean and variance given respectively by  $\varrho_t$  and  $\sigma_t^2$  in (2.38), we can apply the CLT to obtain that

$$\sum_{b=1}^B \frac{1}{\text{SNR}_{g,b}} \xrightarrow{d.} \mathcal{N}(B\varrho_t, B\sigma_t^2), \quad (2.39)$$

where  $d.$  denotes the convergence in distribution and  $\mathcal{N}$  denotes the normal distribution. Therefore,  $T_{\Psi}/\rho = \max_{g \in \Psi} \{\sum_{b=1}^B \rho^{-1}|h_{g,b}|^{-2}\}$  converges in distribution to the maximum of  $G$  i.i.d. normal random variables with mean  $B\varrho_t$  and variance  $B\sigma_t^2$ . Upon defining  $\{X_g\}_{g \in \Psi}$  as a set of  $G$  i.i.d. variables distributed as  $\mathcal{N}(0, 1)$ , we can write that  $T_{\Psi}/\rho \xrightarrow{d.} B\varrho_t + \sqrt{B\sigma_t^2} \max_{g \in \Psi} X_g$ . Since we have defined  $\mathcal{H}_G$  as the expectation of  $\max_{g \in \Psi} X_g$ , we obtain (2.37) from the definition of  $\hat{T}_{\text{ACC}}$  in (2.25).  $\square$

Regarding  $\mathcal{H}_G$  in (2.37), we refer to Lemma 2.6 for more information. Summing up as less as 5 terms in the GHQ method provides a very accurate approximation of  $\mathcal{H}_G$  [51].

We also have the following.

**Lemma 2.11.** *In the large  $K$  regime and for  $m > 2$ , we can approximate the effective coded caching gain of the ACC scheme at low SNR by*

$$\mathcal{G}_{\text{ACC}} \approx \frac{G}{1 + \mathcal{H}_G / \sqrt{B(m-2)}}. \quad (2.40)$$

*Proof.* The result follows by considering Definition 1 and the expressions of  $\mathring{T}_{\text{TDM}}$  in (2.33) and  $\mathring{T}_{\text{ACC}}$  in (2.37).  $\square$

It is direct to see that, when  $B \rightarrow \infty$  or  $m \rightarrow \infty$ , then  $\mathcal{G}_{\text{ACC}}$  converges to  $G$ , which is the optimal gain at high SNR. This means that we recover the nominal gains of coded caching even at low SNR, provided that there are enough users or — as one would expect — a big enough  $m$  (corresponding to, for example, having many receiving antennas).

### Extension — Large Number of Users at any SNR

In the above, we employed the basic low-SNR approximation  $\ln(1 + \text{SNR}) \approx \text{SNR}$ . To remove this approximation, let us employ analysis in the large  $K$  regime. Towards this, we first note that Eq. (2.37) in Lemma 2.10 holds but now  $\varrho_t$  and  $\sigma_t^2$  will respectively denote the mean and variance of  $1/\ln(1 + \text{SNR}_{g,b})$ , instead of  $1/\text{SNR}_{g,b}$ . From the PDF of  $\text{SNR}_{g,b}$  corresponding to Nakagami- $m$  fading, we obtain the following integral forms for  $\varrho_t$  and  $\sigma_t^2$ ,

$$\varrho_t = \frac{1}{\rho^m \Gamma(m)} \int_0^\infty \frac{x^{m-1} e^{-\frac{x}{\rho}}}{\ln(1+x)} dx, \quad (2.41)$$

$$\sigma_t^2 = \frac{1}{\rho^m \Gamma(m)} \int_0^\infty \left( \frac{1}{\ln(1+x)} - \varrho_t \right)^2 x^{m-1} e^{-\frac{x}{\rho}} dx. \quad (2.42)$$

As there are no closed-form solutions for these integrals, we adopt the Gauss-Laguerre quadrature (GLQ) [83], from which the integrals are respectively approximated as

$$\varrho_t \approx \frac{1}{\Gamma(m)} \sum_{v=1}^V \chi_v \frac{y_v^{m-1}}{\ln(1 + \rho y_v)}, \quad (2.43)$$

$$\sigma_t^2 \approx \frac{1}{\Gamma(m)} \sum_{v=1}^V \chi_v y_v^{m-1} \left( \frac{1}{\ln(1 + \rho y_v)} - \varrho_t \right)^2, \quad (2.44)$$

where  $V$ ,  $y_v$  and  $\chi_v$  are the summation terms, sample points and weights of the GLQ, respectively. As for the GHQ, the accuracy of the GLQ is typically very good after summing up a few terms.

### 2.4.3 Numerical Results

In this subsection, we demonstrate the accuracy of the derived expressions through Monte-Carlo simulations. Hereinafter, we assume that the file size is  $F = 8 \times 10^9$  bits (i.e., 1 Gigabyte), and that the bandwidth for each user is 20 MHz (as in 4G standard). In order to implement the Monte-Carlo simulations,  $10^6$  channel states are generated and averaged over Nakagami- $m$  fading channels.

In Fig. 2.12, we plot the delivery time of the MN and ACC schemes versus  $\rho$  for different values of  $m$ . We can observe that the delivery time decreases as  $m$  increases. This happens because a bigger  $m$  allow for enjoying a richer multi-path environment, thereby enhancing the spatial diversity. Indeed, as it is known, if an  $m$ -antenna receiver

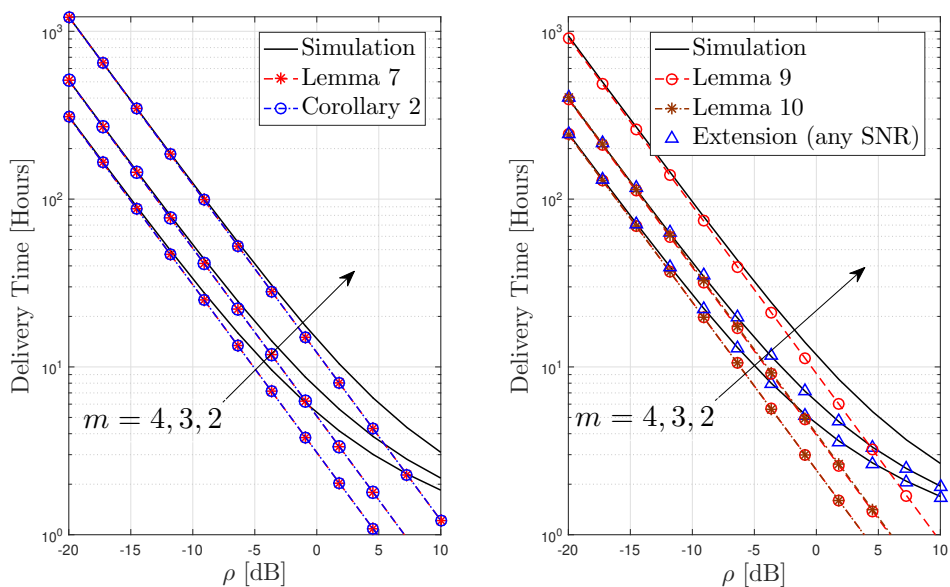


Figure 2.12: Average delivery time of the MN (left) and ACC (right) schemes versus  $\rho$  for  $K = 300$ ,  $B = 5$ ,  $\gamma = 5\%$  and  $V = 7$  in the GHQ and GLQ.

applies MRC over Rayleigh fading, the magnitude of the equivalent channel coefficient after MRC follows a Nakagami- $m$  fading distribution.

Fig. 2.12 also shows the high accuracy of the derived approximations at low SNR. We observe that the SNR regime in which we obtain high accuracy for the low-SNR approximations is reduced as  $m$  increases. This is due to the fact that the bigger the  $m$ , the higher is the effective SNR received at the user for the same  $\rho$ . However, it is worth noting that the large  $K$  approximation (without the additional low-SNR approximation corresponding to (2.43) and (2.44)) tightly approximates the delivery time *even for very small values of  $B$*  (as low as, for example,  $B = 5$ ), for  $m > 2$  and for any SNR.

We plot the effective coded caching gains of the ACC and MN schemes in Fig. 2.13 with the same system settings of Fig. 2.12. Besides validating the correctness of Lemmas 2.8 and 2.11, Fig. 2.13 also shows that the gains of both schemes reach a plateau at very low SNR, but also that the infimum for the ACC scheme (which becomes bigger as  $B$  grows) is larger than the one for the MN scheme. Moreover, the effective gains of both schemes increase as  $m$  increases. Indeed, as stated before, both gains  $G_{\text{ACC}}$  and  $G_{\text{MN}}$  converge to the nominal  $G$  as  $m$  increases.

Fig. 2.14 represents the delivery time improvement of the ACC scheme over the MN scheme (plotted against  $\rho$ ) for different values of  $B$ . As expected, the delivery times of both schemes converge as  $\rho$  increases, while the boost effect of the ACC scheme becomes significant in the low SNR region. It is also obvious that the larger the  $B$ , the better the performance that the ACC scheme can achieve.

Fig. 2.15 illustrates the effective coded caching gains of the MN scheme and of the ACC scheme when  $B = 10$  and for different values of  $m$  and  $G$  as  $\rho \rightarrow 0$ . Naturally the effective gain increases with  $m$ , since a larger  $m$  implies more spatial diversity that

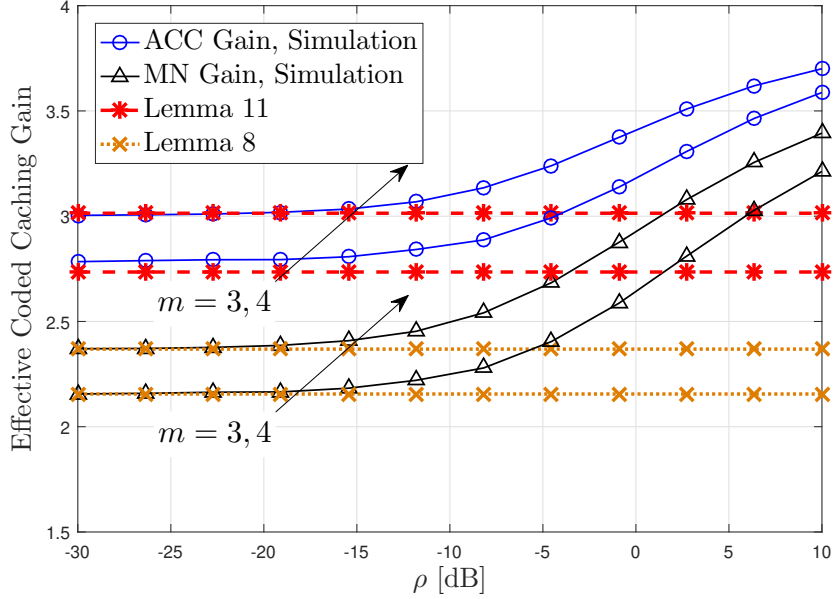


Figure 2.13: Effective coded caching gains of the ACC and MN schemes for  $K = 300$ ,  $B = 5$ ,  $\gamma = 5\%$ , and  $m = 3, 4$ .

alleviates the worst-user effect.

To further show the impact of  $B$  on the effective gain, we plot in Fig. 2.16 the effective gains of the MN (left) and ACC (right) schemes versus  $m$  and  $B$  for  $G = 5$ , and do so in the low-SNR regime. Naturally the effective gain of the ACC scheme increases as  $B$  increases, and the increasing trend becomes more obvious in the small  $m$  region. Indeed, the symmetry of the plot implies a certain equivalence of  $B$  and  $m$ , in the sense that having  $B$  single-antenna users per cache group achieves approximately the same performance as a setting with  $K$   $m$ -antenna receivers that apply MRC.

## 2.5 Extension B: ACC Performance in Ergodic-Fading Scenario With Different Pathloss

In this section, we investigate the use of coded caching in the single-cell downlink scenario where the receiving users are randomly located inside the cell. We first show that, as a result of having users that experience very different path-loss, the real gain of the original (MN) coded caching scheme is severely reduced. We then prove that the use of shared caches — which, we stress, is a compulsory feature brought about by the subpacketization constraint in nearly every practical setting — introduces a powerful spatial-averaging effect that allows us to recover most of the nominal (subpacketization-constrained) gains that coded caching would have yielded in the error-free identical-link setting. For the ergodic-fading scenario with different pathloss, we derive tight approximations of the

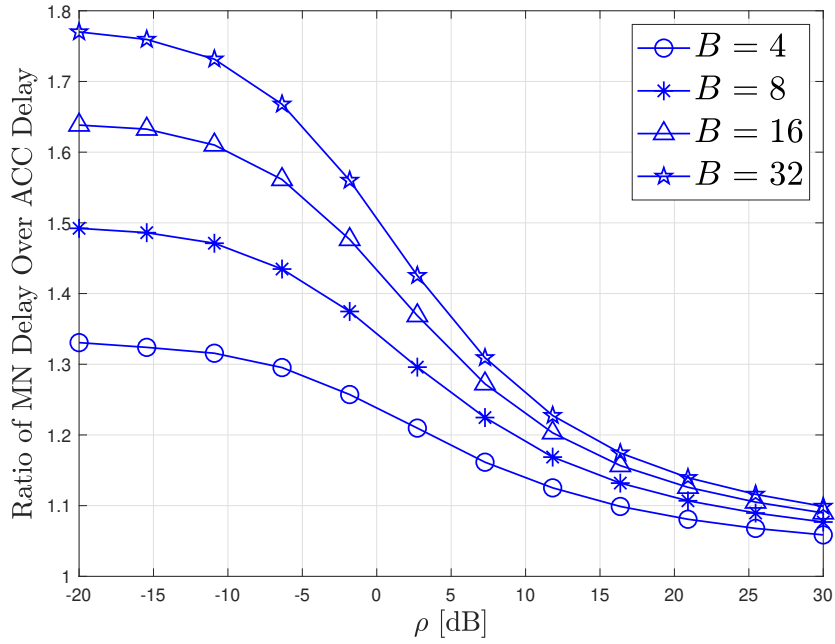


Figure 2.14: Delay Ratio of  $\overset{\circ}{T}_{\text{MN}}$  over  $\overset{\circ}{T}_{\text{ACC}}$  versus  $\rho$  for  $m = 3$ ,  $\gamma = \frac{1}{12}$ , and  $G = 6$ .

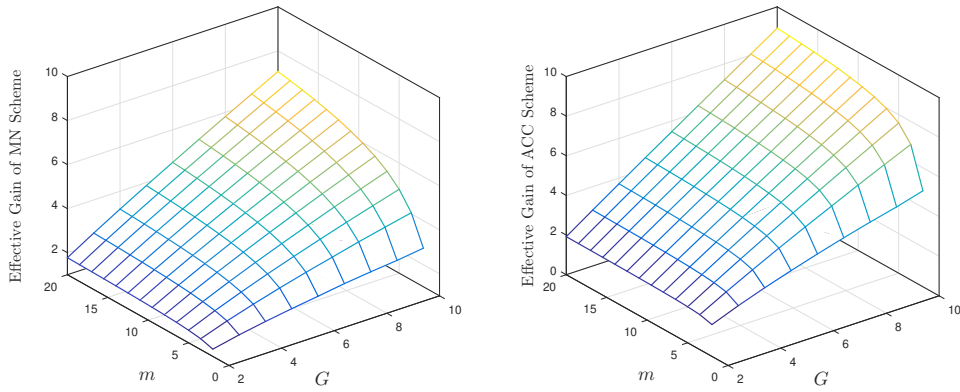


Figure 2.15: Effective coded caching gains of MN (left) and ACC (right) schemes for  $B = 10$  and  $\gamma = 10\%$  as  $\rho \rightarrow 0$ .

average (over the users) rate and of the coded caching gain by means of a basic high-SNR approximation on the point-to-point capacity. These derived expressions prove very accurate even for low SNR. We also provide a result based on the regime of large number of users which is nonetheless also valid for settings with few users.

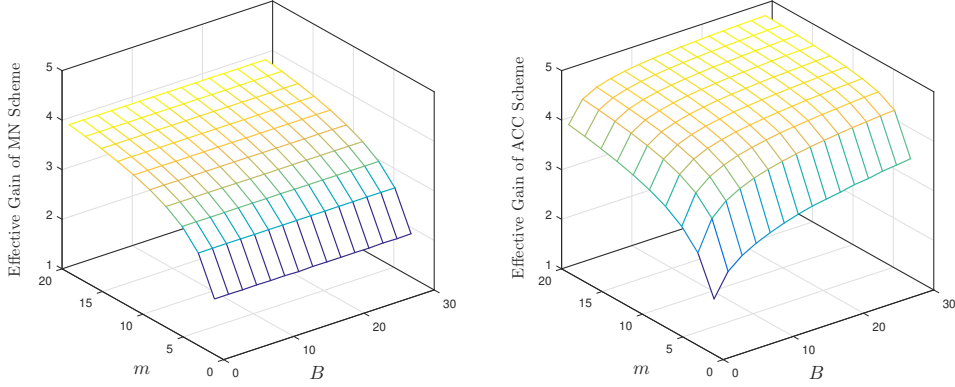


Figure 2.16: Effective coded caching gains of MN (left) and ACC (right) schemes for  $G = 5$  and  $\gamma = 5\%$  as  $\rho \rightarrow 0$ .

### 2.5.1 System Model

We analyze a single-cell setting in which a single-antenna transmitter has, as is common, full access to a library  $\mathcal{F} = \{W_n\}_{n=1}^N$  of  $N$  files  $W_1, \dots, W_N$ , each of size  $F$  bits. The transmitter serves  $K$  cache-aided single-antenna users who are uniformly distributed throughout a ring with inner radius  $D_1$  and outer radius  $D_2$  surrounding the transmitter. Each user benefits from a local cache that can store a fraction  $\gamma \in [0, 1]$  of the library content. We consider the instantaneous SNR at an arbitrary user  $U_i$  to be given by<sup>7</sup>

$$\text{SNR}_i = \frac{P_t}{N_0 B_w N_f \beta_f L_c} |h_i|^2 r_i^{-\eta_0}, \quad (2.45)$$

where  $P_t$ ,  $N_0$ ,  $B_w$ , and  $\eta_0$  are the transmit power, the noise density, the bandwidth, and the path-loss exponent, respectively. In the above,  $h_i$  corresponds to the fast-fading channel coefficient, drawn from a zero-mean unit-variance complex Gaussian distribution, and  $r_i$  is the distance in meters from user  $i$  to the transmitter. In (2.45),  $\beta_f$  is a path-loss component that depends only on the carrier frequency ( $f_{\text{GHz}}$ ), while  $N_f$  denotes the noise figure that measures the practical imperfections of the receiver, and  $L_c$  represents a constant loss term accounting for slow fading and other practical factors (rain, foliage, etc.). To facilitate notation, we define  $\rho \triangleq \frac{P_t}{N_0 B_w N_f \beta_f L_c}$  to incorporate all the terms in  $\text{SNR}_i$  other than the distance and the effect of fast fading. As the users are uniformly distributed within a ring, the PDF of  $r_i$  is (cf. [85])

$$f_{r_i}(r) = \frac{2r}{D_2^2 - D_1^2}, \quad D_1 \leq r \leq D_2. \quad (2.46)$$

Under basic Gaussian signalling assumptions, we consider that single-user transmission through a channel with instantaneous SNR denoted by  $X$  can attain an ergodic rate of

<sup>7</sup>The channel model follows the propagation models defined by the Third Generation Partnership Project (3GPP) [84].

$\mathbb{E}_h\{\ln(1+X)\}$  when averaged over the fast-fading channel coefficients. We will consider the average of such ergodic rate, averaged over the users, or, more precisely, over the random user locations. In this context, we will use  $\mathbb{E}_{h|r}$  to denote the expectation over fast-fading channels for a given location realization  $r$  which remains fixed for the entire transmission, and we will use  $\mathbb{E}_r$  to denote averaging over the user locations  $r$ . Hence, our *average rate* metric corresponds to averaging as  $\mathbb{E}_{h,r}\{\cdot\} \triangleq \mathbb{E}_r\{\mathbb{E}_{h|r}\{\cdot\}\}$ , i.e., averaging the ergodic capacities across user locations.

We consider the basic linear (with respect to  $\ln \rho$ ) approximation  $\ln(1+x) \approx \ln x$  on the capacity, obtained from the fact that  $\ln(1+x) = \ln x + o(1)$ , where as before  $\lim_{x \rightarrow \infty} o(1) = 0$ .

**Definition 2** (Affine approximation of average rate). *The average-rate affine approximation  $\tilde{R}$  is defined as the maximum long-term average achievable rate, averaged first over the fast-fading coefficients and then over user locations, after applying the linear<sup>8</sup> approximation (with respect to  $\ln x$ )  $\ln(1+x) \approx \ln x$ .*

In the case of the ACC scheme (cf. Section 2.2), for a given transmission stage, the average rate achieved by group  $g \in \Psi$  is given by  $\frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b})$ , since the users are served sequentially within a group. Then, the transmission stage ends when all user groups have decoded their intended subfiles. Thus, the (affine-approximated) achievable rate in a given transmission stage takes the form (cf. (2.8))

$$R_\Psi \triangleq G \min_{g \in \Psi} \left\{ \frac{1}{B} \sum_{b=1}^B \ln(\text{SNR}_{g,b}) \right\} \text{ nats/s/Hz}, \quad (2.47)$$

where the factor  $G = \Lambda\gamma + 1$  is due to the fact that the transmitter simultaneously serves a set of  $G$  users.

## 2.5.2 MN Scheme in the Single-Cell Scenario

We first derive the affine-approximated average rate of TDM. We then provide a tight lower bound on the affine-approximated MN rate and a corresponding bound on the approximated MN effective gain.<sup>9</sup> We quickly recall that the affine-approximated rate is simply the sought average rate after substituting  $\ln(1 + \text{SNR}_k)$  by  $\ln \text{SNR}_k$ .

**Lemma 2.12.** *The affine-approximated rate for TDM is*

$$\tilde{R}^{\text{TDM}} = \ln \rho - C + \frac{\eta_0}{2} - \eta_0 \frac{D_2^2 \ln D_2 - D_1^2 \ln D_1}{D_2^2 - D_1^2} \quad (2.48)$$

where  $C = 0.5772\dots$  is the Euler-Mascheroni constant.

<sup>8</sup>As we will see below, this linear approximation will provide affine expressions of the average rate, such that it can be written as  $A \ln \rho + B$ , where  $A$  and  $B$  are independent of  $\rho$ . Such affine approximations have been shown to accurately represent the rate in several scenarios [86, 87].

<sup>9</sup>In a small deviation of notation, and only for this section, we will use  $\text{SNR}_k, \forall k \in [K]$ , to denote the SNR of User  $k$  in both TDM and MN schemes.

*Proof.* The affine-approximated average rate can be written as

$$\tilde{R}^{\text{TDM}} = \mathbb{E}_r \left\{ \mathbb{E}_{h|r} \left\{ \ln \left( \rho r_k^{-\eta_0} |h_k|^2 \right) \right\} \right\}, \quad (2.49)$$

where  $\rho r_k^{-\eta_0}$  is naturally constant with respect to the inner expectation, and where  $|h_k|^2$  follows a unit-mean exponential distribution. It then follows that

$$\tilde{R}^{\text{TDM}} \stackrel{(a)}{=} \mathbb{E}_r \left\{ \ln(\rho r_k^{-\eta_0}) - C \right\} = \ln \rho - C - \eta_0 \mathbb{E}_r \{ \ln r_k \}, \quad (2.50)$$

where (a) holds because, for any  $X$  drawn from a unit-mean exponential distribution,  $-\ln(X)$  follows a standard Gumbel distribution, whose mean equals  $C$  [86]. We then obtain (2.48) from (2.50) by deriving  $\mathbb{E}_r \{ \ln r_k \}$  from the PDF of  $r_k$  in (2.46).  $\square$

We present a lower bound on the affine approximation of the MN scheme in the following lemma, which is derived by means of Jensen's inequality. As we will see in Section 2.5.4, this simple bound offers a tight approximation of the actual rate.

**Lemma 2.13.** *The affine-approximated average rate of the MN scheme is lower bounded by*

$$\tilde{R}^{\text{MN}} \geq G \ln \rho - G \left[ \ln \left( \frac{2G(D_2^{\eta_0+2} - D_1^{\eta_0+2})}{(\eta_0 + 2)(D_2^2 - D_1^2)} \right) + C \right]. \quad (2.51)$$

*Proof.* Since the expression of the average rate of the MN scheme coincides with the one of the ACC scheme for  $B = 1$  (i.e., with dedicated caches) [52], it follows from (2.47) that

$$\tilde{R}^{\text{MN}} = G \ln \rho + G \mathbb{E}_{h,r} \left\{ \ln \left( \min_{k \in \Psi} \left\{ |h_k|^2 r_k^{-\eta_0} \right\} \right) \right\}, \quad (2.52)$$

where we have applied the approximation  $\ln(1+x) \approx \ln x$ . Since  $|h_k|^2 r_k^{-\eta_0}$  follows an exponential distribution with mean  $r_k^{-\eta_0}$  for a fixed location realization,  $\min_{k \in \Psi} \left\{ |h_k|^2 r_k^{-\eta_0} \right\}$  follows an exponential distribution with mean  $1 / \left( \sum_{k \in \Psi} r_k^{\eta_0} \right)$ . Hence, applying the same step as for (a) in (2.50), the expectation of  $\ln \left( \min_{k \in \Psi} \left\{ |h_k|^2 r_k^{-\eta_0} \right\} \right)$  conditioned on the user locations is

$$\mathbb{E}_{h|r} \left\{ \ln \left( \min_{k \in \Psi} \left\{ |h_k|^2 r_k^{-\eta_0} \right\} \right) \right\} = - \ln \left( \sum_{k \in \Psi} r_k^{\eta_0} \right) - C. \quad (2.53)$$

Since  $-\ln x$  is a convex function over  $(0, \infty)$ , we can apply Jensen's inequality to obtain

$$\mathbb{E}_{h,r} \left\{ \ln \left( \min_{k \in \Psi} \left\{ |h_k|^2 r_k^{-\eta_0} \right\} \right) \right\} = \mathbb{E}_r \left\{ - \ln \left( \sum_{k \in \Psi} r_k^{\eta_0} \right) \right\} - C \geq - \ln \left( G \mathbb{E}_r \left\{ r_k^{\eta_0} \right\} \right) - C. \quad (2.54)$$

Finally, we obtain (2.51) by using the PDF of  $r_k$  to derive the closed-form expression for  $\mathbb{E}_r \{ r_k^{\eta_0} \}$  in (2.54).  $\square$

Let us now approximate the effective coded caching gain of the MN scheme (which was introduced in Chapter 2) by the ratio  $\tilde{G}_{\text{MN}}$  of the corresponding average-rate affine approximations. The following is direct from Lemmas 2.12 and 2.13.

**Corollary 2.4.** *The approximate effective coded caching gain for the MN scheme is bounded as*

$$\tilde{G}_{\text{MN}} \geq \frac{G \ln \rho - G \left[ \ln \left( \frac{2G(D_2^{\eta_0+2} - D_1^{\eta_0+2})}{(\eta_0+2)(D_2^2 - D_1^2)} \right) + C \right]}{\ln \rho - C + \frac{\eta_0}{2} - \eta_0 \frac{D_2^2 \ln D_2 - D_1^2 \ln D_1}{D_2^2 - D_1^2}}. \quad (2.55)$$

In the above, the approximation steps only include the use of Jensen's inequality and the use of the basic approximation on capacity. Consequently, as we will show, the above results offer a very accurate evaluation of the actual performance.

### 2.5.3 ACC Scheme in the Single-Cell Scenario

In this subsection, we first derive the analytical expression for the affine-approximated average rate of the ACC scheme. Then, we provide a large- $B$  approximation that precisely characterizes the actual performance even if  $B$  is as low as 2.

#### Affine-Approximated Average Rate

Before presenting the lemma describing the performance of the ACC scheme, let us introduce the notation  $j \triangleq \sqrt{-1}$  and let  $\text{Im}\{\cdot\}$  denote the imaginary part of a complex number.

**Lemma 2.14.** *The affine-approximated average rate of the ACC scheme is given by*

$$\begin{aligned} \tilde{R}^{\text{ACC}} = & G \ln \rho \\ & - \frac{G}{B} \int_0^\infty 1 - \left( \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \text{Im} \left\{ \frac{\exp(jtx)}{t} \left[ \Gamma(1+jt) \frac{j^2(D_2^{2-j\eta_0 t} - D_1^{2-j\eta_0 t})}{(\eta_0 t + j^2)(D_2^2 - D_1^2)} \right]^B \right\} dt \right)^G dx \end{aligned} \quad (2.56)$$

where  $\Gamma(\cdot)$  denotes the Gamma function [76] and  $G = \Lambda\gamma + 1$ .

*Proof.* See Appendix A.6. □

**Remark 2.6.** *We note that (2.56) provides an affine approximation of the rate, where  $G = \Lambda\gamma + 1$  in the first term defines the DoF for both the MN and the ACC scheme. The second term, which does not depend on the transmit power, defines the rate-offset in the affine approximation, and it explains the improved performance of the ACC scheme. This type of affine approximations have been considered at high-SNR [86], but it will be shown later that, for this setting, it also offers an accurate characterization at practical SNR ranges.*

### Large $B$ Approximation

We derive now an approximation based on assuming a large number of users, which is meant to simplify the previous results. This assumption is well justified by current trends in practical dense scenarios, where  $K$  is expected to be as large as 800 in dense urban Micro-cell settings, and up to 4000 in dense urban Macro-cell settings [88, 89]. Then, in addition to the linear approximation on the capacity, we will also apply the Central Limit Theory after assuming that  $B$  is large<sup>10</sup>.

Let us define  $S_{g,b} \triangleq \rho |h_{g,b}|^2 r_{g,b}^{-\eta_0}$ , and let  $\varrho_s$  and  $\sigma_s^2$  denote respectively the mean and the variance of  $\ln(S_{g,b})$ . Furthermore, let  $-\mathcal{H}_G$  denote the expectation of the minimum of  $G$  i.i.d. Gaussian random variables with zero-mean and unit-variance (cf. (2.21)).

**Lemma 2.15.** *For  $B \rightarrow \infty$ , the affine-approximated average rate of the ACC scheme satisfies*

$$\tilde{R}^{\text{ACC}} \approx G \left( \varrho_s - \mathcal{H}_G \sqrt{\sigma_s^2/B} \right), \quad (2.57)$$

where  $X \approx Y$  means that  $X$  can be approximated by  $Y$  with vanishing error as  $B \rightarrow \infty$ , and where  $\varrho_s$  and  $\sigma_s^2$  are given by

$$\varrho_s = \ln \rho - C + \frac{\eta_0}{2} - \eta_0 \frac{D_2^2 \ln D_2 - D_1^2 \ln D_1}{D_2^2 - D_1^2}, \quad (2.58)$$

$$\sigma_s^2 = \frac{\pi^2}{6} + \frac{\eta_0^2}{4} - \eta_0^2 \frac{D_1^2 D_2^2 (\ln D_2 - \ln D_1)^2}{(D_2^2 - D_1^2)^2}. \quad (2.59)$$

*Proof.* See Appendix A.7. □

Now, Lemma 2.15 and the definition of  $\varrho_s$  directly yield the next corollary, which approximates the effective coded caching gain of the ACC scheme by the ratio  $\tilde{G}_{\text{ACC}}$  of the affine-approximated average rates of the ACC scheme and TDM.

**Corollary 2.5.** *The approximate effective coded caching gain of the ACC scheme takes the form*

$$\tilde{G}_{\text{ACC}} \approx \frac{G}{\varrho_s} \left( \varrho_s - \mathcal{H}_G \sqrt{\sigma_s^2/B} \right), \text{ for } B \rightarrow \infty, \quad (2.60)$$

where  $\varrho_s$  is given by (2.58) and  $\sigma_s^2$  by (2.59).

A quick note here is that the ACC scheme can be extended to the case where  $\Lambda$  does not divide  $K$  with minor impact in the performance, and such performance will converge to the result in (2.60) as  $\frac{K}{\Lambda}$  increases.

<sup>10</sup>Numerical results in Fig. 2.19 will clearly illustrate the range of  $B$  for which this large- $B$  approximation holds reasonably well.

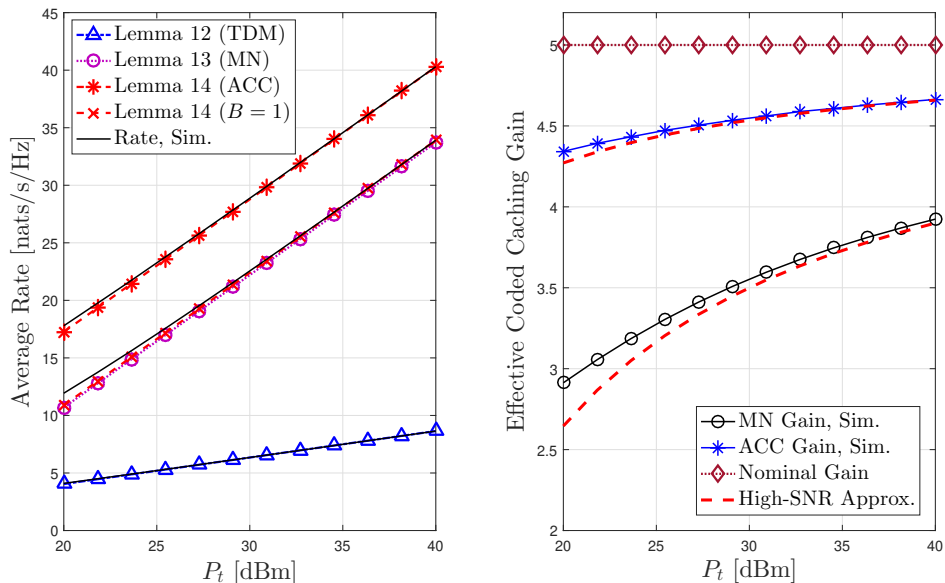


Figure 2.17: Performance in a dense urban Micro-Cell, for the case of  $K = 800$ ,  $\gamma = 5\%$ , and  $\Lambda = 80$  (edge SNR: 15  $\rightarrow$  35dB).

## 2.5.4 Numerical Results

We validate our analytical results through Monte-Carlo simulations for two main 3GPP scenarios: the dense urban Micro-cell setting and the dense urban Macro-cell setting. In accordance to 5G standards [84, 90], we consider  $f_{\text{GHz}} = 3.5$  GHz,  $B_w = 20$  MHz,  $N_0 = -174 \frac{\text{dbm}}{\text{Hz}}$ , and  $N_f = L_c = 10$  dB; following the 3GPP proposition in [84], we consider  $\beta_f(\text{dB}) = 32.4 + 20 \log_{10} f_{\text{GHz}}$  and  $\eta_0 = 2.1$  for the Micro-cell scenario, and  $\beta_f(\text{dB}) = 28 + 20 \log_{10} f_{\text{GHz}}$  and  $\eta_0 = 2.2$  for the Macro-cell scenario. Following the guidelines in [88, 90] for cell sizes, and recalling that users are located in a ring around the transmitter, we consider a distance range of  $[D_1 = 10, D_2 = 100]$  meters in the Micro-cell setting and of  $[D_1 = 35, D_2 = 300]$  meters in the Macro-cell setting.

Regarding the SNR operation range, we note the following: In the Macro-cell setting, a power level of  $P_t = 20$  dBm entails an average SNR at the users on the cell edge ( $D_2 = 300$ ) of 7 dB, while  $P_t = 50$  dBm corresponds to 35 dB of SNR at the edge users. Typical values of transmit power are commonly considered to be  $P_t = 33$  dBm and  $P_t = 40$  dBm for Micro-cell and Macro-cell, respectively [84, 88]. We have broadened the range of power values to provide a wider perspective.

Following [88, 89], we consider  $K = 800$  in the Micro-cell case and  $K = 2000$  in the Macro-cell case. We plot in Figs. 2.17–2.18 the results from Lemmas 2.12–2.14 (dashed lines with markers), and validate these by also presenting the corresponding exact performance computed from Monte-Carlo simulations (solid lines) in Figs. 2.17–2.18. As we can see, the high-SNR analysis closely characterizes the realistic dense Micro-cell and Macro-cell settings for practical  $P_t$  values. As we can see, the ACC scheme considerably

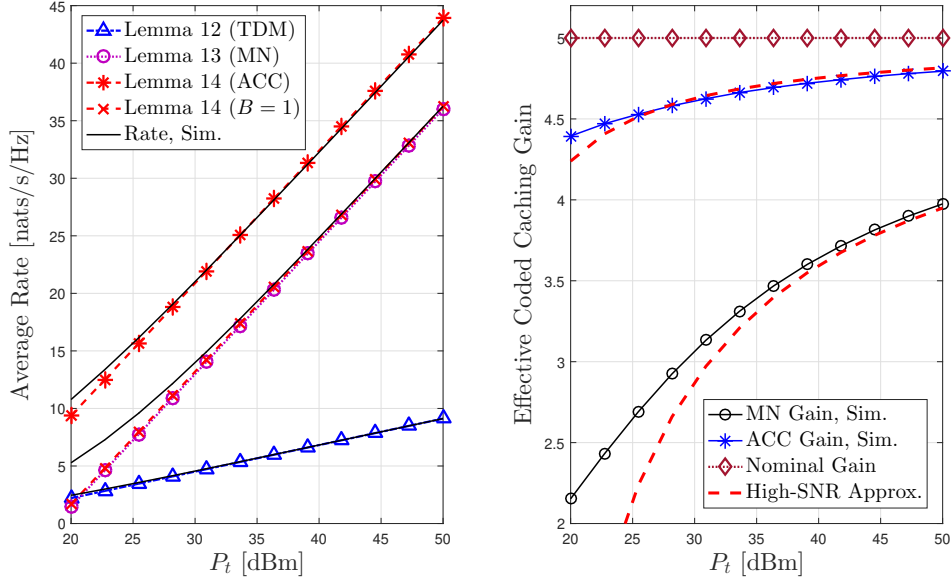


Figure 2.18: Performance in a dense urban Macro-Cell for the case of  $K = 2000$ ,  $\gamma = 5\%$ , and  $\Lambda = 80$  (edge SNR:  $7 \rightarrow 35$ dB).

outperforms the MN scheme, covering most of the nominal (ideal) gains. As  $P_t$  decreases, the performance boost of ACC over MN increases, and the ACC gain approaches the nominal gain of  $\Lambda\gamma + 1$ , much earlier, in terms of SNR, than the MN scheme does.

In Fig. 2.19, we plot the average rate and effective gain versus the number of users per cache state ( $B$ ) for the dense urban Macro-cell scenario. Although the rate approximation in Lemma 2.15 is based on the assumption of large  $B$ , Fig. 2.19 (left) shows that the approximation is very accurate even when  $B$  is as low as  $B = 2$ . As expected, the ACC effective gain increases as  $B$  increases, since  $B$  offers a spatial averaging effect. Finally, Fig. 2.19 (right) reveals a performance gap, between the ACC and MN schemes, that is proportional to  $\Lambda$ .

## 2.6 Extension C: Multi-Antenna ACC

In the previous sections, we have witnessed some of the ACC gains, focusing on the case where the transmitter is equipped with a single antenna. As one would expect, we must now incorporate the ACC idea into the multi-antenna transmitter setting. In this section, we will investigate the ACC performance in the multi-user (MU) multicasting scenario when a multi-antenna transmitter delivers a common message to a set of users. A central point of this section is to describe the ability of a properly modified ACC variant to achieve high performance with a small number of radio frequency (RF) chains. Reducing the number of RF chains is a crucial ingredient in designing modern algorithms, because having many RF chains can be expensive and power consuming.

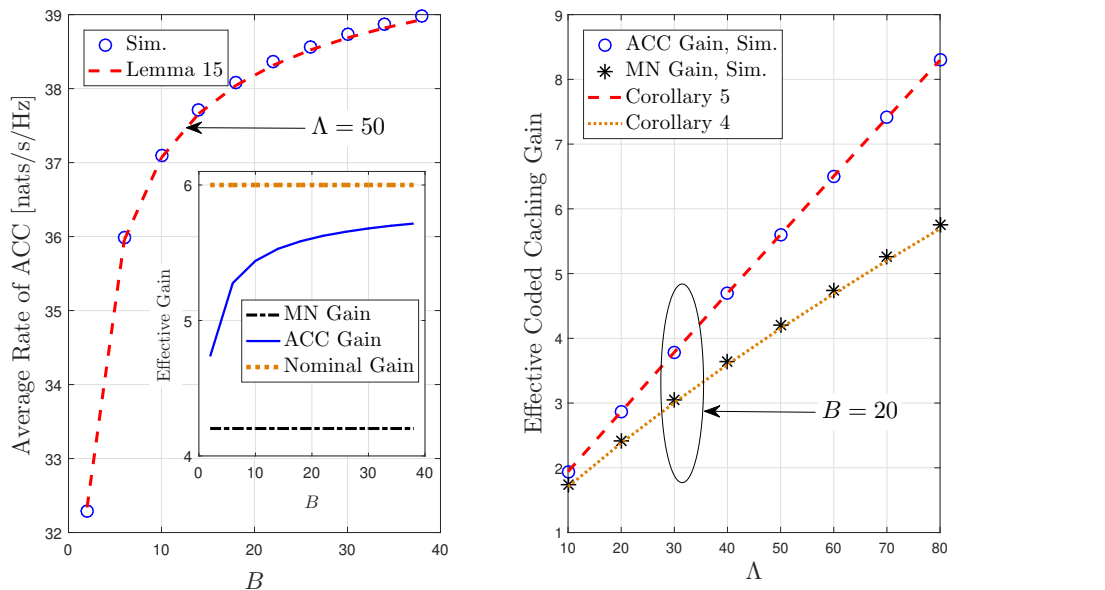


Figure 2.19: Performance in the Macro-Cell setting with  $\gamma = 10\%$  and  $P_t = 40$  dBm.  $\mathcal{H}_G$  is computed through GHQ with 10 terms.

### 2.6.1 Introduction

Consider a BS equipped with  $L$  antennas, serving  $K$  single-antenna users, each requesting a different file from the library  $\mathcal{F}$ . Let us assume that the BS serves  $G$  users at a time (i.e.,  $G$  users over the same time-frequency resource). To achieve this multiplexing capability, in the conventional (cacheless) full-digital (FD) MU-MIMO system, the BS would typically need  $L$  RF chains for unicasting the required  $G$  messages, where these RF chains enable FD precoding techniques such as the widely used zero-forcing (ZF) precoding method. On the other hand, in the cacheless hybrid-precoding MU-MIMO systems, the BS still needs (at least)  $G$  RF chains to deliver  $G$  interference-free signal symbols at a time via a baseband precoder, after which it would employ  $L$  phase shifters for beamforming (RF precoding<sup>11</sup>) [37].

Now, let us consider caching. Assuming that every user has a local cache of normalized size  $\gamma \in [0, 1]$ , then the BS can easily employ a single RF chain to multicast — via coded caching — a common message to  $G$  users at a time, allowing for considerable power-efficiency. As we recall though, the worst-user effect will cause serious performance degradation in this coded multicasting approach. Although there are some works on designing time-efficient and low-complexity transmit beamforming algorithms to boost the delivery rate in a single RF-chain multicasting scenario (cf. [91]) or in an  $L$  RF-chain (FD) scenario (e.g., [92]), these approaches do not explicitly exploit caching, as well as

<sup>11</sup>We note that the RF precoder can only change the phase of the incoming signal, while the baseband precoder can change both the magnitude and phase.

require prohibitively expensive optimization in every transmission.

What we will see below, numerically, is that the ACC idea can in fact compete, with minimal costs, with some of the most expensive existing solutions. Some of the findings are briefly described below.

1. We will show that a cheap hybrid adaptation of ACC can effectively match — over practical SNR values and over symmetric Rayleigh fading channels — the performance of XOR-based (MN) coded caching having a fully optimized FD beamformer. While both approaches multicast to the same number of  $G$  users at a time, and while the FD beamforming approach would require  $L$  RF chains and would need to perform an NP-hard non-convex optimization in every transmission round (cf. [92]), our ACC approach employs a very simple AG beamformer with a single RF chain and no optimization. When the same above comparison is carried over to the case where the  $G$  simultaneously served users experience a different pathloss, then we show that the new ACC scheme (needing a single RF chain) far outperforms the aforementioned XOR-based coded caching approach (that enjoys a fully optimized FD beamformer).
2. We also show that the above same simple ACC-based approach with a single RF chain, can asymptotically (for many users, but for practical SNR values in a realistic, 3GPP proposed, sub-6GHz Macro-cell) matches the performance of the (cacheless) ZF precoder. Here we clarify that both compared methods simultaneously serve the same number of  $G$  users at a time. We also note that the ZF approach would require an inverse of a  $G \times G$  complex-valued matrix in every transmission round.
3. In the setting of multi-antenna BS with a limited number of  $G$  RF chains, it is known that Phased-ZF unicasting to  $G$  users at a time, constitutes a low-complexity hybrid precoder that is also near-optimal [93]. In this work — for typical BS transmit power values, and typical uniformly distributed users in a realistic sub-6GHz Micro-cell — we show that the ACC-aided single-antenna scheme can achieve 80% of the performance provided by the aforementioned near-optimal Phased-ZF (unicasting to these  $G$  users) precoder. Then the performance of this single-antenna ACC variant is validated over realistic mmWave channels. Furthermore, a basic multi-access (NOMA) technique is shown to help this single-antenna ACC effectively accelerate the performance convergence to the ZF/Phased-ZF precoding.

In the following, we first analyze the traditional (MN) cache-aided coded multicasting in the context of a multi-antenna BS with AG/FD beamforming. Then, we design the ACC-aided multicasting, again for the multi-antenna setting, and, as suggested above, we also compare these to the conventional (cacheless) ZF precoding. In the end, some interesting performance comparisons are presented in terms of numerical results.

### 2.6.2 Multi-antenna Cache-aided Coded Multicasting (Multi-antenna MN)

Let  $s \in \mathbb{C}$  with zero-mean and unit-variance be the multicasting data signal (MN XOR) for the simultaneously served  $G$  users in this cache-aided multicasting scenario. Let

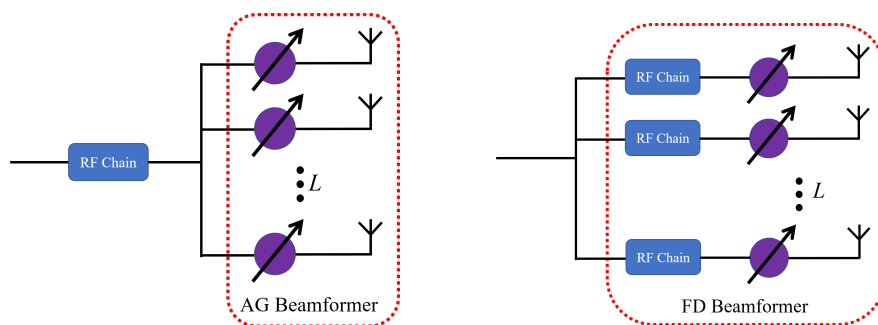


Figure 2.20: Multicasting using a single-RF-chain analog (AG) beamformer (left), and multicasting using an  $L$ -RF-chain full digital (FD) beamformer (right).

$\mathbf{f} \triangleq \mathbb{C}^{L \times 1}$  with  $\|\mathbf{f}\| = 1$  be the beamformer. Let  $\mathbf{x} = \sqrt{P_t} \mathbf{f} s \in \mathbb{C}^{L \times 1}$  be the transmitted signal at the BS, where  $P_t$  is the transmit power at the BS. Let  $\Psi$ , with  $|\Psi| = G$ , be the index set indicating the simultaneously served users during a specific multicasting transmission (i.e., during the delivery of a single XOR). Then, the received signal at user  $U_\psi$  for  $\psi \in \Psi$  takes the form

$$y_\psi = \sqrt{P_t} \mathbf{h}_\psi^T \mathbf{f} s + z_\psi, \quad (2.61)$$

where  $\mathbf{h}_\psi \in \mathbb{C}^{L \times 1}$  is the channel vector from the BS to  $U_\psi$ , and where  $z_\psi$  is the AWGN with power  $N_0$ . For  $\rho = P_t/N_0$  denoting the transmit SNR, then the corresponding received SNR at  $U_\psi$  takes the form

$$\text{SNR}_\psi = \frac{P_t |\mathbf{h}_\psi^T \mathbf{f}|^2}{N_0} = \rho |\mathbf{h}_\psi^T \mathbf{f}|^2. \quad (2.62)$$

As the multicasting rate is always limited by the worst-user, we here aim to separately design the AG and the FD beamformers (cf. Fig. 2.20) that maximize the minimum SNR among the simultaneously served users. The max-min fairness (MMF) optimizations are formulated respectively as

$$\text{AG Beamformer: } \mathbf{f}_{\text{AG}}^* = \arg \max_{\mathbf{f}} \min_{\psi \in \Psi} \left\{ |\mathbf{h}_\psi^T \mathbf{f}|^2 \right\}, \text{ s. t. } |\mathbf{f}(\ell)|^2 = 1/L, \text{ for any } \ell \in [L] \quad (2.63)$$

$$\text{FD Beamformer: } \mathbf{f}_{\text{FD}}^* = \arg \max_{\mathbf{f}} \min_{\psi \in \Psi} \left\{ |\mathbf{h}_\psi^T \mathbf{f}|^2 \right\}, \text{ s. t. } \|\mathbf{f}\| = 1, \quad (2.64)$$

where  $\mathbf{f}(\ell)$  denotes the  $\ell$ -th element of  $\mathbf{f}$ . It is worth noting that both (2.63) and (2.64) are non-convex NP-hard problems [91, 92]. After numerically solving the MMF problems, given the spectrum bandwidth  $B_w$ , the sum rate in this cache-aided multicasting scenario is

$$R_{\text{sum}}^\zeta = G \min_{\psi \in \Psi} \left\{ B_w \log_2 \left( 1 + \rho |\mathbf{h}_\psi^T \mathbf{f}_i^*|^2 \right) \right\} \text{ bits/sec}, \quad (2.65)$$

where  $\iota \in \{\text{AG}, \text{FD}\}$ . As explained before, it is reasonable to assume that the channel remains constant during a packet transmission<sup>12</sup>. At this point, we can write the delivery time for sending a packet to these  $G$  users in the form

$$T_\iota = \frac{F_{\text{sub}}}{\min_{\psi \in \Psi} \left\{ B_w \log_2 \left( 1 + \rho |\mathbf{h}_\psi^T \mathbf{f}_\iota^*|^2 \right) \right\}} \text{ sec, for } \iota \in \{\text{AG}, \text{FD}\} \quad (2.66)$$

where  $F_{\text{sub}}$  denotes the packet (subfile) size in bits. This is for the MN-based approach.

### 2.6.3 ACC-Aided Multicasting

We recall from Section 2.2 that the key idea of the ACC approach is to leverage the need to repeat cache-states across users, in order to achieve a continuous multi-rate transmission and to alleviate the worst-user effect. This also involved the use of nested modulation [94] to achieve multi-rate transmissions to different users, where each user can decode the received (common) signal at its own link's capacity (as if no other user is involved in this transmission). This ability was also based on the fact that each user knows the messages intended for other simultaneously served users (cf. Proposition 2.1). We also recall that when a user completes decoding, another user having the same cache content will replace it automatically, and the BS will generate new multicasting messages that also involves the message for that new user, and thus the number of users served at a time generally does not reduce. We refer to Section 2.2 for more details about the ACC.

We here explore ACC in the presence of a very simple AG (conjugate) beamformer without any optimization<sup>13</sup>. This beamformed ACC approach is represented by

$$\mathbf{f}_{\text{ACC}} = \mathbf{r} \circ \left( \sum_{\psi \in \Psi} \mathbf{h}_\psi^* \right), \quad (2.67)$$

where  $\mathbf{r}$  is the normalization vector such that  $|\mathbf{f}_{\text{ACC}}(\ell)|^2 = 1/L$  for any  $\ell \in [L]$ , and where the operator  $\circ$  denotes the Hadamard product. Specifically,  $\mathbf{r}(\ell) = \left( \sqrt{L} \left| \sum_{\psi \in \Psi} \mathbf{h}_\psi^*(\ell) \right| \right)^{-1}$ . Also the single-link channel capacity to the served user in the cache state  $\psi \in \Psi$  is

$$R_\psi = B_w \log_2 \left( 1 + \rho |\mathbf{h}_\psi \mathbf{f}_{\text{ACC}}|^2 \right) \text{ bits/sec.} \quad (2.68)$$

Let us also recall that we must serve  $B$  users per cache state, thus we are interested in reducing the delivery time when serving  $GB$  users in the selected (cache-state) set  $\Psi$ , where this delivery is outlined in Algorithm 2, where the value of the  $i$ -th element in  $\mathbf{v} \in \mathbb{Z}^{|\Psi|}$  denotes the number of users that have been served in the  $i$ -th cache-state in  $\Psi$ , and where  $W_{d_{\mathbf{v}(i)}}^{\Psi \setminus \Psi(i)}$  denotes<sup>14</sup> the subfile intended for user  $\mathbf{v}(i)$ . Specifically, when a group

<sup>12</sup>Recall that in coded caching, packets are expected to be relatively small, due to the subpacketization requirement.

<sup>13</sup>Interestingly, the impact of the beamformer optimization can be marginal in the ACC-aided delivery, as shown in the numerical results in Fig. 2.26.

<sup>14</sup>Below we will slightly abuse notation when considering the input parameters of encoding function  $\mathcal{X}$  (which is responsible for achieving the performance described in Proposition 2.1) in Algorithm 2.

---

**Algorithm 2:** Delivery time ( $T_{\text{ACC}}$ ) for a cache-state set  $\Psi$  with  $GB$  users
 

---

```

1 Initialize  $T_{\text{ACC}} \leftarrow 0$ ;  $\mathbf{v} \in \mathbb{Z}^{|\Psi|}$  as  $\mathbf{v}(i) \leftarrow 1$  for any  $i \in [|\Psi|]$ ;
2 Initialize Number of finished groups  $\leftarrow 0$ ;
   Packet  $\leftarrow [F_{\text{sub}}, F_{\text{sub}}, \dots, F_{\text{sub}}] \in \mathbb{Z}^{|\Psi|}$ 
3 while Number of finished groups  $\neq |\Psi|$  do
4   Transmit
5      $X_{\Psi, \mathbf{v}} \leftarrow \mathcal{X} \left( \left\{ W_{d_{\mathbf{v}(i)}}^{\Psi \setminus \{\Psi(i)\}} \mid i \in [|\Psi|] \text{ and } \mathbf{v}(i) \leq B \right\} \right)$ 
6     Transmit Beamformer:  $\mathbf{x} \leftarrow \mathbf{f}_{\text{ACC}} X_{\Psi, \mathbf{v}}$ , where
        $\mathbf{f}_{\text{ACC}} \leftarrow \mathbf{r} \circ \left( \sum_{i \in [|\Psi|], \mathbf{v}(i) \leq B} \mathbf{h}_{\Psi(i)}^* \right)$ 
7   until A served user  $\mathbf{v}(i)$ ,  $i \in [|\Psi|]$ , fully obtains its subfile
8   Set  $i^*$  as the index of the group  $\Psi(i^*)$  whose user has decoded its subfile
9   Calculate
10     $T_{i^*} \leftarrow \min_{j \in [|\Psi|]} \frac{\text{Packet}(j)}{B_w \log_2 (1 + \rho |\mathbf{h}_{\Psi(j)}^T \mathbf{f}_{\text{ACC}}|^2)}$ ;  $T_{\text{ACC}} \leftarrow T_{\text{ACC}} + T_{i^*}$ ;
11    Packet( $i$ )  $\leftarrow$  Packet( $i$ )  $- T_{i^*} B_w \log_2 (1 + \rho |\mathbf{h}_{\Psi(i)}^T \mathbf{f}_{\text{ACC}}|^2)$  for any  $i \neq i^*$ ;
12    Packet( $i^*$ )  $\leftarrow F_{\text{sub}}$ 
13  end
14  if  $\mathbf{v}(i^*) = B$  then
15    Number of finished groups  $\leftarrow$  Number of finished groups + 1
16    Packet( $i^*$ )  $\leftarrow +\infty$ 
17   $\mathbf{h}_{\Psi(i^*)} \leftarrow$  New Channel Vector
18   $\mathbf{v}(i^*) \leftarrow \mathbf{v}(i^*) + 1$ 

```

---

updates its served user, the transmitter continues encoding the partially-decoded subfiles, taking into account that there remains only a part of such subfiles to be transmitted. Fig. 2.1 illustrates this.

At this point, we quickly point out an interesting connection between ACC and the desired ability to perform more precise beamforming. In particular, we note that when the users in a cache state  $\psi \in \Psi$  are all served, these users are removed from the transmit beamforming process, and thus with fewer users, we will be able to create a much more focused transmit beam toward the other simultaneously served users, thereby accelerating the delivery to them. Thus even when the degree of multicasting is reduced, the beamforming efficiency is increased.

#### 2.6.4 Cacheless Full-Digital Precoding

A traditional cacheless approach involves the BS employing a hybrid/FD precoder to simultaneously transmit multiple symbols to a set of users (still denoted by  $\Psi$  - which now has nothing to do with multicasting, and which rather simply denotes the set of users served at a given time). This corresponds to conventional (cacheless) MU-MIMO systems,

where indeed each delivered symbol is intended by a single dedicated user. Under the linear precoding scheme, the transmit signal at the BS takes the form  $\mathbf{x} = \sqrt{\rho}\mathbf{W}\mathbf{s}$ , where  $\mathbf{W} \in \mathbb{C}^{L \times G}$  is the linear precoder,  $\mathbf{s} \in \mathbb{C}^G$  is the data vector for the  $G$  users served at a time, and  $\rho = \frac{P_t}{\text{Tr}\{\mathbf{W}^H\mathbf{W}\}}$  regulates the transmit power  $P_t$ . The received signal at a typical user  $U_\psi$  takes the form

$$y_\psi = \sqrt{\rho}\mathbf{h}_\psi^T\mathbf{W}\mathbf{s} + z_\psi = \sqrt{\rho}\mathbf{h}_\psi^T\mathbf{w}_\psi s_\psi + \sqrt{\rho}\sum_{\phi \in \Psi \setminus \psi} \mathbf{h}_\psi^T\mathbf{w}_\phi s_\phi + z_\psi, \quad (2.69)$$

where  $\mathbf{w}_\psi$  denotes the  $\psi$ -th column of  $\mathbf{W}$ , and where  $z_\psi$  is the AWGN with power  $N_0$ . Also the achievable rate for  $U_\psi$  takes the form

$$R_\psi = B_w \log_2 \left( 1 + \frac{\rho |\mathbf{h}_\psi^T \mathbf{w}_\psi|^2}{N_0 + \rho \sum_{\phi \in \Psi \setminus \psi} |\mathbf{h}_\psi^T \mathbf{w}_\phi|^2} \right) \text{ bits/s.} \quad (2.70)$$

Therefore, for a linear precoder  $\mathbf{W}$ , the transmission time for serving the  $G$  users in  $\Psi$ , where each user requires a packet of  $F_{\text{sub}}$  bits, takes the form

$$T_{\mathbf{W}} = \frac{F_{\text{sub}}}{\min_{\psi \in \Psi} B_w \log_2 \left( 1 + \frac{\rho |\mathbf{h}_\psi^T \mathbf{w}_\psi|^2}{N_0 + \rho \sum_{\phi \in \Psi \setminus \psi} |\mathbf{h}_\psi^T \mathbf{w}_\phi|^2} \right)} \text{ sec.} \quad (2.71)$$

### 2.6.5 Performance Metric

To compare the performance boost, we define the effective gain as the ratio of two delivery times for serving  $GB$  users; the numerator is for serving with uncoded TDM transmission (the gain does not take credit for local caching gains), and the denominator corresponds to combining a beamformer with the coded caching scheme of choice. Naturally we average over channel states. Specifically we consider three different gains: the gain of the ACC-aided simple AG beamformer, the gain for the optimal FD beamformer with XOR-based multicasting, and for the optimal AG beamformer for XOR-based multicasting. These gains respectively take the form

$$\mathcal{G}_{\text{ACC}} \triangleq \frac{\mathbb{E}_h\{T_{\text{TDM}}\}}{\mathbb{E}_h\{T_{\text{ACC}}\}}, \quad \mathcal{G}_{\text{FD}} \triangleq \frac{\mathbb{E}_h\{T_{\text{TDM}}\}}{\mathbb{E}_h\{T_{\text{FD}}\}}, \quad \mathcal{G}_{\text{AG}} \triangleq \frac{\mathbb{E}_h\{T_{\text{TDM}}\}}{\mathbb{E}_h\{T_{\text{AG}}\}}, \quad (2.72)$$

where  $T_{\text{ACC}}$  is given in Algorithm 2, while  $T_{\text{AG}}$  and  $T_{\text{FD}}$  can be found via numerically solving (2.63) and (2.64) respectively (e.g., using the built-in function “fminimax” in MATLAB).

For the cacheless FD linear precoding, we consider the ZF precoder as a performance benchmark, which is commonly used in practical MU-MIMO systems, and which takes the form

$$\mathbf{W} = \mathbf{H}_\Psi^H (\mathbf{H}_\Psi \mathbf{H}_\Psi^H)^{-1}, \quad (2.73)$$

where  $\mathbf{H}_\Psi \in \mathbb{C}^{G \times L}$  is the channel matrix from the BS to the users in  $\Psi$ . We also define the effective gain provided by ZF precoding as the ratio of, in the numerator, the delivery

time for serving  $GB$  users using uncoded TDM transmission with a simple (conjugate) AG beamformer (no coded caching), and, in the denominator, the delivery time corresponding to ZF precoding. This gain takes the form

$$\mathcal{G}_{\text{ZF}} \triangleq \frac{\mathbb{E}_h\{T_{\text{TDM}}\}}{\mathbb{E}_h\{T_{\text{ZF}}\}}. \quad (2.74)$$

We refer to Section 2.6.6 for numerical comparisons of  $\mathcal{G}_{\text{ACC}}$ ,  $\mathcal{G}_{\text{FD}}$ ,  $\mathcal{G}_{\text{AG}}$ , and  $\mathcal{G}_{\text{ZF}}$ .

### 2.6.6 Numerical Results

In what follows, we assume i.i.d. complex Gaussian fading with zero-mean and unit-variance, as well as assume an atomic communication packet size of  $F_{\text{sub}} = 50$  bytes. Let us consider Figs. 2.21–2.22. In Fig. 2.21, we can see that the delivery time of the ACC-aided simple AG beamformer converges, in the high SNR region, to the delivery time of the fully optimized FD/AG beamformer with XOR-based multicasting. On the other hand, the performance significantly differs in the low-to-moderate SNR region. This advantage of the fully optimized FD beamformer with XOR-based multicasting is partly due to its ability to employ  $L$  RF chains, albeit at a cost (not recorded here) of a much larger power consumption and a much larger computational complexity from having to continuously solve an NP-hard non-convex optimization problem. It is also worth noting that the ACC-aided delivery with a simple AG beamformer, almost matches the performance of the aforementioned fully optimized FD beamformer, especially in low and high SNRs, especially when we increase the number  $B$  of users per cache state (from  $B = 1$  to  $B = 5$ , in Fig. 2.22).

Then, in Figs. 2.23–2.27, we consider that the users are *uniformly* distributed within a Macro-cell with an inner radius of  $D_1 = 35$  meters and an outer radius of  $D_2 = 500$  meters. When the carrier frequency is 2 GHz, we have the pathloss factor as  $l_0 r^{-\eta_0}$ , where  $l_0 = 10^{-3.53}$  regulates the pathloss at 35 meters, where  $r$  is the distance between the BS and the user, and where  $\eta_0 = 3.76$  is the pathloss exponent (cf. [95]). We also assume that the AWGN density is  $-174$  dBm/Hz. The transmit power at the BS is chosen to lie in  $[20, 60]$  dBm, and the spectrum bandwidth for each user ( $B_w$ ) is chosen at 20 MHz [53, 84, 88].

In particular, Fig. 2.23 plots the delivery time of the ACC-aided simple AG beamformer, the fully optimized AG beamformer with XOR-based multicasting<sup>15</sup>, and the TDM transmission with a conjugate beamformer respectively. We also present the results of the (cacheless) ZF precoding where the BS needs to send  $G$  different data symbols simultaneously over the same time-frequency resource (cf. Section 2.6.4). It is obvious that the ZF precoder has the best performance due to having  $L$  RF chains (FD), despite the largest hardware/software overheads. We also see that the performance of the ACC-aided delivery with a simple AG beamformer is close to that of the ZF precoder, and that the performance converges as  $B$  increases. Fig. 2.24 plots the CDFs of the

<sup>15</sup>As the performance of the fully optimized FD beamformer with XOR-based multicasting are known to be very close to that of the fully optimized AG beamformer (for the same number of served users), we omit the results of this FD beamformer.

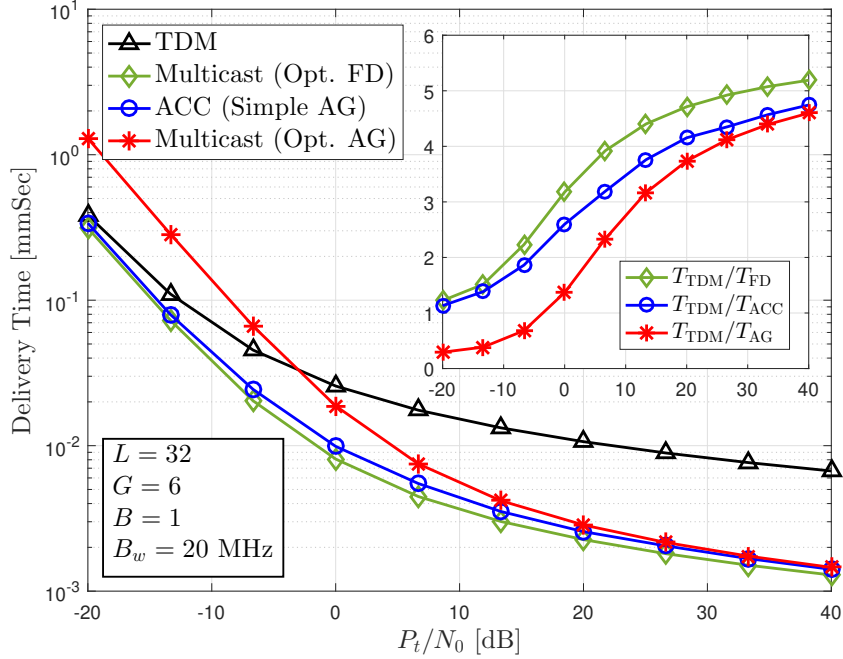


Figure 2.21: Delivery performance over symmetric Rayleigh fading channels.

delivery time of the schemes whose performance is plotted in Fig. 2.23. As shown in Fig. 2.24, for a typical BS transmit power of  $P_t = 40$  dBm in a Macro-cell, compared to the large variance of the fully optimized AG beamformer for XOR-based multicasting, the ACC-aided simple AG beamformer has a much smaller fluctuation around the average delivery time, mainly due to the so-called spatial-averaging effect (cf. Section 2.5), which effectively mitigates and reduces the randomness of channel fading and from the user locations.

In Figs. 2.25–2.26, we see the marginal performance gap between the 4-bit finite resolution phase shifter and the infinite resolution phase shifter, as well as show the negligible gap between the fully optimized FD and AG beamformers in both the (conventional) XOR-based and ACC-aided multicasting scenarios under a realistic Macro-cell setting. In the context of the ACC-aided multicasting, we can also observe that the simple AG beamformer with a 4-bit resolution almost matches the delivery performance of the fully optimized AG beamformer with an infinite resolution. To conclude, the simple AG beamformer with a 4-bit resolution is recommended when practically performing the ACC-aided delivery in this realistic Macro-cell setting.

Fig. 2.27 plots the average rates for the ACC-aided single-antenna BS (referred to as SISO ACC) and for the multi-antenna case with a fully optimized AG beamformer with XOR-based multicasting; both serving  $G$  users at a time. The average rate for a multi-antenna BS with a simple AG beamformer to serve the users one-by-one (i.e., TDM) is here presented as a performance benchmark.

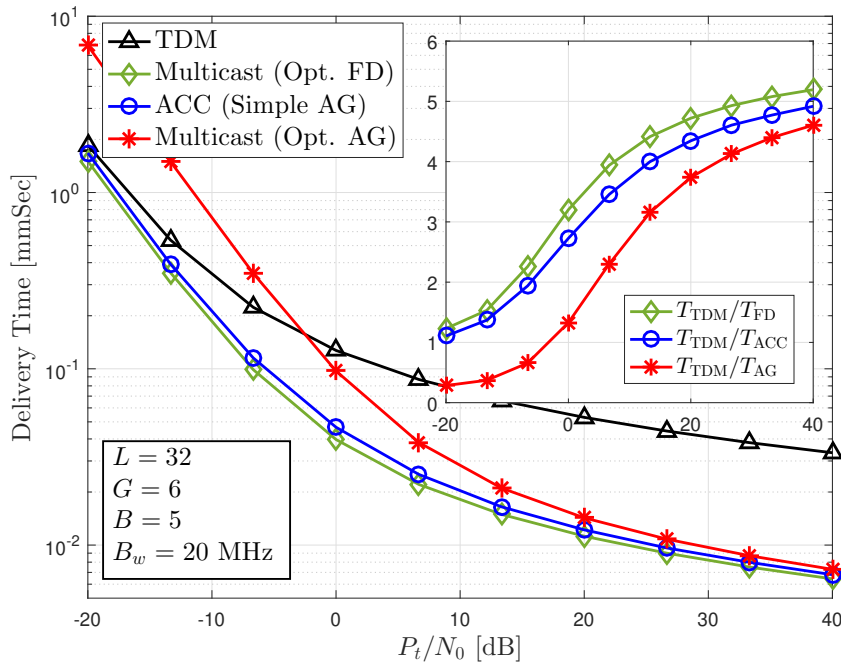
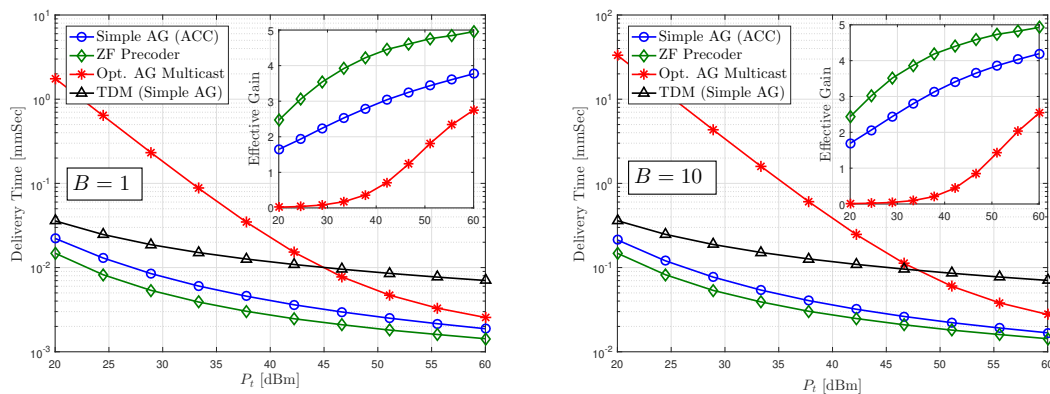


Figure 2.22: Delivery performance over symmetric Rayleigh fading channels.


 Figure 2.23: Delivery performance over non-symmetric Rayleigh fading channels with  $D_1 = 35$ ,  $D_2 = 500$ ,  $L = 128$ ,  $G = 6$ ,  $\eta_0 = 3.76$ , and  $l_0 = 10^{-3.53}$ .

We can observe that the fully optimized AG beamformer outperforms the SISO ACC when  $B = 2$ , but this trend is inverted after  $B$  increases beyond 2, which can imply considerable potential savings in hardware and CSI.

We now seek to further explore SISO ACC and proceed to compare it to the cacheless system where we precode in order to serve the same number of  $G$  users at a time. Toward this, Figs. 2.28–2.29 present the delivery performance of the SISO ACC, the (cacheless) Phased-ZF (PZF) precoder and the (cacheless) ZF precoder. We here select the hybrid

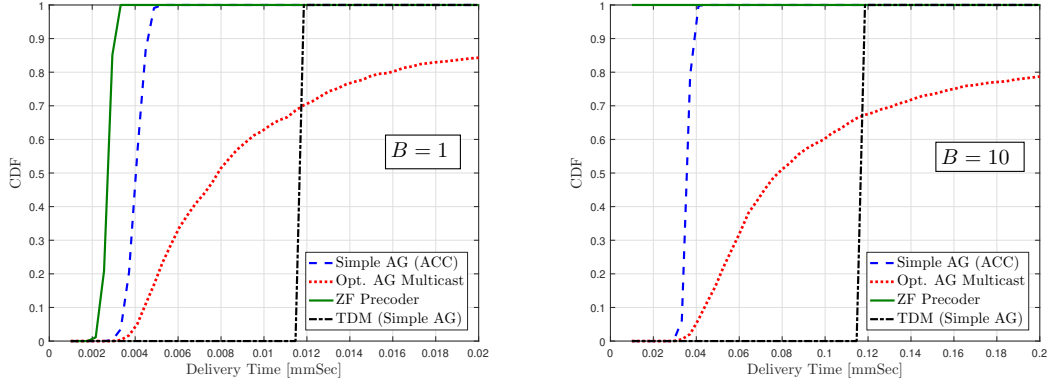


Figure 2.24: Delivery performance over non-symmetric Rayleigh fading channels with  $P_t = 40$  dBm,  $D_1 = 35$ ,  $D_2 = 500$ ,  $L = 128$ ,  $G = 6$ ,  $\eta_0 = 3.76$ , and  $l_0 = 10^{-3.53}$ .

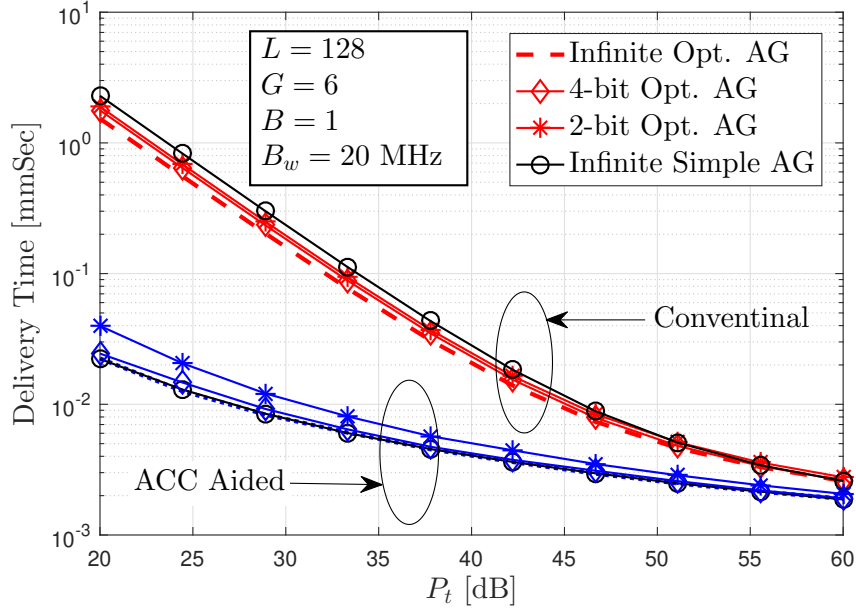


Figure 2.25: Delivery performance over non-symmetric Rayleigh fading channels with  $D_1 = 35$ ,  $D_2 = 500$ ,  $\eta_0 = 3.76$ , and  $l_0 = 10^{-3.53}$ .

PZF precoder proposed in [93] due to its low-complexity and its ability to maintain performance very close to that of the ZF (FD) precoder. The PZF precoder (with  $G$  RF chains) is composed of two cascaded sub-precoders where the first one is responsible for the baseband precoding in order to realize the interference-free transmission to the simultaneously served  $G$  users, while the other one is the RF precoder responsible for analog beamforming. We refer to [93] for more details. Fig. 2.28 plots the delivery

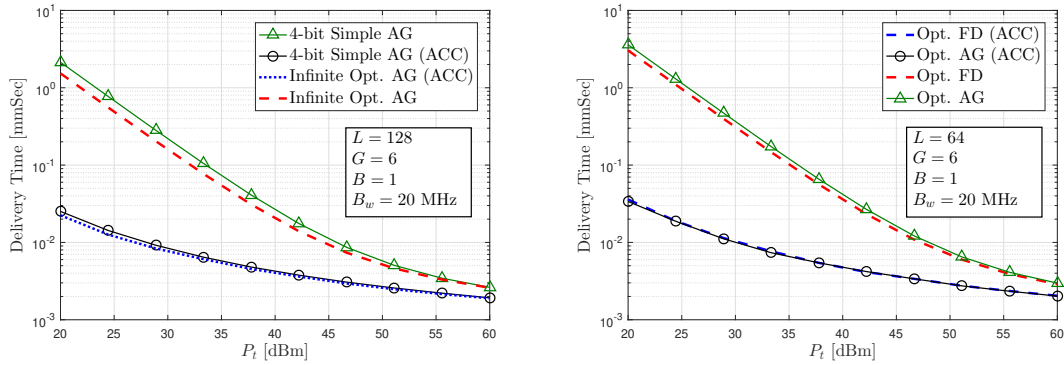


Figure 2.26: Delivery performance over non-symmetric Rayleigh fading channels with  $D_1 = 35$ ,  $D_2 = 500$ ,  $\eta_0 = 3.76$ , and  $l_0 = 10^{-3.53}$ .

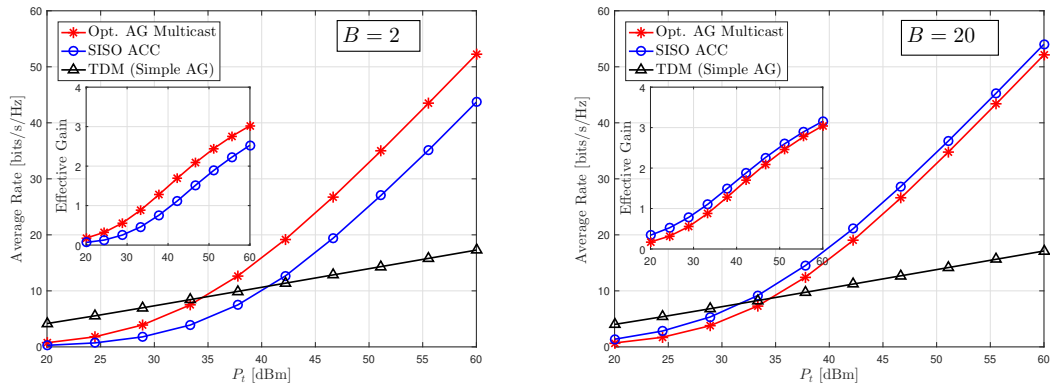


Figure 2.27: Delivery performance over non-symmetric Rayleigh fading channels with  $D_1 = 35$ ,  $D_2 = 500$ ,  $L = 128$ ,  $G = 6$ ,  $\eta_0 = 3.76$ , and  $l_0 = 10^{-3.53}$ .

performance of the SISO ACC, the PZF and the ZF schemes in a realistic Micro-cell setting. As  $B$  increases, the SISO ACC performance improves, and the recorded gain is notable. For example, for a typical BS transmit power of  $P_t = 33$  dBm and for  $B = 10$ , the effective gain of the SISO ACC is here shown to achieve 80% of the (cacheless) PZF gain ( $\Lambda = 50$ ,  $\gamma = 0.1$  and  $K = 500$ ). To cover the remaining gap, we also consider briefly the use of Non-Orthogonal Multiple Access (NOMA) as a boost to the SISO ACC approach. This NOMA-aided ACC brings about some additional gain compared to the original SISO ACC<sup>16</sup>. Fig. 2.28 plots the performance of NOMA-aided SISO ACC, revealing a marginal performance gap between this SISO ACC approach and the (cacheless) ZF/PZF precoder.

<sup>16</sup>We note that the total CSI overheads in the NOMA-aided ACC and in the original ACC should be approximately the same; both approaches use CSI for rate adaptation to  $GB$  users. The only additional complexity brought about by NOMA is that each user has to perform successive interference cancellation (SIC) to decode its own message [96].

Fig. 2.29 validates the advantage of the SISO ACC over mmWave channels, where the channel gain vector  $\mathbf{h}_\psi \in \mathbb{C}^{L \times 1}$  of a uniform linear array (ULA) from the BS to  $U_\psi$  is given by [97]

$$\mathbf{h}_\psi = \sqrt{\frac{L}{N_p}} \sum_{\ell=1}^{N_p} h_\ell^{(\psi)} \mathbf{a}(\varepsilon_\ell^{(\psi)}), \quad (2.75)$$

where  $N_p$  is the number of propagation paths, where  $h_\ell^{(\psi)} \sim \mathcal{CN}(0, 1)$  is the complex gain of the  $\ell$ -th propagation path, where  $\varepsilon_\ell^{(\psi)}$  is the angle of departure (AoD) which is uniformly distributed over  $[-\frac{\pi}{2}, \frac{\pi}{2}]$ , and where  $\mathbf{a}(\varepsilon_\ell^{(\psi)})$  is the array response vector depending only on array structures. When the array structure is the ULA and the transmit antennas are spaced at half wavelength, we can write  $\mathbf{a}(\varepsilon_\ell^{(\psi)})$  as [97]

$$\mathbf{a}(\varepsilon_\ell^{(\psi)}) = \frac{1}{\sqrt{L}} \left[ 1, \exp(j\pi \sin(\varepsilon_\ell^{(\psi)})), \dots, \exp(j\pi(L-1) \sin(\varepsilon_\ell^{(\psi)})) \right]^T. \quad (2.76)$$

In Fig. 2.29, we then assume that there are two LOS paths (i.e.,  $N_p = 2$ ). We note that for the typical far-field LOS propagation,  $N_p = 1$  is recommended [91]. In this case, the channel matrix  $\mathbf{H}_\Psi \in \mathbb{C}^{G \times L}$  does not always have full row rank, which brings substantial additional complexity to the ZF/PZF precoder. We also note that it is practically infeasible to implement the ZF precoding due to the requirement of  $L$  costly RF chains over mmWave channels. In Fig. 2.29, we use  $\text{SNR} \triangleq \frac{P_t}{N_0} L_{\text{PL}}$  in dB as the reference value for simplification in the horizontal axis, which includes the BS transmit power  $P_t$ , the AWGN power  $N_0$  and the pathloss  $L_{\text{PL}}$ . As we can see, the SISO ACC achieves 64% of the (cacheless) PZF effective gain for  $\text{SNR} = 10$  dB (medium SNR), while this figure grows to 87% for  $\text{SNR} = 30$  dB (high SNR). The effective gain of the SISO ACC can be further improved by adopting the NOMA technique (previously shown in Fig. 2.28) over mmWave channels; this is left for future work. It is also worth noting that for 32 transmit antennas, the ACC-aided multicasting with a simple AG beamformer<sup>17</sup> (one RF chain and no optimization) outperforms the (cacheless) PZF precoder ( $G$  RF chains) over the entire SNR region (see Fig. 2.29).

## 2.7 Conclusions

This part of the thesis was motivated by the fact that any attempt to successfully adopt wireless coded caching in large-scale settings, must account for the effects of low-to-moderate SNR fading channels. Our work first revealed that dedicated caches and XOR-based transmissions may no longer be suitable for various realistic SNR regimes. As we have seen, as the SNR becomes smaller, the effective gains of XOR-based schemes collapse, irrespective of the nominal gain or the number of users. We have then proposed a novel dual idea that combines the use of cache replication and a multi-rate transmission

<sup>17</sup>The average (sum) rate of this ACC-aided simple AG beamforming is derived by numerically evaluating the expectation of  $\frac{G B F_{\text{sub}}}{T_{\text{ACC}}}$  over channel states, where  $T_{\text{ACC}}$  is obtained via Algorithm 2.

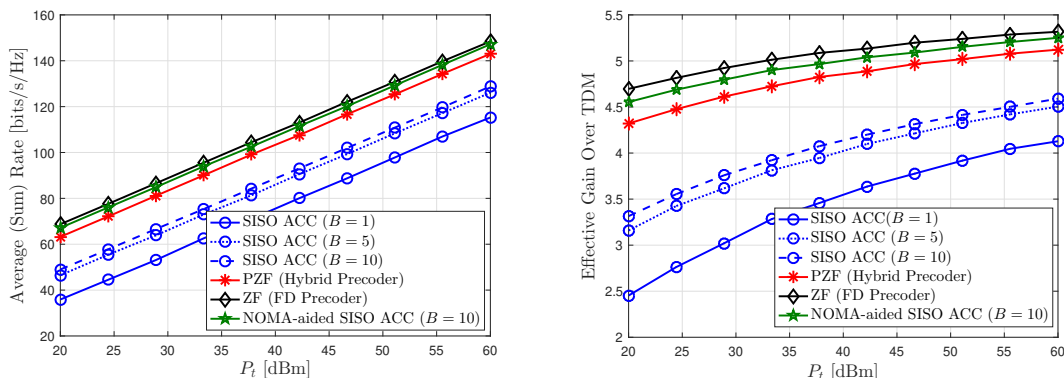


Figure 2.28: Delivery performance over non-symmetric Rayleigh fading channels with  $D_1 = 10$ ,  $D_2 = 100$ ,  $L = 32$ ,  $G = 6$ ,  $\eta_0 = 3$ , and  $l_0 = 10^{-3.7}$ .

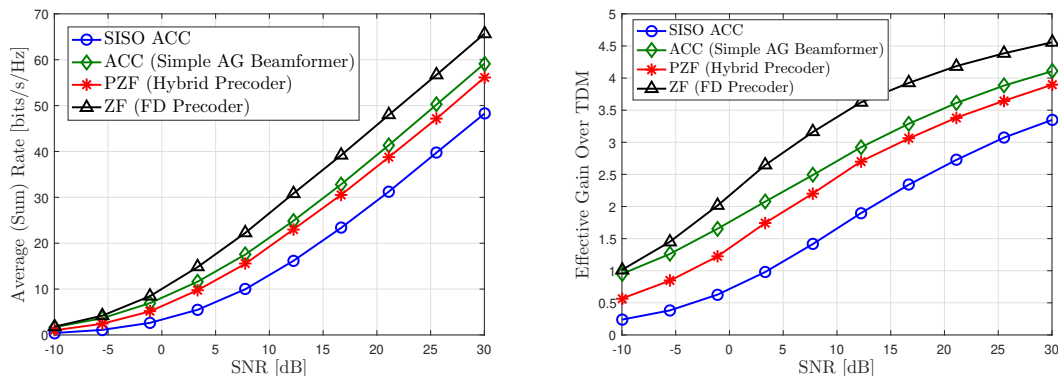


Figure 2.29: Delivery performance over mmWave channels with  $L = 32$ ,  $G = 6$ ,  $B = 5$ ,  $B_w = 200$  MHz,  $N_p = 2$  and  $F_{\text{sub}} = 50$  bytes.

scheme. This approach recovers a big fraction of the lost gains and does so for any SNR value. These gains are fully recovered in the regime of many users, again for any SNR value, thus essentially resolving the worst-user bottleneck. The use of cache replication or shared cache states (which, again, is effectively enforced in practical coded caching settings due to the file-size constraint) is particularly beneficial in lower SNR regimes. We have also shown how having as few as 2 users per cache state allows the proposed scheme to approximately double the effective coded caching gain over symmetric Rayleigh fading channels. As stated before, these gains do not involve user selection, and the corresponding user-grouping is done prior to cache-placement and is oblivious to the demands and of course oblivious to the channel. Subsequently, we have validated the stated gain recovery ability brought about by the ACC scheme in terms of delivery time over symmetric Nakagami- $m$  fading channels, as well as extensively validated it with different pathloss among the served users where the corresponding system parameters adhere to 3GPP recommendations for an urban propagation environment. In the end, the

presented ACC idea applies toward showing that properly designed coded caching has the ability to substantially speed up delivery of multimedia content even in the challenging environment of low-to-moderate SNR fading channels.

In the last section of this chapter, we have modified the ACC scheme to work in conjunction with a multi-antenna transmitter which has been widely considered in current and future wireless networks. In the context of MU multicasting, the ACC-aided simple AG beamformer is shown to be as powerful as the fully optimized FD beamformer for XOR-based multicasting over symmetric Rayleigh fading channels, and to far outperform this FD beamformer when the users are uniformly distributed in a 3GPP proposed sub-6GHz Macro-cell. Interesting, this ACC-aided simple AG beamformer can asymptotically match the performance of the (cacheless) ZF precoder for practical SNR values as the number of users increases in the aforementioned sub-6GHz Macro-cell. To further explore the brilliant advantage brought about by the ACC idea, we have also demonstrated the marginal performance gap in terms of average (sum) rate between the ACC-aided single-antenna BS and the (cacheless) multi-antenna BS employing ZF/PZF precoding in a realistic sub-6GHz Micro-cell and in mmWave channels, which is considerably meaningful in saving hardware/software overheads and CSI costs. These results reveal the fact that with the help of caching, we can use much fewer PHY resources (e.g., RF chains, antenna elements, power consumption, CSI costs and computational complexity) to achieve the same spectral efficiency of the cacheless system.

## Chapter 3

# Vector Coded Caching: Design and Analysis

We introduce the system model and the considered framework in Section 3.1. Subsequently, in Section 3.2, we first adapt the vector coded caching to realistic SNR values, while considering three different linear precoding schemes: ZF, RZF and MF. After doing so, we proceed to employ random matrix theory to analyze (in Theorem 3.1 for MF, Theorem 3.2 for ZF, and Theorem 3.3 for RZF) the achievable throughput of vector coded caching for the three aforementioned precoders.

Subsequently, based on the derived asymptotic performance, in Section 3.3 we optimize both the cacheless as well as the cache-aided algorithms by accounting for the CSI acquisition costs, and by optimizing over the total number of simultaneously served streams (users). This optimization, which is performed as a function of SNR, of  $L$  and of the CSI acquisition costs, can be found in Theorems 3.4, 3.5. The same optimization yields systems that are separately calibrated to better balance multiplexing gains with beamforming gains, in the presence or absence of caching. In this same section we also derive the ratio between the throughputs of the (independently) optimized cache-aided and cacheless systems. This ratio represents the multiplicative throughput boost offered by caching, over optimized cacheless downlink systems with the same power and antenna resources. Subsequently, in Section 3.4 we numerically verify the accuracy of the derived expressions, showing that they characterize very precisely the actual performance. This evaluation allows us to demonstrate the substantial gains from using caching, highlighting realistic regimes of SNR,  $L$ , CSI costs, file sizes and cache sizes. In Section 3.5 we present the main conclusions.

### 3.1 System Model and Problem Description

#### 3.1.1 System Model

We consider a downlink MISO scenario where an  $L$ -antenna base station (BS) serves  $K$  single-antenna cache-aided users. The BS has access to a library of  $N$  equally-sized files, and each user is endowed with a local memory (or cache) such that each user can store a

fraction  $\gamma \in [0, 1)$  of the library content. We denote the library content by  $\mathcal{F}$  and the  $n$ -th file by  $W_n$ , such that  $\mathcal{F} \triangleq \{W_n\}_{n=1}^N$ .

We consider the wireless channel to be modeled as a symmetric Rayleigh fading channel, where all channel coefficients are assumed to be independent and identically distributed (i.i.d.). When describing a general transmission, our notation will often incorporate the subset  $\mathcal{K} \subseteq [K]$  of users that are simultaneously served during that transmission. Consequently, in our communication model, the received signal at the  $k$ -th user in  $\mathcal{K}$  is given by

$$y_{\mathcal{K}(k)} = \mathbf{h}_{\mathcal{K}(k)}^T \mathbf{x}_{\mathcal{K}} + z_{\mathcal{K}(k)}, \quad (3.1)$$

where  $k \in [|\mathcal{K}|]$ , where  $z_{\mathcal{K}(k)} \in \mathbb{C}$  represents the corresponding AWGN with zero-mean and unit-variance, where  $\mathbf{x}_{\mathcal{K}} \in \mathbb{C}^{L \times 1}$  denotes the transmitted signal vector that simultaneously serves the users in  $\mathcal{K}$ , and where  $\mathbf{h}_{\mathcal{K}(k)} \in \mathbb{C}^{L \times 1}$  represents the channel vector for the channel from the BS to the  $k$ -th user in  $\mathcal{K}$ . As mentioned,  $\mathbf{h}_{\mathcal{K}(k)}$  is assumed to be an i.i.d. Gaussian random vector with mean  $\mathbf{0}_L$  and covariance matrix  $\mathbf{I}_L$ . Finally,  $\mathbf{x}_{\mathcal{K}}$  is obtained by applying a specific precoding scheme (which we will detail later on) to the information vector  $\mathbf{s}_{\mathcal{K}} \in \mathbb{C}^{|\mathcal{K}| \times 1}$  intended for the users in  $\mathcal{K}$ , where  $\mathbf{s}_{\mathcal{K}}$  has mean  $\mathbf{0}_{|\mathcal{K}|}$  and covariance matrix  $\mathbf{I}_{|\mathcal{K}|}$ .

We consider an average power normalization, where the power is averaged over both transmit symbols and channel realizations, i.e.,  $\mathbb{E}\{|\mathbf{x}_{\mathcal{K}}|^2\} \leq P_t$ , where  $P_t$  is the average power constraint.

As is common in practical downlink settings, we assume TDD uplink-downlink transmissions, such that the BS estimates the downlink channels through uplink pilot transmissions by applying channel reciprocity.

We proceed to describe the main structure of the scheme, first doing so without specifying the linear precoding class that is used. We will also formally define the main performance metrics investigated in this paper.

### 3.1.2 Signal-Level Vector Coded Caching for Finite SNR

Building on the general vector-clique structure in [98], we are here free to choose the precoding schemes, as well as calibrate at will the dimensionality of each vector clique. This freedom is essential in controlling the impact of CSI costs and of power-splitting across users, both of which directly affect the performance in practical SNR regimes.

We proceed to describe the cache placement phase and the subsequent delivery phase.

#### Placement Phase

The first step involves the partition of each library file  $W_n$  into  $\binom{\Lambda}{\Lambda\gamma}$  non-overlapping equally-sized subfiles  $\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\}$ , each labeled by some  $\Lambda\gamma$ -tuple  $\mathcal{T} \subseteq [\Lambda]$ . As discussed in Chapter 1, the number of cache states  $\Lambda$  is chosen to satisfy the file-size constraint; in our case, the subpacketization is  $\binom{\Lambda}{\Lambda\gamma}$ , which naturally serves as a lower bound on the file sizes. Subsequently the  $K$  users are *arbitrarily* separated into  $\Lambda$  disjoint

groups  $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_\Lambda$ , where the  $g$ -th group, which consists of  $B = \frac{K}{\Lambda}$  users<sup>1</sup>, is given by  $\mathcal{D}_g \triangleq \{b\Lambda + g\}_{b=0}^{B-1} \subseteq [K]$ . The  $\vartheta$ -th user of this  $g$ -th group is denoted by  $U_{g,\vartheta}$ .

At this point, all the users belonging to the same group are assigned the same cache state and thus proceed to cache *identical* content. In particular, for those in the  $g$ -th group, this content takes the form  $\mathcal{Z}_{\mathcal{G}_g} = \{W_n^T : \mathcal{T} \ni g, \forall n \in [N]\}$ . This grouping as well as the entire placement phase, are naturally done before the users' requests take place, and of course well before the channel states are known to the BS.

### Delivery Phase

This phase starts when each user  $\kappa \in [K]$  simultaneously asks for its intended file, denoted here by  $W_{d_\kappa}$ ,  $d_\kappa \in [N]$ . The BS selects  $Q$  users from each group, where  $Q \leq B$  is a variable that will be optimized afterwards and which is the equivalent of the multiplexing gain. By doing so, the BS decides to first 'encode' over the first  $\Lambda Q$  users, and to repeat the encoding process  $B/Q$  times<sup>2</sup>. To deliver to the  $\Lambda Q$  users, the transmitter employs  $\binom{\Lambda}{\Lambda\gamma+1}$  sequential transmission stages. During each such stage, the BS simultaneously serves a unique set  $\Psi$  of  $|\Psi| = \Lambda\gamma + 1$  groups, corresponding to a total of  $Q(\Lambda\gamma + 1)$  users served at a time (i.e., per stage). At the end of the  $\binom{\Lambda}{\Lambda\gamma+1}$  transmission stages, all the  $\Lambda Q$  users obtain their intended files. By repeating this process  $\lceil \frac{B}{Q} \rceil$  times, all the  $K$  users obtain their intended files. As suggested above, the factor  $G \triangleq \Lambda\gamma + 1$  describes the number of user groups that are simultaneously served. Another crucial parameter includes the multiplexing gain  $Q$  which, unlike in [12], will be here subject to optimization.

For example, let us consider a setting with  $G = 3$ ,  $B = 4$ , and  $\Lambda = 40$ , and a choice of  $Q = 2$ . The delivery will be split into  $\frac{B}{Q} = 2$  encoding processes, and each process is split into  $\binom{\Lambda}{G}$  so-called stages. Each stage will involve the transmission to a different set (triplet in this case) of cache groups  $\Psi' \subseteq [\Lambda]$  where  $|\Psi'| = G = 3$ . In each such stage, the BS serves  $Q = 2$  users from each of the above three cache groups, which allows for serving  $GQ = 6$  users at a time. For example, the first stage can correspond to the set  $\Psi = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$ . The difference between the two processes is that in the first delivery process, the BS serves the first  $Q = 2$  users in each cache-group, while in the second process, the BS serves the last  $Q = 2$  users in each cache-group (cf. Fig. 3.1). We refer to the users currently served in a delivery process as active users and to all the other users as passive users in Fig. 3.1.

Let us now focus on a single transmission stage. As mentioned above, at each such stage, we pick a set  $\Psi \subseteq \Lambda$  of  $G = \Lambda\gamma + 1$  groups that will be served simultaneously. From

---

<sup>1</sup>For clarity of exposition, and without limiting the scope of the results, we will consider  $K$  to be a multiple of  $\Lambda$ . The general case can be readily handled (cf. [98]), and in Section 3.4 we provide a related example.

<sup>2</sup>To clarify, what the above says is the following. If there are, e.g.,  $B = 2Q$  users per group and thus  $K = 2\Lambda Q$  users in total, then the algorithm that we describe here will be first applied to the first  $\Lambda Q$  users, and then, after this delivery is done, the same algorithm will apply to the remaining  $\Lambda Q$  users, thus eventually satisfying all  $K$  users. Also note that a small amount of additional subpacketization can easily resolve the case where  $B/Q$  may not be an integer.

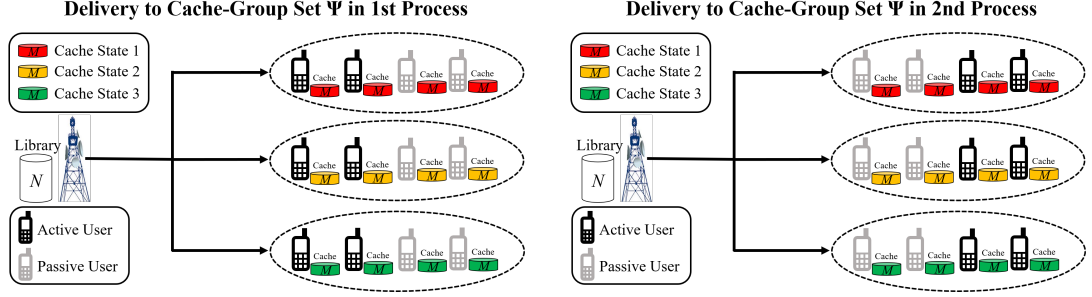


Figure 3.1: An example of vector coded caching with  $G = 3$ ,  $\Lambda = 40$ ,  $B = 4$  and  $Q = 2$ .

within these chosen groups, we will serve  $Q \leq B$  users per group. In particular, for each user  $U_{\psi,\vartheta}$  of some group  $\psi \in \Psi$ , this stage will deliver all subfiles<sup>3</sup>  $s_{\psi,\vartheta}$  by transmitting

$$\mathbf{x}_{\Psi} = \frac{1}{\sqrt{G}} \sum_{\psi \in \Psi} \rho_{\psi} \sum_{\vartheta=1}^Q \mathbf{v}_{\psi,\vartheta} s_{\psi,\vartheta}, \quad (3.2)$$

where  $\mathbf{v}_{\psi,\vartheta} \in \mathbb{C}^{L \times 1}$  denotes the precoder applied to the subfile intended by user  $U_{\psi,\vartheta}$ , and where  $\rho_{\psi}$  denotes the power normalization factor for group  $\psi \in \Psi$ , applied under a total power constraint  $P_t$ . Upon defining  $\mathbf{V}_{\psi} \in \mathbb{C}^{L \times Q}$  as  $\mathbf{V}_{\psi} \triangleq [\mathbf{v}_{\psi,1} | \dots | \mathbf{v}_{\psi,Q}]$  and  $\mathbf{s}_{\psi} \in \mathbb{C}^Q$  as  $\mathbf{s}_{\psi} \triangleq [s_{\psi,1}, \dots, s_{\psi,Q}]^T$ , the above takes the simple form

$$\mathbf{x}_{\Psi} = \frac{1}{\sqrt{G}} \sum_{\psi \in \Psi} \rho_{\psi} \mathbf{V}_{\psi} \mathbf{s}_{\psi}. \quad (3.3)$$

**Remark 3.1.** *It is easy to see that the described scheme simply involves a carefully selected linear combination of  $G$  linear-precoding vectors that are now to be sent simultaneously. It is also easy to see that the above scheme also incorporates the traditional cacheless downlink scenario corresponding to  $\gamma = 0$  which itself corresponds to  $G = |\Psi| = 1$ . In such case, the transmit signal expression reverts to the simpler common expression  $\mathbf{x} = \rho \mathbf{V} \mathbf{s}$ .*

For decoding to work, the subfiles must be chosen carefully. This choice follows the principles of coded caching, and in particular of vector coded caching. Thus, when considering the transmission stage which serves the  $G = \Lambda\gamma + 1$  groups in  $\Psi$ , the subfile transmitted to user  $U_{\psi,\vartheta}$  is here selected to be  $W_{d_{\psi,\vartheta}}^{\Psi \setminus \{\psi\}}$ , simply because this subfile is stored in the cache of each user of every other group in  $\Psi$  except  $\psi$ . Because of this structure, the users of a particular group can remove the inter-group interference from the other  $\Lambda\gamma$  groups by using their cached content. On the other hand, following the principles of vector coded caching, the intra-group interference is handled with linear precoding that ‘separates’ the signals of the users from the same group. Naturally one can imagine that cache-aided removal of interference as well as ‘nulling out’ of interference, both require knowledge of the composite precoder-channel coefficients (cf. (3.5)). These so-called *composite CSI* costs will be explicitly accounted for in our analysis. We proceed to elaborate on the precoders and the transmissions.

<sup>3</sup>In a slight abuse of notation, we use the term “subfile” to refer both to the actual subfile generated after file-splitting, as well as to the corresponding complex-valued information symbol  $s_{\psi,\vartheta}$ .

### 3.1.3 Vector Coded Caching for the Physical Layer

We now emphasize on the physical layer details of the communication scheme. Our description will focus on the transmission that serves a specific set  $\Psi$  of user-groups. First let us recall that  $\mathbf{V}_\psi \in \mathbb{C}^{L \times Q}$  denotes the precoding matrix for the symbols of users in group  $\psi \in \Psi$ . We note that, as is common, our analysis will assume Gaussian signaling. Then let us note that for an average power constraint  $P_t$ , the power normalization factor  $\rho_\psi$  from (3.3), takes the form  $\rho_\psi = \sqrt{\frac{P_t}{\mathbb{E}\{\mathbf{s}_\psi^H \mathbf{V}_\psi^H \mathbf{V}_\psi \mathbf{s}_\psi\}}} = \sqrt{\frac{P_t}{\mathbb{E}\{\text{Tr}\{\mathbf{V}_\psi^H \mathbf{V}_\psi\}}}}$ . Then the subsequent corresponding received signal at user  $U_{\psi,k}$  (i.e., at the  $k$ -th user of group  $\psi \in \Psi$ ), will take the form

$$y_{\psi,k} = \mathbf{h}_{\psi,k}^T \mathbf{x}_\Psi + z_{\psi,k} = \frac{\mathbf{h}_{\psi,k}^T}{\sqrt{G}} \rho_\psi \mathbf{V}_\psi \mathbf{s}_\psi + \underbrace{\frac{\mathbf{h}_{\psi,k}^T}{\sqrt{G}} \sum_{\phi \in \Psi, \phi \neq \psi} \rho_\phi \mathbf{V}_\phi \mathbf{s}_\phi}_{\text{inter-group interference}} + z_{\psi,k}. \quad (3.4)$$

As previously mentioned, the inter-group interference<sup>4</sup> experienced by user  $U_{\psi,k}$ , can be removed from  $y_{\psi,k}$  by exploiting that same user's cached content and that user's composite CSI  $\{\mathbf{h}_{\psi,k}^T \mathbf{v}_{\phi,k'} \rho_\phi\}_{\phi \in \{\Psi \setminus \psi\}, k' \in [Q]}$ . Then, after the cache-aided removal of this inter-group interference, the equivalent received signal at  $U_{\psi,k}$  is given by

$$y'_{\psi,k} = \frac{\rho_\psi}{\sqrt{G}} \mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,k} s_{\psi,k} + \underbrace{\frac{\rho_\psi}{\sqrt{G}} \sum_{\vartheta=1, \vartheta \neq k}^Q \mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,\vartheta} s_{\psi,\vartheta}}_{\text{intra-group interference}} + z_{\psi,k}. \quad (3.5)$$

Consequently, the corresponding SINR for information decoding at  $U_{\psi,k}$ , is given by

$$\text{SINR}_{\psi,k} = \frac{\frac{\rho_\psi^2}{G} |\mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,k}|^2}{1 + \frac{\rho_\psi^2}{G} \sum_{\vartheta=1, \vartheta \neq k}^Q |\mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,\vartheta}|^2}. \quad (3.6)$$

On the other hand, in the cacheless case of  $\gamma = 0$ , the received signal  $y_k = \rho \mathbf{h}_k^T \mathbf{v}_k s_k + \rho \sum_{\vartheta=1, \vartheta \neq k}^Q \mathbf{h}_k^T \mathbf{v}_{\vartheta} s_{\vartheta} + z_k$  at some user  $k$  naturally carries no inter-group interference (as there are no other groups to simultaneously serve), and the SINR takes the standard form  $\text{SINR}_k = \frac{\rho^2 |\mathbf{h}_k^T \mathbf{v}_k|^2}{1 + \rho^2 \sum_{\vartheta=1, \vartheta \neq k}^Q |\mathbf{h}_k^T \mathbf{v}_{\vartheta}|^2}$ . Therefore, the instantaneous rate

$$R_{\psi,k} = \ln(1 + \text{SINR}_{\psi,k}) \quad \text{nats/s/Hz} \quad (3.7)$$

for user  $U_{\psi,k}$  is obtained by evaluating the above, at the SINR value in (3.6).

We will consider the MF, ZF and RZF linear precoding schemes, selected here for being very common, simple, as well as competitive in terms of rate performance [100, 101]. As is known, the corresponding precoding matrices  $\mathbf{V}_\psi$  take the form:

$$\mathbf{V}_\psi = \begin{cases} \mathbf{H}_\psi^H, & \text{MF Precoder} \\ \mathbf{H}_\psi^H (\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}, & \text{ZF Precoder} \\ \mathbf{H}_\psi^H (\mathbf{H}_\psi \mathbf{H}_\psi^H + \alpha \mathbf{I}_Q)^{-1}, & \text{RZF Precoder,} \end{cases} \quad (3.8)$$

<sup>4</sup>As a reminder, the term inter-group interference refers to the received signal component whose power is due to the information meant for users originating from other groups.

where  $\mathbf{H}_\psi \triangleq [\mathbf{h}_{\psi,1} | \mathbf{h}_{\psi,2} | \cdots | \mathbf{h}_{\psi,Q}]^T \in \mathbb{C}^{Q \times L}$  denotes the channel matrix for the channel from the BS to the  $Q$  chosen users belonging to group  $\psi \in \Psi$ , and where  $\alpha$  is the regularization factor of the RZF precoder [100]. It is worth recalling that the RZF precoder reverts to the ZF precoder when  $\alpha = 0$ , and to the MF precoder when  $\alpha \rightarrow \infty$ , and also that  $Q$  is bounded above by  $B$  and, in the case of the ZF/RZF precoding, it is also bounded as  $Q \leq L$ . For simplicity we assume that  $\alpha = L/P_t$ , which is a commonly used assumption throughout the literature [62, 100, 102].

We will henceforth use the term  $(G, Q)$ -vector coded caching, to refer to the vector coded caching scheme when it serves  $G$  groups with  $Q$  users per group. We will also use the term *MF-based  $(G, Q)$ -vector coded caching* to refer to the same scheme when the underlying precoder is MF, and similarly we will use ZF-based or RZF-based  $(G, Q)$ -vector coded caching, for the other two precoders. Let us now formally define some important metrics of interest.

**Definition 3.** (Average sum-rate and effective sum-rate). *For a  $(G, Q)$ -vector coded caching scheme, its average sum-rate is denoted by  $\bar{R}(G, Q)$  and is defined as the total data-transmission rate (before accounting for CSI costs) summed over the  $GQ$  simultaneously served users, and averaged over the fading. Similarly, the effective average sum-rate  $\bar{\mathcal{R}}(G, Q)$  will represent the corresponding average rate after through all CSI costs are duly accounted for.*

**Definition 4.** (Effective gain over MISO). *For a given set of  $L$  and SNR resources, and a fixed underlying precoder class, the effective gain, after accounting for CSI costs, of the  $(G, Q)$ -vector coded caching over the cacheless scenario (corresponding to  $G = 1$ , and an operating multiplexing gain  $Q'$ ), will be denoted as  $\mathcal{G}(G, Q; 1, Q') \triangleq \frac{\bar{\mathcal{R}}(G, Q)}{\bar{\mathcal{R}}(1, Q')}$  in the form of the ratios of the effective rates.*

## 3.2 Analysis of the Average Rate and of the Effective Gain over MISO

In this section, we analyze the average sum-rates and the corresponding effective rates achieved by the cache-aided downlink schemes of Section 3.1.2 for the MF, ZF and RZF linear precoders of interest. After doing so, we also report the effective gains offered by these  $(G, Q)$ -vector coded caching schemes, over the  $(G = 1, Q')$  cacheless equivalents.

We will henceforth consider the ratio  $c \triangleq Q/L$ , while we will often use the notation  $c' \triangleq Q'/L$  when referring explicitly to the cacheless equivalent. The two ratios can be chosen independently. When applying large matrix analysis, we will be assuming a fixed  $c > 0$  and a fixed  $c' > 0$ .

### 3.2.1 MF Precoding

To derive the average sum-rate of vector coded caching with MF precoding, we first recall that the elements of  $\mathbf{H}_\psi$  are i.i.d. Gaussian random variables with zero mean and unit variance, which implies that  $\mathbb{E} \left\{ \text{Tr} \left\{ \mathbf{H}_\psi \mathbf{H}_\psi^H \right\} \right\} = LQ$  (cf. [103]), which then implies that the power normalization factor  $\rho_\psi$  takes the form  $\rho_\psi = \sqrt{\frac{P_t}{\mathbb{E} \left\{ \text{Tr} \left\{ \mathbf{H}_\psi \mathbf{H}_\psi^H \right\} \right\}}} = \sqrt{\frac{P_t}{QL}}$

(cf. [64]). This in turn yields (cf. (3.8), (3.3)) a transmitted signal of the form

$$\mathbf{x}_\Psi = \sqrt{\frac{P_t}{GQL}} \sum_{\psi \in \Psi} \mathbf{H}_\psi^H \mathbf{s}_\psi = \sqrt{\frac{P_t}{GQL}} \sum_{\psi \in \Psi} \sum_{\vartheta=1}^Q \mathbf{h}_{\psi,\vartheta}^* s_{\psi,\vartheta}. \quad (3.9)$$

The corresponding average sum-rate is presented below. We note that Theorem 3.1 focuses on the case of  $Q > 1$ . However, the analysis for  $Q = 1$  is straightforward and follows the same large-matrix properties and principles. The only difference is that for the single-stream scenario, one can deviate from the current scheme, and employ XORs rather than linear combinations over the complex numbers. This is not covered in our work here.

**Theorem 3.1.** *For any given  $P_t$  and  $c = Q/L$ , the average sum-rate  $\bar{R}^{\text{MF}}$  of the MF-based  $(G, cL)$ -vector coded caching scheme in the large  $L$  regime satisfies*

$$\bar{R}^{\text{MF}}(G, cL) \doteq c GL \ln \left( 1 + \frac{1}{c} \frac{P_t}{P_t + G} \right). \quad (3.10)$$

*Proof.* The proof can be found in Appendix B.1.  $\square$

The following directly distills the above result to the cacheless case.<sup>5</sup>

**Corollary 3.1.** *In the limit of large  $L$ , and for any fixed  $P_t$  and  $c'$ , the average sum-rate of the (traditional, cacheless) MF-based MISO BC with  $c'L$  streams satisfies*

$$\bar{R}^{\text{MF}}(1, c'L) \doteq c' L \ln \left( 1 + \frac{1}{c'} \frac{P_t}{P_t + 1} \right). \quad (3.11)$$

### 3.2.2 ZF Precoding

Moving now to the case of ZF-based vector coded caching, and focusing again on a set of groups  $\Psi$  and on the transmission stage corresponding to some group  $\psi \in \Psi$ , the power control factor takes the form  $\rho_\psi^2 = \frac{P_t}{\mathbb{E}\{\text{Tr}\{(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\}\}}$ , while the transmitted signal from (3.3) becomes

$$\mathbf{x}_\Psi = \frac{1}{\sqrt{G}} \sum_{\psi \in \Psi} \rho_\psi \mathbf{H}_\psi^H (\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1} \mathbf{s}_\psi. \quad (3.12)$$

This in turn yields a received signal at user  $U_{\psi,k}$  which — after the cache-aided removal of the inter-group interference (cf. (3.4)) — takes the form

$$y'_{\psi,k} = \frac{1}{\sqrt{G}} \rho_\psi \mathbf{h}_{\psi,k}^T \mathbf{H}_\psi^H (\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1} \mathbf{s}_\psi + z_{\psi,k}$$

<sup>5</sup>It is worth noting that while there have been various works (cf. [61–64]) analyzing the MF sum-rate in traditional massive MIMO systems, the result derived in this work here entails less assumptions. For example, focusing on the large- $L$  regime, the result in [61] directly assumes a tight Jensen’s bound, while the result in [63] is under a so-called “near deterministic” assumption in low/high SNRs. On the other hand, our method here draws from the uplink analysis in [65], and only employs a large- $L$  assumption to derive the exact asymptotic optimality for any value of SNR.

$$= \frac{1}{\sqrt{G}} \rho_\psi (\mathbf{1}_k^T \mathbf{H}_\psi) \mathbf{H}_\psi^H (\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1} \mathbf{s}_\psi + z_{\psi,k} = \frac{1}{\sqrt{G}} \rho_\psi s_{\psi,k} + z_{\psi,k}, \quad (3.13)$$

where  $\mathbf{1}_k \in \mathbb{C}^{Q \times 1}$  denotes the vector whose components are all zero except for the  $k$ -th element, which equals 1. After then considering that all intra-group interference is canceled by means of ZF precoding, we can write the SINR at user  $U_{\psi,k}$  as

$$\text{SINR}_{\psi,k}^{\text{ZF}} = \frac{P_t}{G \mathbb{E}\{\text{Tr}\{(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\}\}}. \quad (3.14)$$

With this in place, we proceed with the following theorem.

**Theorem 3.2.** For  $c = \frac{Q}{L} \in (0, 1)$ , the average sum-rate  $\bar{R}_{\text{sum}}^{\text{ZF}}$  of the ZF-based  $(G, Q)$ -vector coded caching scheme, takes the form

$$\bar{R}^{\text{ZF}}(G, Q) = QG \ln \left( 1 + \frac{P_t}{G} \left( \frac{1}{c} - 1 \right) \right). \quad (3.15)$$

*Proof.* Directly from [104], and from the fact that  $\mathbf{H}_\psi \mathbf{H}_\psi^H$  is a Wishart matrix with  $L$  degrees of freedom, we know that  $\mathbb{E}\{\text{Tr}\{(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\}\} = \frac{Q}{L-Q}$  for  $L > Q$ . Naturally,  $\text{SINR}_{\psi,k}^{\text{ZF}}$  is deterministic and constant across all simultaneously served users. By summing the average rate of each of the  $GQ$  served users, we obtain (3.15).  $\square$

### 3.2.3 RZF Precoding

We finally consider our third precoder, and do so in the asymptotic regime of large  $L$  and fixed  $c$ . We first note that the received signal at  $U_{\psi,k}$  — after cache-aided removal of the inter-group interference — takes the form

$$y'_{\psi,k} = \frac{\rho_\psi}{\sqrt{G}} \sum_{\vartheta=1}^Q \mathbf{h}_{\psi,k}^T (\alpha \mathbf{I}_L + \mathbf{H}_\psi^H \mathbf{H}_\psi)^{-1} \mathbf{h}_{\psi,\vartheta}^* s_{\psi,\vartheta} + z_{\psi,k}. \quad (3.16)$$

For  $\mathbf{H}_{\psi,-k}$  denoting the matrix resulting from  $\mathbf{H}_\psi$  after removing its  $k$ -th row, we can define

$$A_{\psi,k} \triangleq \mathbf{h}_{\psi,k}^T (\alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k})^{-1} \mathbf{h}_{\psi,k}^*, \quad (3.17)$$

$$B_{\psi,k} \triangleq \mathbf{h}_{\psi,k}^T (\alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k})^{-1} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} (\alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k})^{-1} \mathbf{h}_{\psi,k}^*. \quad (3.18)$$

With these in place, we can now derive the SINR at user  $U_{\psi,k}$  to be

$$\text{SINR}_{\psi,k}^{\text{RZF}} = \frac{A_{\psi,k}^2 \frac{\rho_\psi^2}{G}}{(1 + A_{\psi,k})^2 + \frac{\rho_\psi^2}{G} B_{\psi,k}}, \quad (3.19)$$

where the proof of (3.19) is relegated to Appendix B.2.

We can now present the asymptotic deterministic equivalent of the sum-rate of our proposed scheme when RZF is applied. We recall that in the limit of large  $L$ , the deterministic value  $\hat{X}$  represents the *asymptotic deterministic equivalent* of  $X$  if  $X \xrightarrow{a.s.} \hat{X}$ .

**Theorem 3.3.** *In the large- $L$  regime with fixed  $c = Q/L$ , the average sum-rate  $\bar{R}^{\text{RZF}}$  of RZF-based  $(G, Q)$ -vector coded caching, takes the form*

$$\bar{R}^{\text{RZF}}(G, Q) \doteq \hat{R}^{\text{RZF}}(G, cL) \triangleq cGL \ln \left( 1 + \frac{a_{\psi,k}^2 p_{\psi}^2 / G}{(1 + a_{\psi,k})^2 + P_t / G} \right), \quad (3.20)$$

where  $\hat{R}^{\text{RZF}}$  is the deterministic equivalent of  $\bar{R}^{\text{RZF}}$ ,<sup>6</sup> and where

$$a_{\psi,k} \triangleq \frac{1}{2} \left[ \sqrt{(1-c)^2 P_t^2 + 2(1+c)P_t + 1} + (1-c)P_t - 1 \right], \quad (3.21)$$

$$p_{\psi}^2 \triangleq \frac{P_t}{a_{\psi,k} - \frac{P_t}{2} \left( \frac{P_t(c-1)^2 + c + 1}{\sqrt{P_t^2(c-1)^2 + 2(c+1)P_t + 1}} + 1 - c \right)}. \quad (3.22)$$

*Proof.* The proof is based on the derivation of the asymptotic deterministic equivalent of the SINR, and it is presented in Appendix B.2.  $\square$

### 3.2.4 Accounting for the CSI Costs

To account for the cost of CSI acquisition under TDD, we consider a basic CSI-acquisition effort where at the beginning of each transmission stage, the  $GQ$  served users send uplink orthogonal pilot symbols, from which the BS can estimate the downlink channel matrix, under the assumption of channel reciprocity. Then the CSI-acquisition process engages downlink training, of similar complexity, in order to communicate the composite CSI that here allows our receivers to perform cache-aided cancellation of the inter-group interference (cf. (3.4)) from their signal. This acquisition process for gathering composite CSI, with the same aforementioned complexity per served user, is standard in a variety of traditional communications techniques such as SIC-based approaches. For additional details, please refer to [105]. To account for this CSI-acquisition overhead, we directly extend the commonly-used approach in [106–109], that easily allows us to calculate the effective average sum-rate (cf. Definition 3) for each precoder  $i \in \{\text{MF}, \text{ZF}, \text{RZF}\}$ , to be

$$\bar{\mathcal{R}}^i = \left( 1 - \frac{\beta_{\text{tot}} G Q}{T_c W_c} \right) \bar{R}^i = (1 - c \zeta_{G,Q}) \bar{R}^i, \quad (3.23)$$

where  $\beta_{\text{tot}}$  is the number of resources per user and per block used for pilot transmission,  $\bar{R}^i$  is the previously calculated average sum-rate before accounting for CSI costs, where  $T_c$  and  $W_c$  are the coherence time and coherence bandwidth, respectively, and where  $\zeta_{G,Q} \triangleq \frac{\beta_{\text{tot}} G L}{T_c W_c}$ . For completeness we report the effective rates in the following corollary. The proof is direct as it merely involves applying (3.23) in the expressions from Theorems 3.1–3.3. We recall that  $a_{\psi,k}$  and  $p_{\psi}$  are defined in Theorem 3.3.

<sup>6</sup>This entails a small abuse of terminology, as it is  $\hat{R}^{\text{RZF}}/L$  that is the deterministic equivalent of  $\bar{R}^{\text{RZF}}/L$ .

**Corollary 3.2.** *The effective rates of the proposed vector coded caching schemes under MF, ZF and RZF precoding, respectively take the form*

$$\bar{\mathcal{R}}^{MF}(G, Q) \doteq (1 - c\zeta_{G,Q}) c GL \ln \left( 1 + \frac{1}{c} \frac{P_t}{P_t + G} \right), \quad (3.24)$$

$$\bar{\mathcal{R}}^{ZF}(G, Q) = (1 - c\zeta_{G,Q}) QG \ln \left( 1 + \frac{P_t}{G} \left( \frac{1}{c} - 1 \right) \right), \quad (3.25)$$

$$\bar{\mathcal{R}}^{RZF}(G, Q) \doteq (1 - c\zeta_{G,Q}) c GL \ln \left( 1 + \frac{a_{\psi,k}^2 p_{\psi}^2 / G}{(1 + a_{\psi,k})^2 + P_t/G} \right). \quad (3.26)$$

### 3.2.5 Effective Gains over Cacheless MISO Systems

At this point, with Theorems 3.1, 3.2, 3.3 in place, and in conjunction with Corollary 3.2, we can directly report the effective gains over cacheless MISO. For each of the three precoder classes, MF, ZF, and RZF, and for a fixed set of antenna and SNR resources, we will be reporting the effective gain  $\mathcal{G}(G, Q; 1, Q') = \frac{\bar{\mathcal{R}}(G, Q)}{\bar{\mathcal{R}}(1, Q')}$  (cf. Definition 4) of the  $(G, Q)$ -vector coded caching schemes, over the cacheless scenario ( $G = 1$ ) with some chosen number of streams  $Q'$ . These effective gains are collected together in the following corollary.

**Corollary 3.3.** *The effective gains of the proposed vector coded caching schemes under MF, ZF and RZF precoding, respectively take the form*

$$\mathcal{G}_{MF}(G, Q; 1, Q') \triangleq \frac{\bar{\mathcal{R}}^{MF}(G, Q)}{\bar{\mathcal{R}}^{MF}(1, Q')} \doteq \xi_{\text{csi}} \frac{GQ \ln \left( 1 + \frac{L}{Q} \frac{P_t}{P_t + G} \right)}{Q' \ln \left( 1 + \frac{L}{Q'} \frac{P_t}{P_t + 1} \right)}, \quad (3.27)$$

$$\mathcal{G}_{ZF}(G, Q; 1, Q') \triangleq \frac{\bar{\mathcal{R}}^{ZF}(G, Q)}{\bar{\mathcal{R}}^{ZF}(1, Q')} = \xi_{\text{csi}} \frac{GQ \ln \left( 1 + \frac{P_t}{G} \left( \frac{L}{Q} - 1 \right) \right)}{Q' \ln \left( 1 + P_t \left( \frac{L}{Q'} - 1 \right) \right)}, \quad (3.28)$$

$$\mathcal{G}_{RZF}(G, Q; 1, Q') \triangleq \frac{\bar{\mathcal{R}}^{RZF}(G, Q)}{\bar{\mathcal{R}}^{RZF}(1, Q')} \xrightarrow{a.s.} \xi_{\text{csi}} \frac{\mathring{R}^{RZF}(G, cL)}{\mathring{R}^{RZF}(1, c'L)}, \quad (3.29)$$

where  $\mathring{R}^{RZF}(\cdot, \cdot)$  is defined in (3.20), and where  $\xi_{\text{csi}} \triangleq \frac{(L - Q\zeta_{G,Q})}{(L - Q'\zeta_{1,Q'})}$ .

## 3.3 Optimizing Physical Layer Vector Coded Caching

Theorems 3.1-3.3 reveal the important dependence of vector coded caching on the number of streams,  $Q$ , that we choose to activate. This dependence strikes at the very core of the problems stemming from power-splitting and CSI overheads. Indeed, while an increased  $Q \leq L$  allows for a higher DoF at lower subpacketization, this increase in the number of streams may not be beneficial in practice as it entails less power per stream as well as more CSI to be communicated.

For this reason, we here proceed to analytically optimize our schemes over the choices of  $Q$ . This optimization is tractable partly due to the simplicity of the achievable-rate expressions derived in the previous theorems<sup>7</sup>, and while some of these expressions involve asymptotic approximations, they will, as we will verify numerically, be very precise (see for example Fig. 3.2). Our analysis of the optimal  $c^*$  will assume a variable  $c = Q/L$  that is continuous and unbounded. As noted before, the optimization takes into account the impact of CSI acquisition under TDD.

Let us first focus on deriving the optimal  $c^*$  for MF, where  $c \in (0, \infty)$  and  $\Omega \triangleq \frac{P_t}{P_t + G}$ .

**Theorem 3.4.** *In the MF-based  $(G, Q)$ -vector coded caching with non-negligible CSI costs, the optimal  $c^*$  that maximizes  $\bar{\mathcal{R}}^{MF}$  in the asymptotic sense, is given by the solution to the following:*

$$(1 - 2\zeta_{G,Q}c^*) \ln \left( 1 + \frac{\Omega}{c^*} \right) - \frac{\Omega(1 - \zeta_{G,Q}c^*)}{\Omega + c^*} = 0. \quad (3.30)$$

*Proof.* See Appendix B.3. □

Next, we consider ZF-based cache-aided precoding, for which we have the following.

**Theorem 3.5.** *In the ZF-based  $(G, Q)$ -vector coded caching with non-negligible CSI costs, the optimal  $c^*$  that maximizes  $\bar{\mathcal{R}}^{ZF}$ , is given by the solution to the following equation:*

$$(1 - 2\zeta_{G,Q}c^*) \ln \left( 1 + \frac{P_t}{G} \left( \frac{1}{c^*} - 1 \right) \right) - \frac{(1 - \zeta_{G,Q}c^*)P_t/G}{(1 - P_t/G)c^* + P_t/G} = 0. \quad (3.31)$$

*Proof.* See Appendix B.4. □

**Remark 3.2.** *As  $P_t \rightarrow \infty$ , we can write (B.28) as  $\frac{\partial \bar{\mathcal{R}}^{ZF}}{\partial c} = GL \left[ \ln \left( \frac{P_t}{G} \right) + \ln \left( \frac{1-c}{1-c} \right) - \frac{1}{1-c} \right] + o(1)$ , where  $\lim_{P_t \rightarrow \infty} o(1) = 0$ . Therefore, in the high-SNR regime and without taking CSI costs into account, the optimal value of  $c$  that maximizes  $\bar{\mathcal{R}}^{ZF}$  (and thus  $\bar{\mathcal{R}}^{RZF}$ , since both converge at high-SNR) is given by  $c^* = \left( 1 + \frac{1}{\mathcal{W}(P_t/(eG))} \right)^{-1}$ , upon omitting an  $o(1)$  additive term, and upon using  $\mathcal{W}(\cdot)$  to denote the Lambert  $W$ -Function. This expression can serve as a good approximation in those moderate-to-high SNR scenarios where the dimensionality of the problem implies a relatively small CSI cost. As one can see, as the SNR becomes very large, the above  $c^*$  converges, as is known, to 1, corresponding to  $Q \approx L$ .*

Having derived the above optimal  $c^*$ , we can now consider the ratio

$$\mathcal{G}^* \triangleq \frac{\max_{Q \in \mathbb{Z}^+} \bar{\mathcal{R}}^i(G, Q)}{\max_{Q' \in \mathbb{Z}^+} \bar{\mathcal{R}}^i(G = 1, Q')}, \quad (3.32)$$

which describes the performance boost due to caching, over (independently) optimized downlink cacheless systems, after accounting for CSI costs. These gains  $\mathcal{G}_{MF}^*$ ,  $\mathcal{G}_{ZF}^*$ ,  $\mathcal{G}_{RZF}^*$  are

<sup>7</sup>The derivation of the optimal point for RZF is omitted due to the fact that, although we can obtain the derivative of the sum-rate, the equation to find the optimal  $Q$  provides little insight and we would need to obtain the solution numerically (cf. Appendix III in [110]).

Table 3.1: Derived Theorems (Thms.) and Corollaries (Cors.) in Chapter 3

Thm. 3.1	Thm. 3.2	Thm. 3.3	Thm. 3.4	Thm. 3.5	Cor. 3.1	Cor. 3.2	Cor. 3.3
Average sum-rate MF	Average sum-rate ZF	Average sum-rate RZF	Optimal $Q$ for MF	Optimal $Q$ for ZF	Average sum-rate in cacheless MF	Effective rates in MF/ZF/RZF	Effective gains in MF/ZF/RZF

reported for the three precoders of interest. As one would expect, this comparison is done under a fixed set of SNR and antenna resources. The transition from the continuous  $c$  to the operating  $Q$ , will follow by simply considering  $Q^* = \arg \max_{Q \in \{\lfloor c^* L \rfloor, \lfloor c^* L \rfloor + 1\}} \{\mathcal{R}(Q)\}$ , where  $\lfloor \cdot \rfloor$  denotes the nearest integer less than or equal to the argument.

### 3.4 Numerical Results

We proceed to numerically demonstrate the achieved effective rates as well as the effective gains that an optimized vector coded caching scheme provides over the independently optimized cacheless downlink solution<sup>8</sup>. We note that the simulated results employ no approximations (for example, the corresponding SINR is taken directly from (3.6)). For ease of exposition, we list in Table 3.1 the derived theorems and corollaries.

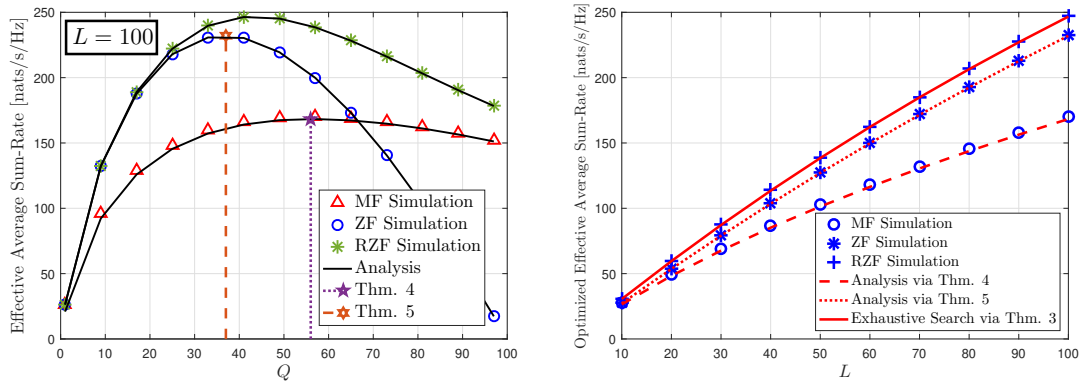
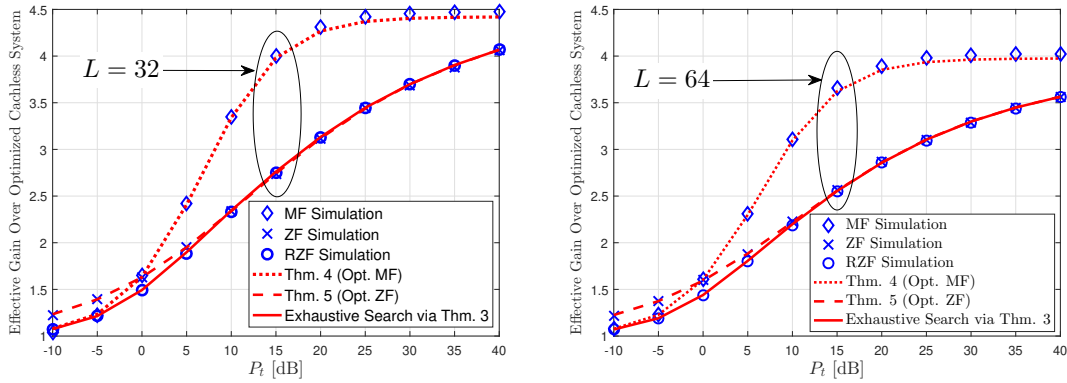
The following figures build on the analysis of the effective sum rates and effective gains of Section 3.2, as well as on the analysis of the optimized gains of Section 3.3. These figures incorporate the CSI costs in the realistic scenario of having  $\beta_{\text{tot}} = 10$ ,  $T_c = 0.04$  seconds and  $W_c = 300$  kHz (cf. (3.23)), which captures the common scenario of low-mobility users consuming videos<sup>9</sup>. We note that  $\beta_{\text{tot}} = 10$  is high enough to allow us to neglect the impact of CSI estimation noise [108, 109]. Fig. 3.2 (left) describes the effective rate of the different cache-aided schemes, for different values of  $Q$ . The plot highlights the tightness of the results of Theorems 3.1–3.3 (after accounting for CSI costs: see Corollary 3.2), where we see that indeed the derived asymptotically-approximate expressions have no discernible distance from the actual (simulated) performance. The vertical lines indicate the optimal  $Q$  derived in Theorems 3.4–3.5. These optimal points indeed match the actual maximum point of the curves. Fig. 3.2 (right) extends this illustration of the tightness of the results, to Theorems 3.4–3.5, by illustrating the optimized (over all  $Q$  choices) effective rate performance of the three precoders, comparing the derived results to the actual performance. We note that, for the case of RZF precoding, we represent the result of Theorem 3.3 (Corollary 3.2) by considering a  $c^*$  value that is obtained from an exhaustive search based on these derived expressions.

Fig. 3.3 focuses on the effective gains over optimized cacheless downlink systems. As before, the theoretical and simulated results match fully. Here the theoretical results reflect the effective gain ratio  $\mathcal{G}^*$  in (3.32), where the derived effective-rate expressions are from Corollary 3.2 (and the corresponding Theorems 3.1–3.3), and where the optimized  $c^*$  are directly from<sup>10</sup> Theorems 3.4–3.5.

<sup>8</sup>For the convenience of annotation in the simulation figures, we omit the chapter labels of theorems, lemmas, corollaries, propositions and equations. As the numerical results are independent across different chapters, this kind of omission does not bring about any confusion.

<sup>9</sup>We note that  $\beta_{\text{tot}}$  could be decreased down to 2 at high SNR, hence reducing the CSI overhead.

<sup>10</sup>We recall that, for the RZF case, in Fig. 3.3 we numerically evaluate  $c^*$  from Theorem 3.3.


 Figure 3.2: Effective rate  $\bar{\mathcal{R}}$  and optimized effective rate for  $P_t = 10$  dB and  $G = 5$ .

 Figure 3.3: Effective gain  $\mathcal{G}^*$  over optimized cacheless system for  $L \in \{32, 64\}$  and  $G = 6$ .

It is notable that, despite the fact that Theorem 3.1 and Theorem 3.3 are obtained from asymptotic analysis, they closely characterize the real performance obtained from simulations. This is also reflected in Fig. 3.4.

Under the above realistic coherence periods and coherence bandwidths, realistic CSI costs, as well as realistic values of SNR and  $L$ , the multiplicative boosts over the achievable rates of optimized downlink systems are quite notable. For example, for 64 transmit antennas, a receiver-side SNR of 20 dB, the same  $W_c = 300$  kHz and  $T_c = 40$  ms, and under realistic file-size and cache-size constraints that allow us to assume  $G = 6$ , vector coded caching is here shown to offer a multiplicative boost of about 280% in ZF/RZF precoding and 380% over MF-based cacheless systems, whereas for the case of 32 antennas the gain elevates to 310% for ZF and to a 430% multiplicative boost in the performance of already optimized MF-based cacheless systems<sup>11</sup>. As one would expect, this same figure reveals that the gains  $\mathcal{G}^*$  grow monotonically with the SNR, and often come very close to the theoretical upper bound of  $G$ .

<sup>11</sup>In addition to the speedup factor reported here, the use of caches can also lead to additional — albeit marginal — reductions in delivery-time, complements of the so-called local caching gain, which is though of no particular interest to this study.

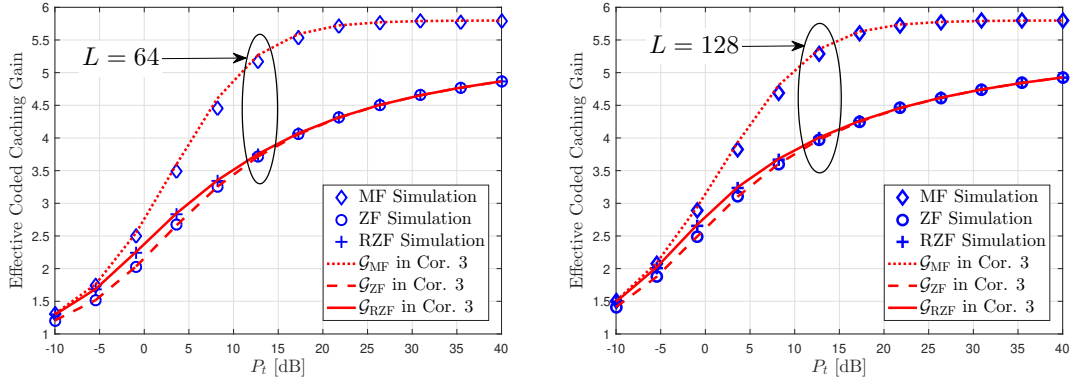


Figure 3.4: Hardening-constrained effective gain over a constrained classical downlink system.  $Q$  is fixed for both systems at  $Q = 8$ , while  $G = 6$ .

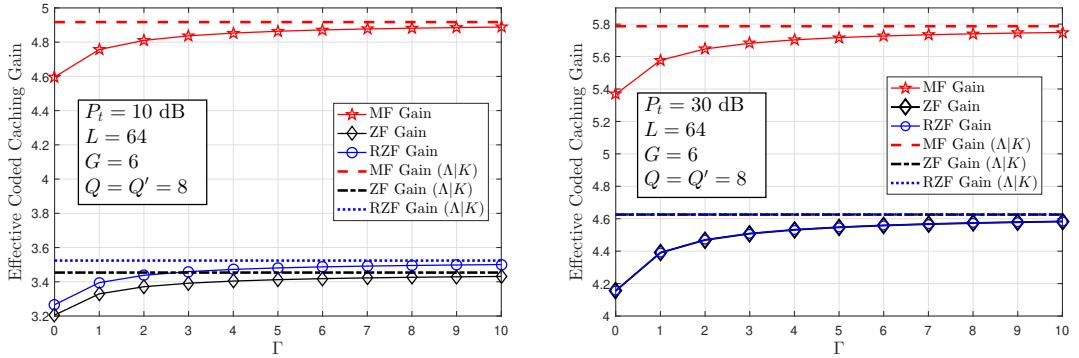


Figure 3.5: Effective gain versus  $\Gamma$  in medium SNR (10 dB) and high SNR (30 dB).

Another interesting comparison is shown in Fig. 3.4, where we ask that the cache-aided and cacheless scenarios share the same exact multiplexing gain  $Q$ . The motivation for this comparison traces back to the idea of channel hardening, which refers to the fact that as long as  $L$  is sufficiently large, and as long as  $Q/L$  is sufficiently small, the channel converges to a deterministic value, thus making CSI acquisition easier. While this paper is not about the channel hardening properties of the cache-aided downlink, this Fig. 3.4 — which plots the effective gain  $\mathcal{G}(G, Q; 1, Q) = \bar{\mathcal{R}}(G, Q)/\bar{\mathcal{R}}(1, Q)$  — offers a first indication of yet another benefit of vector coded caching, which now allows us to serve more users at a time, but do so with a controlled ratio  $Q/L$  that guarantees certain channel hardening conditions. Focusing on the case of a fixed  $Q = 8$  for both the cache-aided ( $G = 6$ ), as well as the cacheless case ( $G = 1$ ), Fig. 3.4 reveals that under the same  $W_c, T_c$  and under realistic SNR values of, for example, approximately 15 dB, the effective gains (over cacheless equivalent systems with the same  $Q/L$ ) approach 400% for the ZF-based precoders, and even go beyond 540% when using MF-based precoding. Similar gains are recorded in the larger scenario with  $L = 128$  transmit antennas.

So far, for the sake of clarity of exposition, we have considered the case where  $K$  is a multiple of  $\Lambda$ . The impact of deviating from this assumption is indeed very small. Let us briefly discuss this. Let  $\underline{B} \triangleq \lfloor K/\Lambda \rfloor$ , in which case  $K - \Lambda\underline{B}$  cache groups will have  $\underline{B} + 1$  users each, while the remaining  $\Lambda(\underline{B} + 1) - K$  cache groups will have  $\underline{B}$  users. Then for the first  $\frac{\underline{B}+1}{Q} - 1$  delivery processes, the effective gain will be the same as before, while for the remaining processes this will be slightly reduced. Let us consider the worst case where there are only  $\underline{B}$  users in each cache group in a specific cache-group set  $\Psi$ . In this case, we can have that the number of users in each cache group in  $\Psi$  is  $\Gamma Q + (Q - 1)$  where  $\Gamma = \frac{\underline{B}+1}{Q} - 1$ . The corresponding effective gain averaged over the entire  $\Gamma + 1$  delivery processes for serving the cache-group set  $\Psi$  is then

$$\mathcal{G}(G, Q; 1, Q') = \frac{1}{\Gamma + 1} \left( \Gamma \cdot \frac{\bar{\mathcal{R}}(G, Q)}{\bar{\mathcal{R}}(1, Q')} + \frac{\bar{\mathcal{R}}(G, Q - 1)}{\bar{\mathcal{R}}(1, Q')} \right), \quad (3.33)$$

where  $\bar{\mathcal{R}}(\cdot)$  was introduced in Definition 3. We illustrate this result in Fig. 3.5, which plots this effective gain in (3.33), comparing it to the corresponding gain under the assumption that  $\Lambda$  divides  $K$  (denoted by  $\Lambda|K$  in Fig. 3.5). We can easily see that the effective gain gap decreases as  $\Gamma$  increases, and eventually becomes negligible for a reasonable value of  $\Gamma$ , e.g., gap  $\leq 2\%$  for  $\Gamma = 4$  in both the medium and the high SNR regimes.

The above numerical illustrations refer to theoretical gains of  $G = 5$  and  $G = 6$ . To better understand the implications that such values entail, we provide the following simplifying example scenario in which we explain how the considered values are obtained in some realistic use cases.

**Example 3.** *Let us consider the Netflix library, focusing on movies, and let us make the educated speculation that the popularity distribution of the library content follows a Zipf distribution with exponent parameter 1.4 (cf. [111]). Assume that we choose to apply coded caching on the part of the library that captures 90% of the traffic, such that on average 90% of the Netflix traffic will experience a streaming volume reduction by a (theoretical) factor of  $G$ . Thus in a Netflix library of approximately 3700 movies, coded caching is applied to the 100 most popular ones. The remaining 10% of the traffic is sent in an uncoded manner.*

*The subpacketization constraint will be largely defined by the latency requirements, which will ask from us, before subpacketization, to first split each movie into files that — in order to guarantee smooth streaming — will have to be sufficiently small. Assume a latency of two minutes, which can be seamlessly handled with a small buffer. This, with the extra assumption that movies last around 90 minutes, implies file (sub-movie) sizes of approximately  $\frac{2\text{min}}{90\text{min}} = 1/45$  of the movie size. Let us now consider several possible scenarios that we can encounter in practice.*

**First setting** *Let us assume that the receiving devices are each endowed with a cache of size equal to 25GB, and let us assume that they stream HD movies whose size is approximately 1.3GB. From this, we obtain a file size of approximately  $\frac{1.3\text{GB}}{45} = 28.8\text{MB}$ .*

*Under the assumption of atomic (indivisible) communication packets of size equal to 50 bytes, this brings us to a subpacketization of  $\frac{28.8\text{MB}}{50\text{B}} \approx 6 \cdot 10^5$ . This level of*

subpacketization, together with the corresponding  $\gamma = \frac{25\text{GB}}{100 \cdot 1.3\text{GB}} \approx 0.19$ , allows for a theoretical gain of  $G = 7$  (since  $\binom{\Lambda}{0.19\Lambda} \leq 6 \cdot 10^5$  and  $G = \Lambda\gamma + 1$ ). Recalling our example of the hardening-constrained setting with  $Q = 8$  (cf. Fig. 3.4), to attain the promised gain of  $G = 7$ , we require at least  $Q\Lambda \approx 240$  receiving nodes/antennas, which could represent 60 users with 4 receive antennas each.

**Second setting** Under approximately the same conditions, but for Full-HD movies of size 2.47GB, the corresponding scenario implies  $\gamma < 0.10$  and can allow for a gain close to  $G = 6$ . Recalling the same setting with  $Q = 8$  of Fig. 3.4, under the Full-HD assumption, we see that attaining the promised gain of  $G = 6$  requires a network with at least  $Q\Lambda \approx 400$  receiving nodes/antennas, which could represent  $K = 100$  users with 4 receive antennas each.

**Third setting** Let us now assume that each cache has a size equal to 5GB, and let us consider Standard Definition (SD-480p) streaming. Hence, the file (sub-movie) sizes become  $\frac{400.5\text{MB}}{45} = 8.9\text{MB}$  and  $\gamma = \frac{5\text{GB}}{100 \cdot 400.5\text{GB}} \approx 0.125$ . With an atomic communication packet size of 200 bytes, we have subpacketization  $4.5 \cdot 10^4$ , with a theoretical gain of  $G = 5$ . In this SD small-cache scenario, considering  $Q = 8$  corresponds to  $K = 256$  single antenna users, or 128 users with 2 antennas each.

### 3.5 Conclusions

In this chapter, we have investigated the performance of vector coded caching in MU-MISO systems, where three different linear precoding schemes (i.e., MF/ZF/RZF) that can be applied to vector coded caching are considered. We have derived simple but very tight closed-form expressions for the average sum-rates in MF, ZF and RZF based vector coded caching schemes respectively with the aid of random matrix theory. These derived expressions allow us to investigate the corresponding effective coded caching gain over the standard (without coded caching) MIMO system with the same system parameters. We have also provided the optimal number of users that must be served simultaneously to maximize the average sum-rate as a function of the number of transmit antennas, taking into account the impact of CSI acquisition at the BS.

Numerical results have shown a very substantial effective gain in the average sum-rate, for a class of precoders, which generally is expected to hold for an even larger class of precoders. The idea is simple: instead of sending precoded vectors, one after the other, we have now the ability to linearly combine  $G$  such vectors. In practice, such  $G$  should range between 4 and 6, and perhaps if  $\gamma$  is larger maybe 7. This  $G$  is bounded by the well-known subpacketization constraint. Numerical results have showed that the derived expressions hold tight in realistic (non-asymptotic) scenarios. It is also worth noting that MF/ZF/RZF precoding can recover most of the nominal gain in realistic SNR values, where we consider a variety of practical issues such as power dissemination across signals, realistic SNR values, as well as CSI costs. This work provides another example of how coded caching techniques can provide actual gains in realistic scenarios, and motivates the

analysis of other aspects such as the impact of multiple receive antennas, non-symmetric users and power allocation.



## Chapter 4

# More Practical Considerations in Vector Coded Caching

Chapter 3 showed how vector coded caching could provide a multiplicative boost in the throughput of MU-MISO systems over symmetric Rayleigh fading channels. In this chapter, we further investigate the performance of vector coded caching in MU-MIMO systems under various additional realistic considerations, that include path-loss, max-min fairness (MMF) and multi-antenna receivers. Specifically, in Section 4.1, we will consider vector coded caching in the presence of multi-antenna receivers, users with different path-loss, and block-diagonalization (BD) precoding with maximal ratio combining (MRC). In Section 4.2, we will derive the analytical expression for the overall throughput and for the effective coded caching gain, accounting for CSI costs and considering optimal power allocation for MMF. Moreover, some closed-form expressions are derived in some special cases of interest, that include massive MIMO. Subsequently, in Section 4.3, we consider a simple ZF precoder and derive lower and upper bounds on the downlink overall throughput. These bounds are numerically shown to be excellent approximations to the actual performance. Then in Section 4.4 we present additional numerical results and various comparisons that reveal the significant performance boost that vector coded caching offers to existing cacheless MU-MIMO systems, and finally in Section 4.5 we conclude this chapter.

Before the main content, let us define two new notations. For a set of matrices  $\{\mathbf{A}_{a_1, a_2} : a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2\}$  where each matrix  $\mathbf{A}_{a_1, a_2}$  with the same number of rows ( $A_r$ ) is labeled by a unique two-tuple subscript  $(a_1, a_2)$  for some  $a_1 \in \mathcal{A}_1$  and  $a_2 \in \mathcal{A}_2$ , we use  $\{\mathbf{A}_{a_1, a_2}\}_{a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2}$  to denote a matrix with  $A_r$  rows, which is composed by

$$\{\mathbf{A}_{a_1, a_2}\}_{a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2} \triangleq \begin{bmatrix} \mathbf{A}_{\mathcal{A}_1(1), \mathcal{A}_2(1)} | \mathbf{A}_{\mathcal{A}_1(1), \mathcal{A}_2(2)} | \cdots | \mathbf{A}_{\mathcal{A}_1(1), \mathcal{A}_2(|\mathcal{A}_2|)} \\ \mathbf{A}_{\mathcal{A}_1(2), \mathcal{A}_2(1)} | \mathbf{A}_{\mathcal{A}_1(2), \mathcal{A}_2(2)} | \cdots | \mathbf{A}_{\mathcal{A}_1(2), \mathcal{A}_2(|\mathcal{A}_2|)} | \cdots | \cdots | \cdots \\ \mathbf{A}_{\mathcal{A}_1(|\mathcal{A}_1|), \mathcal{A}_2(1)} | \mathbf{A}_{\mathcal{A}_1(|\mathcal{A}_1|), \mathcal{A}_2(2)} | \cdots | \mathbf{A}_{\mathcal{A}_1(|\mathcal{A}_1|), \mathcal{A}_2(|\mathcal{A}_2|)} \end{bmatrix}.$$

In the above,  $\mathcal{A}_1(i)$  denotes the  $i$ -th element of  $\mathcal{A}_1$ , and similarly for  $\mathcal{A}_2(i)$ . We note that the elements in  $\mathcal{A}_1$  and  $\mathcal{A}_2$  are both ordered, and that the matrices in the set

$\{\mathbf{A}_{a_1, a_2} : a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2\}$  may have different number of columns. We also use  $\text{Diag}\{a_1\}_{a_1 \in \mathcal{A}_1}$  to denote a diagonal matrix whose main diagonal is composed by the elements in the ordered set  $\mathcal{A}_1$ .

## 4.1 System Model and Problem Description

In a cache-aided downlink MU-MIMO system, a BS equipped with  $L$  antennas serves  $K$  cache-aided multi-antenna users where the users requests, as always in this thesis, different files (i.e., the worst-case) from a library  $\mathcal{F}$  of  $N$  ( $N \gg K$ ) equal-sized files in total. The BS has full access to the library  $\mathcal{F}$  while each user can only cache a fraction  $\gamma \in [0, 1]$  of library content. To accelerate delivery, we slightly modify the vector coded caching scheme originally proposed in [12], which has been illustrated in Section 3.1.2. Due to the finite file-size constraint, there are only  $\Lambda$  ( $\Lambda \ll K$ ) different cache states in vector coded caching, which leads to having  $B = \frac{K}{\Lambda}$  users sharing the same cache state and thus caching the corresponding content during the placement phase. During the delivery phase of vector coded caching, there are  $G \triangleq \Lambda\gamma + 1$  user groups (each group with their own cache state) selected for simultaneous service in each transmission stage, where  $Q \in [B]$  users in each selected cache state are active for receiving messages, which in turn implies that there are as many as  $GQ$  users served at a time (see Fig. 3.1 for a simple illustration).

We use  $\Psi$  to denote the selected  $G$  user groups chosen for simultaneous service, and we use  $U_{\psi, k}$  to denote the  $k$ -th ( $k \in [Q]$ ) active user from user group  $\psi \in \Psi$ . We assume that each served user  $U_{\psi, k}$  is equipped with  $M_{\psi, k}$  receiving antennas. We use  $M_\psi = \sum_{k \in [Q]} M_{\psi, k}$  to denote the total number of active receive antennas in the user group  $\psi \in \Psi$ . Thanks to multi-antenna receivers, the BS can send  $J_{\psi, k}$  symbols to  $U_{\psi, k}$  simultaneously, thereby enhancing the throughput toward  $U_{\psi, k}$ , where the maximum achievable value of  $J_{\psi, k}$  naturally depends on the number of transmit and receive antennas, the number of users per cache, and of course the precoding scheme. We use  $\mathbf{s}_{\psi, k}^T \triangleq \{s_{\psi, k, q}\}_{q \in [J_{\psi, k}]} \in \mathbb{C}^{J_{\psi, k}}$  and  $\mathbf{P}_{\psi, k} \triangleq \text{Diag}\{\sqrt{P_{\psi, k, q}}\}_{q \in [J_{\psi, k}]} \in \mathbb{C}^{J_{\psi, k} \times J_{\psi, k}}$  to denote the data vector to  $U_{\psi, k}$  and the corresponding power allocation matrix respectively, where the independent variables  $\{s_{\psi, k, q} : q \in [J_{\psi, k}]\}$  have zero-mean and unit-power. We also use the precoding matrix  $\mathbf{V}_{\psi, k} \triangleq \{\mathbf{v}_{\psi, k, q}\}_{q \in [J_{\psi, k}]} \in \mathbb{C}^{L \times J_{\psi, k}}$  to split different signal streams to  $U_{\psi, k}$  (which is mathematically written as  $\mathbf{V}_{\psi, k} \mathbf{s}_{\psi, k}$ ), where the unit-norm vector  $\mathbf{v}_{\psi, k, q} \in \mathbb{C}^L$  precodes  $s_{\psi, k, q}$ . The transmitted signal  $\mathbf{x}_\Psi \in \mathbb{C}^L$  for the selected user-group set  $\Psi$  at the BS is designed as

$$\mathbf{x}_\Psi = \sum_{\psi \in \Psi} \sum_{k \in [Q]} \mathbf{V}_{\psi, k} \mathbf{P}_{\psi, k} \mathbf{s}_{\psi, k}. \quad (4.1)$$

Given  $\mathbf{x}_\Psi$  in (4.1), the received signal vector  $\mathbf{y}_{\psi, k} \in \mathbb{C}^{M_{\psi, k}}$  at the typical user  $U_{\psi, k}$  takes the form

$$\mathbf{y}_{\psi, k} = \mathbf{H}_{\psi, k}^T \mathbf{x}_\Psi + \mathbf{z}_{\psi, k} = \mathbf{H}_{\psi, k}^T \mathbf{V}_{\psi, k} \mathbf{P}_{\psi, k} \mathbf{s}_{\psi, k} + \mathbf{z}_{\psi, k}$$

$$\begin{aligned}
 & + \underbrace{\mathbf{H}_{\psi,k}^T \sum_{k' \in [Q] \setminus k} \mathbf{V}_{\psi,k'} \mathbf{P}_{\psi,k'} \mathbf{s}_{\psi,k'}}_{\text{intra-group interference}} + \underbrace{\mathbf{H}_{\psi,k}^T \sum_{\phi \in \Psi \setminus \psi} \sum_{\vartheta \in [Q]} \mathbf{V}_{\phi,\vartheta} \mathbf{P}_{\phi,\vartheta} \mathbf{s}_{\phi,\vartheta}}_{\text{inter-group interference}}, \\
 \end{aligned} \tag{4.2}$$

where  $\mathbf{z}_{\psi,k} \sim \mathcal{CN}(\mathbf{0}_L, N_0 \mathbf{I}_L)$  denotes the AWGN, and where  $\mathbf{H}_{\psi,k} \in \mathbb{C}^{L \times M_{\psi,k}}$  denotes the channel matrix from the BS to  $U_{\psi,k}$ . The elements of  $\mathbf{H}_{\psi,k}$  will be i.i.d. complex Gaussian random variables with zero-mean and variance  $\beta_{\psi,k}$  if we consider Rayleigh fading channels. We note that  $\beta_{\psi,k}$  here accounts for the large-scale shadowing and/or pathloss. Let  $\mathbf{R}_{\psi,k} \triangleq \{\mathbf{r}_{\psi,k,q}\}_{q \in [M_{\psi,k}]} \in \mathbb{C}^{M_{\psi,k} \times M_{\psi,k}}$  with each unit-norm column  $\mathbf{r}_{\psi,k,q} \in \mathbb{C}^{M_{\psi,k}}$  be the channel-dependent decoding matrix at  $U_{\psi,k}$ . As  $U_{\psi,k}$  knows (has access to, from their cache; as we can recall from Section 3.1.2) the messages  $\{\mathbf{s}_{\phi,\vartheta} : \phi \in \Psi \setminus \psi, \vartheta \in [Q]\}$  intended by the active users of other user groups in  $\Psi$ , then the inter-group interference in (4.2) can be removed by using the cached content in  $U_{\psi,k}$  and the composite CSI  $\{\mathbf{H}_{\psi,k}^T \mathbf{V}_{\phi,\vartheta} \mathbf{P}_{\phi,\vartheta} : \phi \in \Psi \setminus \psi, \vartheta \in [Q]\}$ , the cost of which we will account for in our analysis. After removing the inter-group interference via vector coded caching and using the decoding matrix  $\mathbf{R}_{\psi,k}$ , the signal vector for decoding at  $U_{\psi,k}$  is

$$\mathbf{y}'_{\psi,k} = \mathbf{R}_{\psi,k}^H \mathbf{H}_{\psi,k}^T \mathbf{V}_{\psi,k} \mathbf{P}_{\psi,k} \mathbf{s}_{\psi,k} + \mathbf{R}_{\psi,k}^H \sum_{k' \in [Q] \setminus k} \mathbf{H}_{\psi,k}^T \mathbf{V}_{\psi,k'} \mathbf{P}_{\psi,k'} \mathbf{s}_{\psi,k'} + \mathbf{z}'_{\psi,k}, \tag{4.3}$$

where  $\mathbf{z}'_{\psi,k} \triangleq \mathbf{R}_{\psi,k}^H \mathbf{z}_{\psi,k} \sim \mathcal{CN}(\mathbf{0}_L, N_0 \mathbf{I}_L)$  in view of the property of multi-variate Gaussian distribution.

Building on the general vector-clique structure in [12], we are here free to choose the precoding schemes, as well as calibrate at will the dimensionality of each vector clique. This freedom is essential in controlling the impact of CSI costs and of power-splitting across users, both of which directly affect the performance in practical SNR regimes [66–68]. We refer to Section 3.1.2 for the cache placement phase and the subsequent delivery phase in vector coded caching.

#### 4.1.1 BD Precoding and MRC Combining

As pointed out in [112], complete channel diagonalization (e.g., ZF) at the BS is suboptimal since each multi-antenna user is able to coordinate the processing of its own receiver outputs. We thus alternatively consider the well-known BD precoding method [112]. Toward this, we first define the matrix  $\mathbf{H}_{\psi,-k} \triangleq \{\mathbf{H}_{\psi,\iota}\}_{\iota \in [Q] \setminus k} \in \mathbb{C}^{L \times (M_{\psi} - M_{\psi,k})}$ . To cancel the intra-group interference in (4.2),  $\mathbf{V}_{\psi,k}$  must lie in the null-space of  $\mathbf{H}_{\psi,-k}^*$  such that the product of  $\mathbf{H}_{\psi,-k}^T$  and  $\mathbf{V}_{\psi,k}$  is a matrix with all elements equaling zero for any  $k \in [Q]$ . Furthermore, to successfully eliminate the inter-symbol interference in  $U_{\psi,k}$ , we must have that

$$J_{\psi,k} \triangleq \min \left\{ \text{Rank}(\mathbf{H}_{\psi,-k}^*), \text{Rank}(\mathbf{H}_{\psi,k}) \right\} \tag{4.4}$$

where  $J_{\psi,k}$  denotes the maximum allowable number of symbols that can be simultaneously transmitted to  $U_{\psi,k}$ . For independent Rayleigh fading channels, then  $J_{\psi,k} = \min \{L - (M_{\psi} - M_{\psi,k}), M_{\psi,k}\}$ . We use  $Q_{\psi,\max}$  to denote the maximum  $Q$  in the user-group  $\psi$ ,

which equals  $\lfloor \frac{M+L-1}{M} \rfloor$  if all the served users have the same number of receive antennas  $M$ . If the users are equipped with different numbers of antennas, we can determine  $Q_{\psi, \max}$  via solving  $\frac{\sum_{k \in [Q]} M_{\psi, k}}{1-L+\sum_{k \in [Q]} M_{\psi, k}} - Q = 0$ , which implies that  $Q_{\psi, \max}$  depends on how we select the users. As we consider an equal  $Q$  for each group, the aforementioned maximum allowable  $Q$  takes the form  $Q_{\max} = \min_{\psi \in \Psi} Q_{\psi, \max}$ .

Let  $\mathbf{T}_{\psi, -k} \triangleq \mathbf{I}_L - \mathbf{H}_{\psi, -k}^* \left( \mathbf{H}_{\psi, -k}^T \mathbf{H}_{\psi, -k}^* \right)^{-1} \mathbf{H}_{\psi, -k}^T$  be the projection matrix which maps any vector  $\mathbf{m} \in \mathbb{C}^L$  into the null-space of  $\mathbf{H}_{\psi, -k}^*$ . We note that  $\mathbf{T}_{\psi, -k}^2 = \mathbf{T}_{\psi, -k} = \mathbf{T}_{\psi, -k}^H$  according to the projection matrix properties. The BD precoding matrix dedicated to  $U_{\psi, k}$  can be written as

$$\mathbf{V}_{\psi, k} = \left\{ \frac{\mathbf{T}_{\psi, -k} \mathbf{m}_{\psi, k, q}}{\|\mathbf{T}_{\psi, -k} \mathbf{m}_{\psi, k, q}\|} \right\}_{q \in [J_{\psi, k}]} \in \mathbb{C}^{L \times J_{\psi, k}}. \quad (4.5)$$

By using this BD precoding, the signal vector at  $U_{\psi, k}$  in (4.3) becomes

$$\begin{aligned} \mathbf{y}'_{\psi, k} &= \mathbf{R}_{\psi, k}^H \mathbf{H}_{\psi, k}^T \mathbf{V}_{\psi, k} \mathbf{P}_{\psi, k} \mathbf{s}_{\psi, k} + \mathbf{z}'_{\psi, k} \\ &= \left[ \mathbf{r}_{\psi, k, 1}^H \mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, 1} \sqrt{P_{\psi, k, 1}} s_{\psi, k, 1}, \dots, \mathbf{r}_{\psi, k, J_{\psi, k}}^H \mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, J_{\psi, k}} \sqrt{P_{\psi, k, J_{\psi, k}}} s_{\psi, k, J_{\psi, k}} \right]^T + \mathbf{z}'_{\psi, k}. \end{aligned} \quad (4.6)$$

We have that  $|\mathbf{r}_{\psi, k, q}^H \mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, q}|^2 \leq \|\mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, q}\|^2$ , where equality is achieved only if  $\mathbf{r}_{\psi, k, q} = \theta \mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, q}$  for a constant  $\theta \neq 0$  according to the well-known Cauchy-Schwarz inequality. Here, we set  $\theta = 1/\|\mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, q}\|$  to normalize  $\mathbf{r}_{\psi, k, q}$ , which corresponds to the MRC receiver. Therefore, under the common Gaussian signaling, the signal-to-interference plus noise ratio (SINR) for decoding  $s_{\psi, k, q}$  at  $U_{\psi, k}$  is

$$\begin{aligned} \text{SINR}_{\psi, k, q} &= \frac{P_{\psi, k, q}}{N_0} |\mathbf{r}_{\psi, k, q}^H \mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, q}|^2 \\ &\stackrel{\text{MRC}}{\leq} \frac{P_{\psi, k, q}}{N_0} \|\mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, q}\|^2 = \frac{P_{\psi, k, q}}{N_0} \mathbf{v}_{\psi, k, q}^H \mathbf{H}_{\psi, k}^* \mathbf{H}_{\psi, k}^T \mathbf{v}_{\psi, k, q} \\ &= \frac{P_{\psi, k, q}}{N_0} \frac{\mathbf{m}_{\psi, k, q}^H \mathbf{T}_{\psi, -k} \mathbf{H}_{\psi, k}^* \mathbf{H}_{\psi, k}^T \mathbf{T}_{\psi, -k} \mathbf{m}_{\psi, k, q}}{\|\mathbf{T}_{\psi, -k} \mathbf{m}_{\psi, k, q}\|^2}. \end{aligned} \quad (4.7)$$

Toward providing CSI estimates to the BS and the users, we will consider the common TDD uplink-downlink pilot transmission, as this applies to MU-MIMO systems. Let  $T_c$  be the coherence block time (in symbols), and let  $\beta_{\text{tot}}$  be the number of resources per user's antenna and per block used for pilot transmission. The corresponding effective rate for  $U_{\psi, k}$  is then

$$R_{\psi, k} = \xi_{G, Q} \sum_{q=1}^{J_{\psi, k}} \ln(1 + \text{SINR}_{\psi, k, q}) \text{ nats/s/Hz}, \quad (4.8)$$

where  $\xi_{G, Q} \triangleq 1 - \beta_{\text{tot}} (\sum_{\psi \in \Psi} \sum_{k \in [Q]} M_{\psi, k}) / T_c$  accounts for CSI costs [109]. Without loss of generality, we sort  $\{s_{\psi, k, q} : q \in [J_{\psi, k}]\}$  in descending order according to the corresponding  $\{\text{SINR}_{\psi, k, q} : q \in [J_{\psi, k}]\}$  at  $U_{\psi, k}$ . In the following, we optimize the BD precoder  $\{\mathbf{v}_{\psi, k, q} : \psi \in \Psi, k \in [Q], q \in [J_{\psi, k}]\}$  in order to yield the maximum effective rate for each served user.

**Lemma 4.1.** *The optimal precoding vector  $\mathbf{v}_{\psi,k,q}$  for  $\mathbf{U}_{\psi,k}$  to decode  $s_{\psi,k,q}$  under the BD-MRC scheme is*

$$\mathbf{v}_{\psi,k,q}^* = \frac{\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}}{\|\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}\|}, \quad (4.9)$$

where  $\mathbf{t}_{\psi,k,q} \in \mathbb{C}^{M_{\psi,k}}$  is the eigenvector associated with the  $q$ -th largest (non-zero) eigenvalue  $\lambda_{\psi,k,q}$  of  $\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \in \mathbb{C}^{M_{\psi,k} \times M_{\psi,k}}$ . The corresponding SINR when decoding  $s_{\psi,k,q}$  takes the form

$$\text{SINR}_{\psi,k,q}^{\text{BD-MRC}} = \frac{P_{\psi,k,q}}{N_0} \lambda_{\psi,k,q}, \quad (4.10)$$

and the resulting effective rate for  $\mathbf{U}_{\psi,k}$  then takes the form

$$R_{\psi,k} = \xi_{G,Q} \sum_{q=1}^{J_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q}}{N_0} \lambda_{\psi,k,q} \right). \quad (4.11)$$

*Proof.* The proof is relegated to Appendix C.1.  $\square$

Let  $P_{\psi,k} = \sum_{q=1}^{J_{\psi,k}} P_{\psi,k,q}$  be the transmit power allocated to  $\mathbf{U}_{\psi,k}$ . From (4.11), we know that when  $P_{\psi,k}$  is fixed, adjusting power allocation among the symbols intended by  $\mathbf{U}_{\psi,k}$ , does not affect the effective rates for other users. Let us define  $(x)^+ \triangleq \max(x, 0)$ . The following corollary considers the use of water filling when assigning power to the symbols intended by  $\mathbf{U}_{\psi,k}$ , under the power constraint  $P_{\psi,k}$ , in order to maximize the effective rate in (4.11).

**Corollary 4.1.** *The optimal power allocated to each symbol  $s_{\psi,k,q}$  for maximizing the effective rate in (4.11), takes the form*

$$P_{\psi,k,q} = \left( \frac{1}{\alpha_{\psi,k}} - \frac{N_0}{\lambda_{\psi,k,q}} \right)^+, \quad (4.12)$$

where, under a power constraint  $P_{\psi,k}$ , the Lagrange multiplier  $\alpha_{\psi,k}$  is the solution to

$$\sum_{q=1}^{J_{\psi,k}} \left( \frac{1}{\alpha_{\psi,k}} - \frac{N_0}{\lambda_{\psi,k,q}} \right)^+ = P_{\psi,k}. \quad (4.13)$$

Then, the corresponding optimal effective rate for  $\mathbf{U}_{\psi,k}$ , takes the form

$$R_{\psi,k}^*(P_{\psi,k}) = \xi_{G,Q} \sum_{q=1}^{J_{\psi,k}} \ln \left( 1 + \left( \frac{\lambda_{\psi,k,q}}{N_0 \alpha_{\psi,k}} - 1 \right)^+ \right). \quad (4.14)$$

*Proof.* The proof is direct from the water-filling algorithm (cf. [113, Ch. 10]).  $\square$

Let us now consider max-min fairness, where the minimum effective rate among the simultaneously served users is maximized via power allocation under a specific precoding scheme. The MMF problem for serving the users  $\{\mathbf{U}_{\psi,k} : \psi \in \Psi, k \in [Q]\}$  can be formulated as

$$\mathcal{S}_1 : \begin{cases} \max_{\mathbf{P}_\Psi} \min_{\psi \in \Psi} \min_{k \in [Q]} \xi_{G,Q} \sum_{q=1}^{J_{\psi,k}} \ln \left( 1 + \text{SINR}_{\psi,k,q} \right) \\ \text{s.t. } P = \text{Tr}\{\mathbf{P}_\Psi\} = \sum_{\psi \in \Psi} \sum_{k \in [Q]} P_{\psi,k} \leq P_{\text{tot}}, \end{cases} \quad (4.15)$$

where the diagonal matrix  $\mathbf{P}_\Psi \in \mathbb{C}^{GQ \times GQ}$  collects the power values allocated to the served  $GQ$  users, and where  $P_{\text{tot}}$  is the maximum allowable transmit power at the BS.

### 4.1.2 Main Performance Metrics

We will henceforth use the term  $(G, Q)$ -vector coded caching, to refer to the vector coded caching scheme when it serves  $G$  groups with  $Q$  users per group. We will also use the term *BD-based  $(G, Q)$ -vector coded caching* to refer to the same scheme when the underlying precoder is the BD precoder, and similarly we will use ZF-based  $(G, Q)$ -vector coded caching, when considering the ZF precoder. Let us now formally define some important metrics of interest.

**Definition 5.** (Effective sum-rate). *For a  $(G, Q)$ -vector coded caching scheme, its effective (instantaneous) sum-rate is denoted by  $R(G, Q)$  and is defined as the total effective rate (after accounting for CSI costs) summed over the  $GQ$  simultaneously served users. Moreover,  $\bar{R}(G, Q)$  represents  $R(G, Q)$  averaged over channel fading.*

**Definition 6.** (Effective gain over MIMO). *For a given set of SNR and  $L$  resources, and a fixed underlying precoder class, the effective gain, after accounting for CSI costs, of the  $(G, Q)$ -vector coded caching scheme over the cacheless scenario (corresponding to  $G = 1$ , and an operating multiplexing gain  $Q'$ ), will be denoted as  $\mathcal{G} \triangleq \frac{\bar{R}^*(G, Q)}{\bar{R}^*(1, Q')}$ , where  $\bar{R}^*(G, Q)$  describes the rate  $R(G, Q)$  that is first optimized via power allocation under the MMF criterion (cf. (4.15)), and then averaged over channel fading. We also call  $\mathcal{G}^* \triangleq \frac{\max_Q \bar{R}^*(G, Q)}{\max_{Q'} \bar{R}^*(1, Q')}$  as the effective gain of optimized rates, where  $Q$  and  $Q'$  are also independently optimized.*

## 4.2 BD-MRC Analysis for Multi-Antenna Receivers

### 4.2.1 Effective Sum-Rate and Effective Gain: the case of BD-MRC

We first note that the set  $\{\lambda_{\psi, k, q} : \psi \in \Psi, k \in [Q], q \in [J_{\psi, k}]\}$  is a function of the channel gains but not of the power allocation policy. By using the effective rate expression in (4.11), the MMF optimization problem in (4.15), under the BD-MRC scheme for downlink power allocation, can be transformed into

$$\mathcal{S}_2 : \begin{cases} \max_{\mathbf{P}_\Psi} \min_{\psi \in \Psi} \min_{k \in [Q]} \xi_{G, Q} \sum_{q=1}^{J_{\psi, k}} \ln \left( 1 + \frac{P_{\psi, k, q}}{N_0} \lambda_{\psi, k, q} \right) \\ \text{s.t. } P_t = \sum_{\psi \in \Psi} \sum_{k \in [Q]} \sum_{q \in [J_{\psi, k}]} P_{\psi, k, q} = P_{\text{tot}}. \end{cases} \quad (4.16)$$

The following theorem addresses the optimization problem in (4.16) and derives the effective instantaneous sum-rate  $R_{\text{BD-MRC}}^*$  and the effective gain of optimized rates (cf. Definition 6). In the following,  $f_{\psi, k}^{-1}(\cdot)$  will denote the inverse function of  $R_{\psi, k}^*(P_{\psi, k})$  in (4.14) (which is a monotonically increasing function w.r.t.  $P_{\psi, k}$ ).

**Theorem 4.1.** *The effective instantaneous sum-rate  $R_{\text{BD-MRC}}^*$  under optimal power allocation for the MMF problem in (4.16) is the solution to*

$$\sum_{\psi \in \Psi} \sum_{k \in [Q]} f_{\psi, k}^{-1} \left( \frac{R_{\text{BD-MRC}}^*}{GQ} \right) = P_{\text{tot}}, \quad (4.17)$$

and then the corresponding effective gain of optimized rates under the BD-MRC scheme, takes the form

$$\mathcal{G}_{BD-MRC}^* = \frac{\max_{Q \in [Q_{\max}]} \mathbb{E}_h \{R_{BD-MRC}^*(G, Q)\}}{\max_{Q' \in [Q_{\max}]} \mathbb{E}_h \{R_{BD-MRC}^*(1, Q')\}}. \quad (4.18)$$

Furthermore, the optimal rate is bounded as

$$\tilde{R}_{BD-MRC}^* \leq R_{BD-MRC}^* \leq \hat{R}_{BD-MRC}^*$$

where  $\tilde{R}_{BD-MRC}^*$  and  $\hat{R}_{BD-MRC}^*$  are respectively the solutions to

$$\sum_{\psi \in \Psi} \sum_{k \in [Q]} \frac{J_{\psi,k} N_0}{\left(\min_{q \in [J_{\psi,k}]} \lambda_{\psi,k,q}\right)} \left( \exp \left( \frac{\tilde{R}_{BD-MRC}^*}{\xi_{G,Q} J_{\psi,k} G Q} \right) - 1 \right) = P_{\text{tot}}, \quad (4.19)$$

$$\sum_{\psi \in \Psi} \sum_{k \in [Q]} \frac{J_{\psi,k} N_0}{\left(\max_{q \in [J_{\psi,k}]} \lambda_{\psi,k,q}\right)} \left( \exp \left( \frac{\hat{R}_{BD-MRC}^*}{\xi_{G,Q} J_{\psi,k} G Q} \right) - 1 \right) = P_{\text{tot}}. \quad (4.20)$$

*Proof.* As the power allocation among the symbols intended by  $U_{\psi,k}$  does not affect the power allocation to other served users, the effective rate for  $U_{\psi,k}$  must reach its optimal bound under the power constraint  $P_{\psi,k}^*$  (optimal power allocated to  $U_{\psi,k}$  for MMF), which has been solved in Corollary 4.1. Then it is easy to see that  $R_{\psi,k}^*(P_{\psi,k})$  in (4.14) is a monotonically increasing function w.r.t.  $P_{\psi,k}$ . When the optimum for (4.16) is achieved, if there exists one user whose effective rate is higher than the smallest rate, this user can “borrow” some power to the user with the smallest rate until their rates are the same, without affecting the rates for other users, which enhances the smallest effective rates, and which is contradictory to the optimal power allocation assumption. Therefore, the users must have the same effective rate (equalling  $R_{BD-MRC}^*/G/Q$ ) under the optimal power allocation in (4.16). Considering  $\sum_{\psi \in \Psi} \sum_{k \in [Q]} P_{\psi,k}^* = P_{\text{tot}}$  and the inverse function of  $R_{\psi,k}^*(P_{\psi,k})$ , we can obtain (4.17).

To derive the lower-bound  $\tilde{R}_{BD-MRC}^*$ , we first consider that

$$R_{\psi,k} = \xi_{G,Q} \sum_{q=1}^{J_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q}}{N_0} \lambda_{\psi,k,q} \right) \geq \xi_{G,Q} J_{\psi,k} \ln \left( 1 + \frac{P_{\psi,k,q}}{N_0} \left( \min_{q \in [J_{\psi,k}]} \lambda_{\psi,k,q} \right) \right), \quad (4.21)$$

which is substituted into (4.16) as the objective function. Using similar analysis as that leading to (4.17), we can easily obtain (4.19). The upper-bound  $\hat{R}_{BD-MRC}^*$  follows the same procedure as  $\tilde{R}_{BD-MRC}^*$ , with a difference being that upper bounding  $R_{\psi,k}$  now uses  $\max_{q \in [J_{\psi,k}]} \lambda_{\psi,k,q}$ .  $\square$

We also have the following, which considers the commonly-assumed symmetric case where each user receives an equal number of symbols.

**Corollary 4.2.** *In the symmetric case where  $J_{\psi,k} = J$  for any  $\psi \in \Psi$  and  $k \in [Q]$ , the lower and upper bounds to the optimal rate  $R_{BD-MRC}^*$  take the form*

$$\tilde{R}_{BD-MRC}^* \triangleq \xi_{G,Q} G Q J \ln \left( 1 + \frac{P_{\text{tot}}}{N_0 J \sum_{\psi \in \Psi} \sum_{k \in [Q]} (\min_{q \in [J_{\psi,k}]} \{\lambda_{\psi,k,q}\})^{-1}} \right), \quad (4.22)$$

$$\hat{R}_{BD-MRC}^* \triangleq \xi_{G,Q} G Q J \ln \left( 1 + \frac{P_{\text{tot}}}{N_0 J \sum_{\psi \in \Psi} \sum_{k \in [Q]} (\max_{q \in [J_{\psi,k}]} \{\lambda_{\psi,k,q}\})^{-1}} \right). \quad (4.23)$$

*Proof.* We can easily derive the expressions of  $\tilde{R}_{BD-MRC}^*$  and  $\hat{R}_{BD-MRC}^*$  from (4.19) and (4.20) respectively, after setting  $J_{\psi,k} = J$ ,  $\forall \psi \in \Psi, k \in [Q]$ .  $\square$

**Remark 4.1.** *We note that Lemma 4.1, Corollary 4.1, Theorem 4.1 and Corollary 4.2, are all valid for any propagation channel model, including of course Rayleigh fading, Rician-K fading, and Keyhole channels. The same results also hold for scenarios that involve non full-rank channel matrix product  $\mathbf{H}_{\psi,k}^T \mathbf{H}_{\psi,k}^*$  for any  $\psi \in \Psi$  and  $k \in [Q]$ , in which case we apply the pseudo-inverse of  $\mathbf{H}_{\psi,-k}^T \mathbf{H}_{\psi,-k}^*$  in the projection matrix  $\mathbf{T}_{\psi,-k}$ .*

#### 4.2.2 Special Case (i): Massive MIMO Regime Over Rayleigh Fading Channels

In this subsection, we consider a very large number of antennas  $L$  and a relatively small  $M_{\psi,k}$ . This will allow us to simplify our analysis on the eigenvalues of  $\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \in \mathbb{C}^{M_{\psi,k} \times M_{\psi,k}}$ . When  $L \gg \sum_{k \in [Q]} M_{\psi,k}$  for  $\forall \psi \in \Psi$ , we can reasonably assume that  $J_{\psi,k}$  in (4.4) is always  $M_{\psi,k}$  (for Rayleigh fading channels). Lemma 4.2 shows the simplified results of Theorem 4.1.

**Lemma 4.2.** *In the massive MIMO regime, the effective instantaneous sum-rate  $R_{BD-MRC}^*$  with CSI costs considerations, and under the BD-MRC scheme and optimal power allocation for MMF over Rayleigh fading channels, can be obtained via numerically solving*

$$\sum_{\psi \in \Psi} \sum_{k \in [Q]} \frac{N_0 M_{\psi,k}}{\beta_{\psi,k} (L - \sum_{k' \in [Q] \setminus k} M_{\psi,k'})} \left( \exp \left( \frac{R_{BD-MRC}^*}{\xi_{G,Q} M_{\psi,k} G Q} \right) - 1 \right) = P_{\text{tot}}. \quad (4.24)$$

*The corresponding optimal power allocation policy that yields  $R_{BD-MRC}^*$  takes the form*

$$P_{\psi,k,q} = \frac{N_0 M_{\psi,k}}{\beta_{\psi,k} (L - \sum_{k' \in [Q] \setminus k} M_{\psi,k'})} \left( \exp \left( \frac{R_{\psi,k}^*}{\xi_{G,Q} M_{\psi,k} G Q} \right) - 1 \right), \text{ for } \forall \psi \in \Psi, k \in [Q], \quad (4.25)$$

*When  $M_{\psi,k} = M$  for  $\forall \psi \in \Psi, k \in [Q]$ , then*

$$R_{BD-MRC}^*(G, Q) = \xi_{G,Q} G Q M \ln \left( 1 + \frac{P_{\text{tot}} (L - (Q-1)M)}{N_0 M \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right). \quad (4.26)$$

*Note that (4.24), (4.25) and (4.26) are independent of channel fading.*

*Proof.* The proof is relegated to Appendix C.2.  $\square$

### 4.2.3 Special Case (ii): Single-Antenna Receivers in BD

We here consider BD precoding with optimal beamforming for each single-antenna receiver. Let us first define  $J' \triangleq L - Q + 1$ ,  $\nu \triangleq GQJ'$ , as well as

$$\Delta_\nu^{(1)} \triangleq \left\{ \left( 1 + \frac{J'}{\beta_{\psi,k}} \right)_{J'} : \psi \in \Psi, k \in [Q] \right\} \quad (4.27)$$

$$\Delta_\nu^{(2)} \triangleq \left\{ \left( \frac{J'}{\beta_{\psi,k}} \right)_{J'} : \psi \in \Psi, k \in [Q] \right\}, \quad (4.28)$$

where  $(\cdot)_{J'}$  stands for repeating  $J'$  times the enclosed argument. We also use  $G_{\cdot, \cdot}(\cdot)$  to denote the Meijer's G-function (cf. [76, Eq. (9.301)]), and naturally define  $\rho$  as  $\rho \triangleq \frac{P_{\text{tot}}}{N_0}$ . We proceed with the following lemma.

**Lemma 4.3.** *The optimal effective sum-rate of BD precoding for single-antenna users under the power allocation for MMF over Rayleigh fading channels takes the form*

$$\bar{R}_{BD}^*(G, Q) = GQ\xi_{G,Q} \left[ \prod_{\psi \in \Psi, k \in [Q]} \left( \frac{J'}{\beta_{\psi,k}} \right)_{J'} \right] \int_0^\infty \ln \left( 1 + \frac{\rho}{x} \right) G_{\nu, \nu}^{\nu, 0} \left( \exp(-x) \middle| \begin{matrix} \Delta_\nu^{(1)} \\ \Delta_\nu^{(2)} \end{matrix} \right) dx. \quad (4.29)$$

When all the served  $GQ$  users have the same path-loss  $\beta_{\text{PL}}$ , the effective sum-rate in (4.29) can be simplified as

$$\bar{R}_{BD}^*(G, Q) = \frac{\rho GQ\xi_{G,Q}}{\beta_{\text{PL}}(\nu - 1)!} G_{2,3}^{2,2} \left( \frac{\rho}{\beta_{\text{PL}}} \middle| \begin{matrix} 0, 0 \\ 0, \nu-1, -1 \end{matrix} \right). \quad (4.30)$$

In the high-SNR regime of  $\rho \rightarrow \infty$ , this takes the form

$$\bar{R}_{BD}^*(G, Q) = GQ\xi_{G,Q} \ln \rho - GQ\xi_{G,Q} \left( \ln \beta_{\text{PL}} - C + \sum_{\ell=1}^{\nu-1} \frac{1}{\ell} \right) + o(1), \quad (4.31)$$

where  $C = 0.5772\dots$  denotes the Euler-Mascheroni constant, and where  $\sum_{\ell=1}^{\nu-1} \frac{1}{\ell} = 0$  if  $\nu = 1$ . The second term  $\mathcal{L}^\infty \triangleq -GQ\xi_{G,Q} \left( \ln \beta_{\text{PL}} - C + \sum_{\ell=1}^{\nu-1} \frac{1}{\ell} \right)$  in (4.31), represents the high SNR power offset due to path-loss and fading [86].

*Proof.* The proof is relegated to Appendix C.3. □

## 4.3 ZF Precoding Analysis for Multi-Antenna Receivers

In this section, we analyze the effective sum-rate achieved by the cache-aided downlink schemes of Section 3.1.2 (over independent Rayleigh fading channels) for the case of the ZF linear precoder. After doing so, we also report the effective gains offered by the ZF-based  $(G, Q)$ -vector coded caching scheme, where the gains are over the  $(G = 1, Q')$  cacheless equivalent. In contrast to BD precoding where both the precoding matrix optimization and power allocation should be adjusted according to the instantaneous

channel fading (recall Theorem 4.1), we here in (cf. Theorem 4.2) simply perform channel matrix inversion without any precoder optimization, and we simply calibrate the power allocation as a function of the (large-scale) path-loss which of course changes much slower than fading does. Our numerical results show that the performance gap between the simpler BD precoder and the fully optimized ZF precoder is negligible for practical values of receive antennas (e.g.,  $M_{\psi,k} \leq 4$ ).

Following the conventional ZF precoding for single-antenna receivers (cf. [114]), we completely separate the transmitted  $M_\psi = \sum_{k \in [Q]} M_{\psi,k}$  symbol streams such that there is no intra-group and inter-group interference. Therefore, the  $M_{\psi,k}$  symbols simultaneously sent to  $U_{\psi,k}$ , are fully separated (using complete channel diagonalization at the BS), and user  $U_{\psi,k}$  independently decodes the intended  $M_{\psi,k}$  symbols without interference from other users in its cache-group  $\psi$ . This corresponds to a decoding matrix  $\mathbf{R}_{\psi,k} = \mathbf{I}_{M_{\psi,k}}$  at user  $U_{\psi,k}$ . Specifically, after defining the channel matrix  $\mathbf{H}_\psi \triangleq \{\mathbf{H}_{\psi,k}\}_{k \in [Q]} \in \mathbb{C}^{L \times \sum_{k \in [Q]} M_{\psi,k}}$  representing the channel between the BS and the active users in cache-group  $\psi$ , the ZF variant for multi-antenna receivers is designed as

$$\mathbf{V}_\Psi = \left\{ \mathbf{H}_\psi^* (\mathbf{H}_\psi^T \mathbf{H}_\psi^*)^{-1} \circ \mathbf{D}_{\text{ZF},\psi} \right\}_{\psi \in \Psi} \in \mathbb{C}^{L \times \sum_{\psi \in \Psi} \sum_{k \in [Q]} M_{\psi,k}}, \quad (4.32)$$

where  $\{\mathbf{D}_{\text{ZF},\psi} : \psi \in \Psi\}$  are the normalization matrices which guarantee that the norm-2 of each column in  $\mathbf{V}_\Psi$  is equal to 1, and where each of these matrices is given by

$$\mathbf{D}_{\text{ZF},\psi} \triangleq \left\{ \sqrt{\left( [(\mathbf{H}_\psi^T \mathbf{H}_\psi^*)^{-1}]_{k(q),k(q)} \right)^{-1}} \mathbf{1}_L \right\}_{k \in [Q], q \in [M_{\psi,k}]} \in \mathbb{C}^{L \times \sum_{k \in [Q]} M_{\psi,k}}, \quad (4.33)$$

where  $k(q) \triangleq q + \mathbb{I}\{k > 1\} \sum_{k'=1}^{k-1} M_{\psi,k'}$  which varies with  $\psi$ . In the above,  $\mathbb{I}\{\cdot\}$  denotes the indicator function.

We first present the upper and lower bounds for the effective rate at any given user, averaged over channel fading, under ZF precoding. In what follows, we use  $M_\psi \triangleq \sum_{k \in [Q]} M_{\psi,k}$ .

**Proposition 4.1.** *The effective average rate at a typical user  $U_{\psi,k}$ ,  $\psi \in \Psi, k \in [Q]$  under ZF-based precoding, is bounded as  $\tilde{R}_{\psi,k}^{\text{ZF}} \leq \bar{R}_{\psi,k}^{\text{ZF}} \leq \hat{R}_{\psi,k}^{\text{ZF}}$ , where*

$$\tilde{R}_{\psi,k}^{\text{ZF}} \triangleq \xi_{G,Q} \sum_{q=1}^{M_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q} (L - M_\psi) \beta_{\psi,k}}{N_0} \right), \quad (4.34)$$

$$\hat{R}_{\psi,k}^{\text{ZF}} \triangleq \xi_{G,Q} \sum_{q=1}^{M_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q} (L - M_\psi + 1) \beta_{\psi,k}}{N_0} \right). \quad (4.35)$$

*Proof.* The proof is relegated to Appendix C.4.  $\square$

In the following, we separately optimize the bounds in Proposition 4.1. The MMF optimization problems are respectively of the form

$$\mathcal{S}_3 : \begin{cases} \max_{\mathbf{P}_\Psi} \min_{\psi \in \Psi} \min_{k \in [Q]} \xi_{G,Q} \sum_{q=1}^{M_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q} (L - M_\psi) \beta_{\psi,k}}{N_0} \right) \\ \text{s. t. } P_t = \text{Tr}\{\mathbf{P}_\Psi\} = \sum_{\psi \in \Psi} \sum_{k \in [Q]} \sum_{q \in [M_{\psi,k}]} P_{\psi,k,q} \leq P_{\text{tot}}. \end{cases} \quad (4.36)$$

$$\mathcal{S}_4 : \begin{cases} \max_{\mathbf{P}_\Psi} \min_{\psi \in \Psi} \min_{k \in [Q]} \xi_{G,Q} \sum_{q=1}^{M_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q}(L-M_\psi+1)\beta_{\psi,k}}{N_0} \right) \\ \text{s. t. } P_t = \text{Tr}\{\mathbf{P}_\Psi\} = \sum_{\psi \in \Psi} \sum_{k \in [Q]} \sum_{q \in [M_{\psi,k}]} P_{\psi,k,q} \leq P_{\text{tot}}. \end{cases} \quad (4.37)$$

Obviously,  $P_t$  should reach its upper-bound  $P_{\text{tot}}$  when the optimum in (4.36) and (4.37) is achieved. In accordance with the water-filling algorithm, it is easy to see that equal power allocation among  $\{s_{\psi,k,q} : q \in [M_{\psi,k}]\}$  for  $\forall \psi \in \Psi$  and  $k \in [Q]$  is optimal, which means that we get  $P_{\psi,k,q} = P_{\psi,k}/M_{\psi,k}$ . To solve the MMF problems in (4.36) and (4.37), we have the following theorem.

**Theorem 4.2.** *The optimal MMF-constrained optimal effective sum-rate  $\bar{R}_{ZF}^*$  is bounded as  $\tilde{R}_{ZF}^* \leq \bar{R}_{ZF}^* \leq \hat{R}_{ZF}^*$ , where  $\tilde{R}_{ZF}^*, \hat{R}_{ZF}^*$  are respectively the solutions to*

$$\sum_{\psi \in \Psi} \sum_{k \in [Q]} \frac{N_0 M_{\psi,k}}{\beta_{\psi,k}(L-M_\psi)} \left( \exp \left( \frac{\tilde{R}_{ZF}^*}{\xi_{G,Q} G Q M_{\psi,k}} \right) - 1 \right) = P_{\text{tot}} \quad (4.38)$$

$$\sum_{\psi \in \Psi} \sum_{k \in [Q]} \frac{N_0 M_{\psi,k}}{\beta_{\psi,k}(L-M_\psi+1)} \left( \exp \left( \frac{\hat{R}_{ZF}^*}{\xi_{G,Q} G Q M_{\psi,k}} \right) - 1 \right) = P_{\text{tot}}. \quad (4.39)$$

In the symmetric case of  $M_{\psi,k} = M, \forall \psi \in \Psi, k \in [Q]$ , the optimal effective sum-rate is lower and upper bounded respectively by

$$\tilde{R}_{ZF}^* = \xi_{G,Q} G Q M \ln \left( 1 + \frac{P_{\text{tot}}(L-QM)}{M N_0 \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right) \quad (4.40)$$

$$\hat{R}_{ZF}^* = \xi_{G,Q} G Q M \ln \left( 1 + \frac{P_{\text{tot}}(L-QM+1)}{M N_0 \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right). \quad (4.41)$$

*Proof.* As was the case in Theorem 4.1, here also all  $GQ$  users accept the same lower-bound on their effective rate under the optimal power allocation in (4.36). Considering that  $\sum_{\psi \in \Psi} \sum_{k \in [Q]} P_{\psi,k} = P_{\text{tot}}$  and also considering Proposition 4.1, we can easily obtain (4.38). The derivation of (4.39) follows in similar steps.  $\square$

**Corollary 4.3.** *Given ZF-based precoding at the BS and given  $M$ -antenna receivers, the optimal effective gain  $\mathcal{G}_{ZF}^*$  of our ZF-based vector coded caching, is bounded by  $\tilde{\mathcal{G}}_{ZF}^* \leq \mathcal{G}_{ZF}^* \leq \hat{\mathcal{G}}_{ZF}^*$ , where*

$$\tilde{\mathcal{G}}_{ZF}^* \triangleq \frac{\max_{Q \in [Q_{\text{max}}]} \tilde{R}_{ZF}^*(G, Q)}{\max_{Q' \in [Q_{\text{max}}]} \tilde{R}_{ZF}^*(1, Q')} = \frac{\max_{Q \in [Q_{\text{max}}]} \xi_{G,Q} G Q M \ln \left( 1 + \frac{P_{\text{tot}}(L-QM)}{M N_0 \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right)}{\max_{Q' \in [Q_{\text{max}}]} \xi_{1,Q'} Q' M \ln \left( 1 + \frac{P_{\text{tot}}(L-Q'M+1)}{M N_0 \sum_{k' \in [Q']} \beta_{k'}^{-1}} \right)} \quad (4.42)$$

$$\hat{\mathcal{G}}_{ZF}^* \triangleq \frac{\max_{Q \in [Q_{\text{max}}]} \hat{R}_{ZF}^*(G, Q)}{\max_{Q' \in [Q_{\text{max}}]} \hat{R}_{ZF}^*(1, Q')} = \frac{\max_{Q \in [Q_{\text{max}}]} \xi_{G,Q} G Q M \ln \left( 1 + \frac{P_{\text{tot}}(L-QM+1)}{M N_0 \sum_{\psi \in \Psi} \sum_{k \in [Q]} \beta_{\psi,k}^{-1}} \right)}{\max_{Q' \in [Q_{\text{max}}]} \xi_{1,Q'} Q' M \ln \left( 1 + \frac{P_{\text{tot}}(L-Q'M)}{M N_0 \sum_{k' \in [Q']} \beta_{k'}^{-1}} \right)}, \quad (4.43)$$

and where  $Q_{\max} = \frac{L-1}{M}$ .

*Proof.* We can derive Corollary 4.3 directly from Theorem 4.2.  $\square$

## 4.4 Numerical Results

This section presents various numerical results that validate our analysis as well as provide insightful comparisons. We here consider relatively low mobility users, and assume a coherence time of 0.05 s and a coherence bandwidth of 300 kHz, corresponding to a coherence block of 15000 symbols. We consider CSI pilot length of  $\beta_{\text{tot}} = 10$ , which is expected to be sufficient for providing near-perfect CSI at both the BS and the users [109]. The AWGN spectral density is considered to be  $-174$  dBm/Hz, and the spectrum bandwidth for each user is 20 MHz. We generate 1000 realizations of users' locations, based on the assumption of uniformly-distributed users across the cell. We consider the Macro-cell scenario with an inner radius of 35 meters and an outer radius of 500 meters, as well as consider the Micro-cell case with an inner radius of 10 meters and an outer radius of 100 meters. Assuming a carrier frequency of 2 GHz, in the Macro-cell case, the pathloss is modeled as  $\beta_{\psi,k} = l_0 r_{\psi,k}^{-\eta_0}$  [84], where  $r_{\psi,k}$  is the distance between the BS and  $U_{\psi,k}$ , where  $\eta_0 = 3.76$  is the path-loss exponent, and where  $l_0 = 10^{-3.53}$  regulates the channel attenuation at 35 meters. For the Micro-cell scenario, we note that the pathloss model can often differ when considering delivery distance between 10 and 40 meters, compared to when considering a distance in the [40, 100] meters range. For simplicity though, we here use the pathloss model for [10, 40] meters over the entire delivery range in the Micro-cell, and thus consider  $l_0 = 10^{-3.7}$  and  $\eta_0 = 3$  (cf. [115, Table II]).

Fig. 4.1 plots the effective rate for a typical user versus  $P_{\text{tot}}$  for different numbers of receiver antennas, where both simulated and analytical results are presented side by side in order to validate the accuracy of derived expressions. In Fig. 4.1, the analytical result for the BD-MRC is derived based on Lemma 4.1, while we derive the BD-MRC approximation by substituting the approximation in (C.9) into Lemma 4.1. It is obvious that the effective rate increases as  $M$  grows under both ZF and BD-MRC schemes because the BS can send more symbols at a time. It is also worth noting that the performance gap between ZF and BD-MRC schemes decreases with decreasing  $M$ , and almost vanishes for  $M = 2$  (a practical receiver-antenna numbers). Fig. 4.2 shows the effective sum-rate versus  $P_{\text{tot}}$ , where the setting is the same as in Fig. 4.1, with the exception though that we consider MMF power allocation. In Fig. 4.2, we plot the analytical result for the BD-MRC from Lemma 4.1 and after using the one-dimensional dichotomous search for  $R_{\text{BD-MRC}}^*$ , while the BD-MRC approximation plotted is from Lemma 4.2. In the simulation results, we use the built-in function “fminmax” in MATLAB to numerically solve the MMF optimization in (4.16). We can see that the effective coded caching gain is an increasing function w.r.t.  $P_{\text{tot}}$ , while it is a decreasing function w.r.t.  $M$ . It is also obvious that the ZF variant precoding always provides lower bounds for both effective sum-rate and gain.

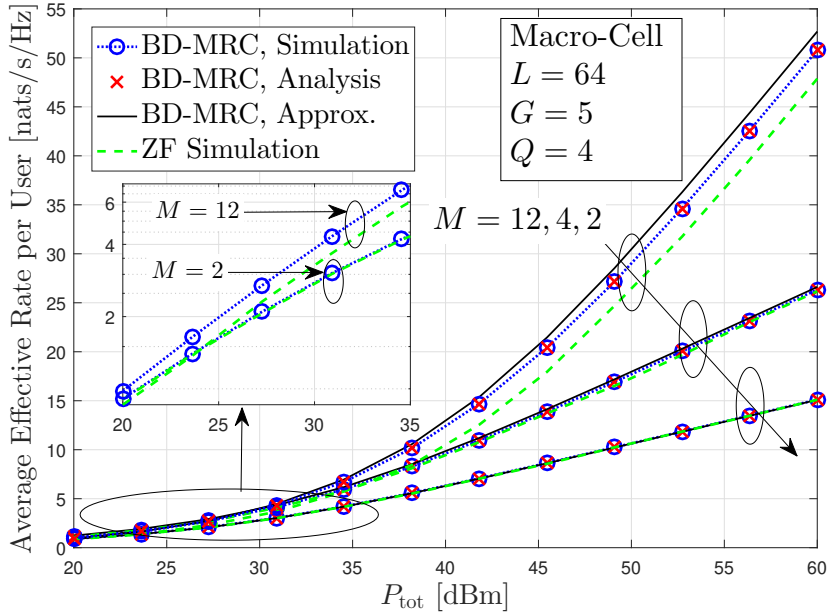


Figure 4.1: Effective rate versus  $P_{\text{tot}}$  in the Macro-cell setting for  $J = M$  with equal-power allocation.

Fig. 4.3 shows both the effective sum-rate and the corresponding gain versus  $P_{\text{tot}}$  for  $L = 128$  and  $M = 4$ . Compared to the performance of  $M = 4$  in Fig. 4.2, we observe modest increases in both the effective sum-rate and gain when we increase the number of transmit antennas from 64 to 128. The delivery performance with the same DoF (i.e.,  $Q' = GQ$ ) in vector coded caching and in the cacheless scenario is plotted in Fig. 4.4; this corresponds to the same setting (except for the values of  $Q$  and  $Q'$ ) as in Fig. 4.3. In contrast to the increasing effective gain w.r.t.  $P_{\text{tot}}$  that we experience in Fig. 4.3, the effective gain in Fig. 4.4 decreases with  $P_{\text{tot}}$ , and eventually approaches 1 (all gains disappear) as  $P_{\text{tot}}$  increases.

We then plot the optimal effective gain versus  $P_{\text{tot}}$ , where in Fig. 4.5 we do so for  $G = 4$  and in Fig. 4.6 we do so for  $G = 6$ . In both cases, we do so for  $L = 128$  and  $M = 4$ .

One note is that while the gain from vector coded caching (over the corresponding traditional (cacheless) MU-MIMO scenario, is very substantial, it is indeed less than in the symmetric case (same pathloss) that we have seen in Chapter 3. This is mainly due to the heavy worst-user effect (or the near-far effect) which considerably limits the effective sum-rate, especially in a large cell (e.g., a Macro-cell), even after optimal power allocation. What makes things worse is that the users are uniformly distributed within this Macro-cell, which implies that most of the users tend to locate on the cell edge. For example, again for the Macro-cell case, we see that 64.32% of the users are more than 300 meters away from the BS. For the typical BS transmit power of  $P_t = 40$  dBm in a Macro-cell, the received (average) SNR at these users is below 12.55 dB (cf. Table 4.1).

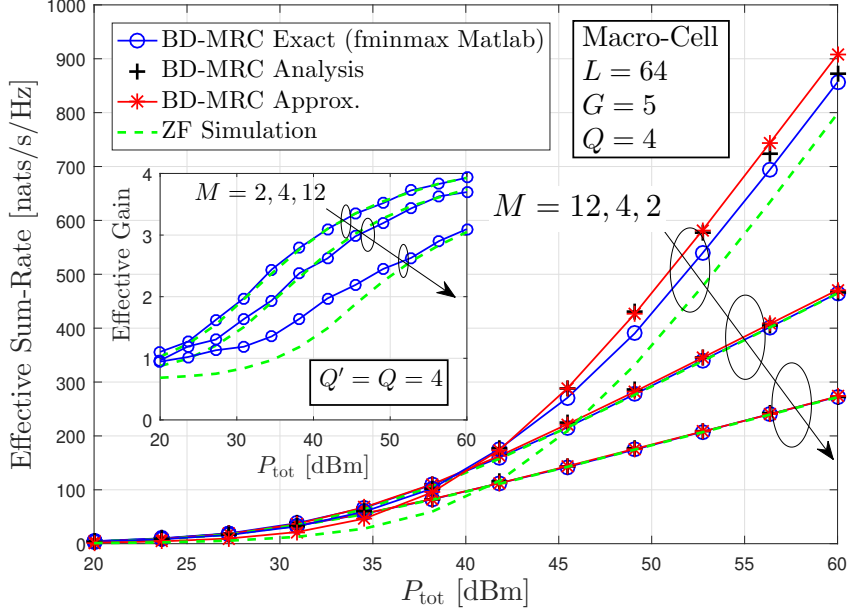

 Figure 4.2: Effective rate versus  $P_{\text{tot}}$  in the Macro-cell setting under MMF.

 Table 4.1: User Percentage and Receiver SNR Range in a Macro Cell with  $D_1 = 35$ ,  $D_2 = 500$ ,  $\eta = 3.76$ , and  $l_0 = 10^{-3.53}$ 

Delivery distance	$\geq 50$ m	$\geq 100$ m	$\geq 200$ m	$\geq 300$ m	$\geq 400$ m	$\geq 450$ m
Percentage in total users	99.49%	96.47%	84.41%	64.32%	36.18%	19.09%
Received SNR = $\frac{P_t}{N_0} l_0 r^{-\eta}$	$\leq P_t$ [dB] + 31.81	$\leq P_t$ [dB] + 20.49	$\leq P_t$ [dB] + 9.17	$\leq P_t$ [dB] + 2.55	$\leq P_t$ [dB] - 2.15	$\leq P_t$ [dB] - 4.07

Things are very different in the Micro-cell setting (cf. Fig. 4.7) where the effective gain of the ZF-based vector coded caching is again very notable. For example, the recorded gain is 410% for the reasonable BS transmit power of  $P_{\text{tot}} = 33$  dBm in a Micro-cell.

## 4.5 Conclusions

We have investigated vector coded caching under various realistic considerations such as having variable path-loss, multi-antenna receivers, practical decoders, and MMF. Specifically, we have derived analytical expressions of the effective sum-rate and the effective gain under BD-MRC and ZF schemes respectively, as well as we have provided various closed-form expressions under various simplifying assumptions. Numerical results validate very clearly the tightness of the derived expressions, revealing again notable effective gains that vector coded caching can provide in realistic scenarios. These gains are particularly high in Micro-cell environments. For example, we have seen that vector coded caching can offer more than a 410% boost in the overall throughput over the conventional (cacheless) MU-MIMO system, under the aforementioned realistic assumptions. We note that all recorded gains attributed to vector coded caching, are taken to reflect the

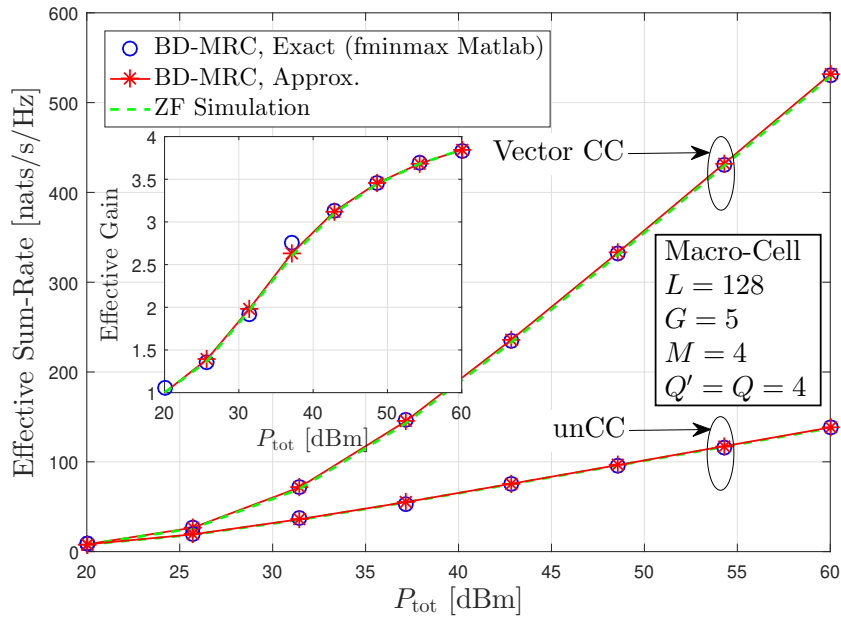


Figure 4.3: Effective rate versus  $P_{\text{tot}}$  in a Macro-cell under MMF.

improvement over *optimized* traditional MU-MIMO systems (where for example, such traditional MU-MIMO system is optimized w.r.t. the operational multiplexing gain).

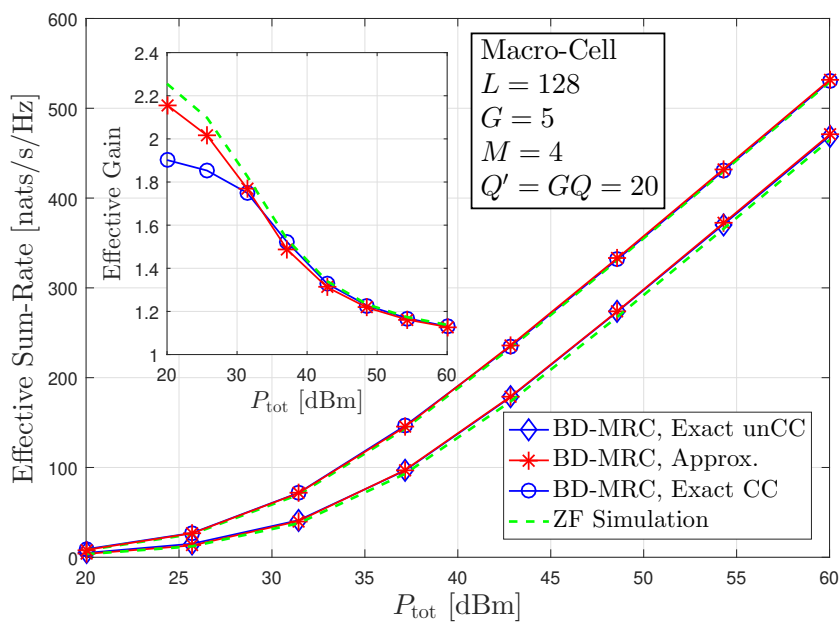


Figure 4.4: Effective rate versus  $P_{\text{tot}}$  in a Macro-cell under MMF.

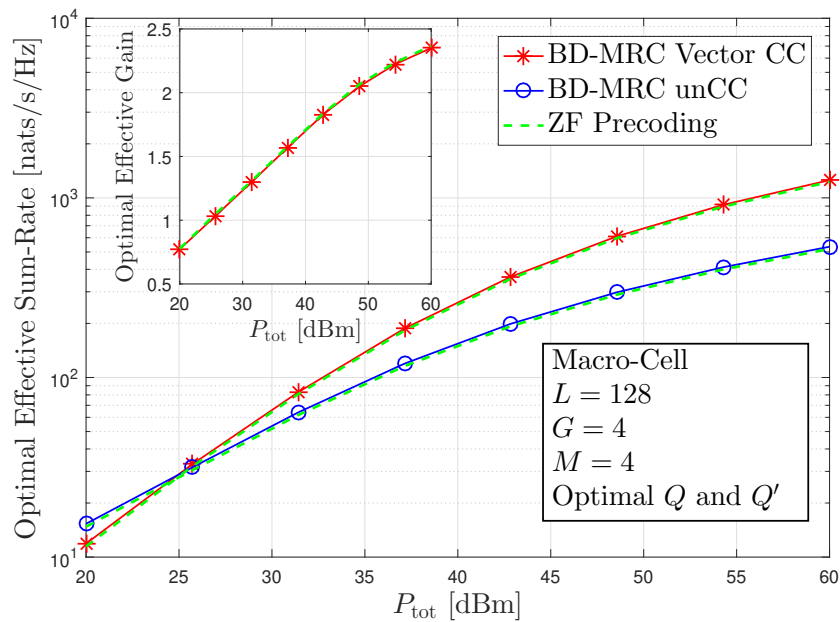


Figure 4.5: Optimal effective gain versus  $P_{\text{tot}}$  in a Macro-cell under MMF.

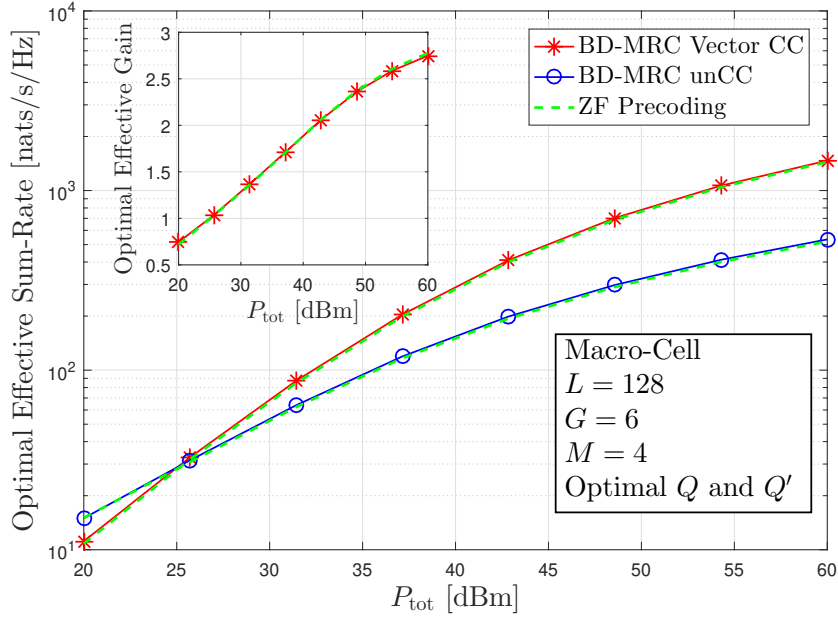


Figure 4.6: Optimal effective gain versus  $P_{\text{tot}}$  in a Macro-cell under MMF.

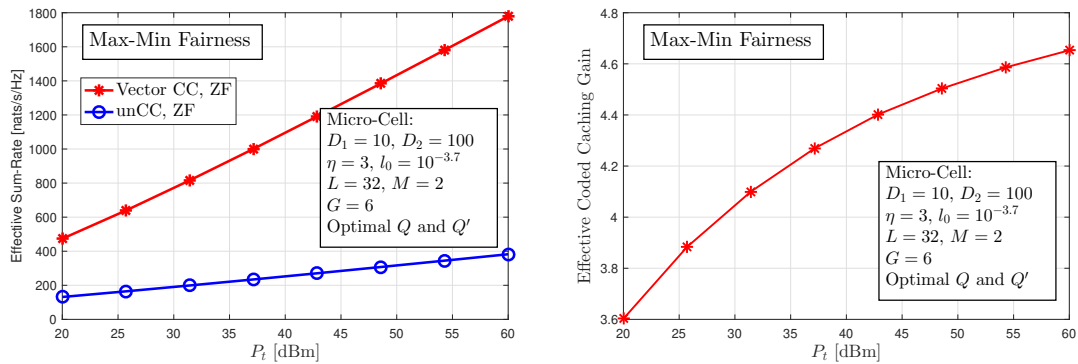


Figure 4.7: Delivery Performance of vector coded caching in a Micro-cell



## Chapter 5

# ACC-aided Land Mobile Satellite System

In this chapter, we investigate the performance of coded caching in land mobile-satellite (LMS) systems, where a satellite station with full access to a content library, serves  $K$  cache-aided land users. As is known, LMS systems experience low-to-moderate SNR due to their long propagation distances, and thus the traditional coded caching gains may suffer in these more realistic SNR regimes. In this thesis, these gains are first analyzed for the traditional MN solution, and then for the new ACC solution. In particular, we first analyze the extent to which the MN coded caching gains are preserved in LMS systems. We model the satellite-terrestrial channels through the widely adopted Shadowed-Rician fading model, and we show that — interestingly, and unlike what we saw before in the Rayleigh fading scenario — the coded caching gains are partially preserved even in the low-SNR limit due to the existence of line-of-sight (LOS) components. In particular, here MN coded caching will allow us — under realistic assumptions — to double the throughput even at low SNR. These results illustrate the potential of MN coded caching in LMS systems but also motivate us to further investigate the delivery performance of our new ACC scheme, which will be here shown to boost the effective MN gain, under realistic assumptions, by a multiplicative factor of approximately 2.5. For the two considered coded caching schemes, MN and ACC, we derive the corresponding analytical expressions for the average rate and for the effective gain, as well as simplify these results in the low SNR regime. For ACC, we also provide a result in the regime of large  $K$ , which is nonetheless shown to be robust even when  $K$  is moderate to small. To provide a general analysis of the ACC performance, we further consider a mixture Gamma (MG) channel model which is known to include most of existing and practical channel models. Finally, these derived expressions are extensively validated using Monte-Carlo simulations.

### 5.1 Introduction

The use of LMS systems has been rapidly growing in recent years [116], as they provide persistent and now lower-cost services, while being able to benefit from an enormous

coverage area. Despite this progress, satellite is still somewhat limited compared to fiber-based content delivery services, and for this reason we here seek to explore how such systems can benefit from coded caching. Some works that jointly consider satellite and cache-aided terrestrial networks can be found in [117–122]. As we have discussed, coded caching is an attractive method to exploit caches on the ground, as it is able to provide, in theory, a considerable multiplicative performance boost [1] which cannot be found in uncoded caching approaches.

*Worst-User Bottleneck:*

However, the satellite communications setting exacerbates the *worst-user bottleneck* that results from the fact that the achievable rate in multicast transmissions is constrained by the user with the worst channel state among the  $K\gamma+1$  simultaneously served users [23]. *This bottleneck is, as mentioned above, unfortunately exacerbated in satellite setting where the SNR is decreased.* For such SNR regimes, under some channel assumptions, we have seen the gain from XOR-based (MN) coded multicasting to vanish [31, 51, 52]. In our context here, we note that, for example, the received SNR at a terrestrial user is typically below 10 dB in home TV broadcasts from geosynchronous (GEO) satellites with 36 MHz of bandwidth and several hundred of watts of power [123]. Our aim here is to explore this bottleneck, in the satellite setting which naturally accepts different channel statistics [71] that reflect the existence of LOS components and shadowing.

The remainder of this chapter is organized as follows: Section 5.2 defines the system model and the studied problem. The average rates and effective gains of MN and ACC are investigated respectively in Sections 5.3 and 5.4, where we also derive several tight approximations. Some numerical results and comparisons are presented in Section 5.5, and finally Section 5.6 concludes the chapter.

## 5.2 System Model

We consider a scenario in which a satellite, such as a GEO satellite used for transmitting video on demand, having full access to a library  $\mathcal{F}$  with  $N$  equal-size files<sup>1</sup>, serves a set of  $K$  cache-aided land users, where each user requests a different file from the intended library  $\mathcal{F}$ . To speed up delivery, these  $K$  users pre-store a fraction  $\gamma$  of the library during off-peak hours. As discussed in Chapter 2, due to the finite file size constraint, we expect to have only some  $\Lambda \ll K$  different cache states when implementing coded caching, which effectively forces  $B = K/\Lambda$  users to share the cache state, i.e., to store identical content.

We recall that coded caching typically follows a clique-based approach such that the transmission is divided into transmission stages that experience a clique-based side information pattern. This implies that any desired subfile of some served user can be found in the cache of every other user involved in that same transmission stage<sup>2</sup>. We also recall that for a transmission stage to a specific user-group set  $\Psi$ , the received signal

<sup>1</sup>This access is probably maintained due to the fact that the satellite is connected to a terrestrial satellite gateway via a feeder link which can often sustain extremely high rates. This satellite gateway has a wired connection to the core network and thus access to the library [70] is expected to be seamless.

<sup>2</sup>We refer to Section 2.2.1 for more details about the clique-based cache placement and content delivery in ACC and MN.

at the  $b$ -th user of group  $g \in \Psi$  (i.e., to user  $U_{g,b}$ ) takes the form

$$y_{g,b} = h_{g,b}X_{ts} + z_{g,b}, \quad (5.1)$$

where  $X_{ts}$  is the transmitted signal symbol, satisfying an average power constraint  $\mathbb{E}\{|X_{ts}|^2\} = \rho$ , where  $z_{g,b}$  is the unit-power AWGN, and where  $h_{g,b}$  is the channel gain between  $S$  and  $U_{g,b}$ . Obviously,  $\text{SNR}_{g,b} = \rho|h_{g,b}|^2$  is the instantaneous SNR at  $U_{g,b}$ .

### 5.2.1 LMS Channel Model

In order to accurately describe the fluctuation of the signal envelope, we consider the widely adopted Rician-Shadowed fading [71] to model the satellite-terrestrial channel. This very general model can be nicely calibrated to capture both fixed and mobile land terminals, and it can be applied for all types of orbits and for a variety of frequency bands including S-band, L-band, Ku-band, and Ka-band [71]. In Rician-Shadowed fading, the instantaneous lowpass-equivalent complex signal envelope is written as

$$\mathcal{S} = \mathcal{E} \exp(j\varsigma) + \mathcal{V} \exp(j\varsigma_0), \quad (5.2)$$

where  $j \triangleq \sqrt{-1}$  is the imaginary unit,  $\varsigma$  and  $\varsigma_0$  are the stationary random phase with uniform distribution over  $[0, 2\pi)$  and the deterministic phase of the LOS component, respectively. In (5.2),  $\mathcal{E}$  and  $\mathcal{V}$  are the amplitudes of the scatter and the LOS component, respectively. Specifically,  $\mathcal{E}$  is modeled by a Rayleigh distribution, while  $\mathcal{V}$  follows a Nakagami- $m$  distribution. Therefore, the PDFs of  $\mathcal{E}$  and  $\mathcal{V}$  are respectively

$$f_{\mathcal{E}}(x) = \frac{x}{b_0} \exp\left(-\frac{x^2}{2b_0}\right), \quad f_{\mathcal{V}}(x) = \frac{2m_0^{m_0}}{\Gamma(m_0)\aleph^{m_0}} x^{2m_0-1} \exp\left(-\frac{m_0x^2}{\aleph}\right),$$

where  $2b_0$  and  $\aleph$  are the means of  $\mathcal{E}^2$  and  $\mathcal{V}^2$  respectively, and where  $m_0 \triangleq \frac{\mathbb{E}^2\{\mathcal{V}^2\}}{\text{Var}\{\mathcal{V}^2\}}$  reflects the (average) obstruction of the LOS component (i.e., the blockage of the LOS by buildings, trees, hills, etc.). We refer to [71] for more details.

Although the arbitrary moment of  $|\mathcal{S}|$  has been derived in [71, Eq. (5)], we here present more concise closed-form expressions for the average and the variance of  $|\mathcal{S}|^2$  in Proposition 5.1 to facilitate the analysis in Sections 5.3 and 5.4.

**Proposition 5.1.** *The mean and variance of the channel power gain  $|\mathcal{S}|^2$  in Rician-Shadowed fading channels, are respectively of the form*

$$\mathbb{E}\{|\mathcal{S}|^2\} = 2b_0 + \aleph, \quad (5.3)$$

$$\text{Var}\{|\mathcal{S}|^2\} = 4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0}. \quad (5.4)$$

*Proof.* See Appendix D.1 □

## 5.2.2 Adopted Assumptions and Preliminaries

To facilitate the analysis, we assume the same channel statistics (i.e., same  $m_0$ ,  $b_0$  and  $\aleph$ ) among the  $K$  users, which is reasonable when the users are uniformly distributed within a disk [124]. We consider that the users are located within a disk of radius equal to several kilometers [53]. As the radius is negligible compared to the height of the GEO satellite, we assume that all the users have the same pathloss. Therefore, we consider *statistically symmetric users*. Furthermore, as a small difference in  $m_0$  reflects a similar (average) obstruction of the LOS component, we consider that  $m_0$  is a positive integer for mathematical tractability, which is a simplifying assumption that is widely adopted in many existing works [120, 125].

GEO satellites are static with respect to an observer from Earth, which makes the Doppler spread negligible for static terrestrial users [125]. Thus, we assume that the coherence time is large, and we do not consider an ergodic channel. Instead, we assume that the channel experiences quasi-static fading, which generally comes about in the presence of longer coherence periods and shorter latency constraints, and which nicely models low-mobility scenarios which nicely capture coded-caching use-cases where slowly moving or stationary users are consuming video content.

In the following, let us present the PDF of  $\text{SNR}_{g,b}$  for a positive integer  $m_0$ . Upon defining  $\Xi(i) \triangleq \binom{m_0-1}{i} \frac{\delta_0^i}{i!}$ ,  $\alpha_0 \triangleq \left(\frac{2b_0m_0}{2b_0m_0+\aleph}\right)^{m_0}$ ,  $\varphi_0 \triangleq \frac{1}{2b_0}$ , and  $\delta_0 \triangleq \frac{\aleph}{2b_0(2b_0m_0+\aleph)}$ , we have the following proposition.

**Proposition 5.2.** *For any integer  $m_0 \geq 1$ , the PDF of the SNR at  $U_{g,b}$  over Rician-Shadowed fading channels can be simplified as*

$$f_{\text{SNR}_{g,b}}(x) = \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} x^i \exp\left(-\frac{\varphi_0 - \delta_0}{\rho} x\right), \quad x \geq 0. \quad (5.5)$$

*Proof.* To see this, we simply note that in Rician-Shadowed fading, the PDF of  $\text{SNR}_{g,b}$  is given by [71, Eq. (6)]

$$f_{\text{SNR}_{g,b}}(x) = \frac{\alpha_0}{\rho} \exp\left(-\frac{\varphi_0}{\rho} x\right) \cdot {}_1F_1\left(m_0; 1; \frac{\delta_0}{\rho} x\right), \quad x \geq 0, \quad (5.6)$$

where  ${}_1F_1(\cdot; \cdot; \cdot)$  denotes the generalized hypergeometric function [76]. For  $m_0 = 1, 2, \dots$ , by using [126, Eq. (24)], the PDF of  $\text{SNR}_{g,b}$  can be simplified as (5.5).  $\square$

By integrating the PDF expression in (5.5), we can easily obtain the corresponding CDF of  $\text{SNR}_{g,b}$  as [125, Eq. (13)]

$$F_{\text{SNR}_{g,b}}(x) = 1 - \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} \sum_{j=0}^i \frac{i!}{j!} \left(\frac{\rho}{\varphi_0 - \delta_0}\right)^{i+1-j} x^j \exp\left(-\frac{\varphi_0 - \delta_0}{\rho} x\right). \quad (5.7)$$

To facilitate the analysis in Sections 5.3 and 5.4, we present a simplified CDF of  $\text{SNR}_{g,b}$  (or equivalently, the outage probability) in Proposition 5.3 that follows.

**Proposition 5.3.** *The CDF of  $\text{SNR}_{g,b}$ , or equivalently the outage probability at  $U_{d,b}$ , over Rician-Shadowed fading channels, can be simplified as*

$$F_{\text{SNR}_{g,b}}(x) = 1 - \alpha_0 \exp\left(-\frac{\varphi_0 - \delta_0}{\rho}x\right) \sum_{j=0}^{m-1} \left( \sum_{\ell=j}^{m-1} \frac{\Xi(\ell)\ell!}{j!} \frac{\rho^{-j}}{(\varphi_0 - \delta_0)^{\ell-j+1}} \right) x^j. \quad (5.8)$$

*Proof.* By exchanging the summation orders of  $i$  and  $j$  in (5.7) and after some simple mathematical manipulations, we can easily derive (5.8).  $\square$

### 5.3 Average Rate and Effective Gain of the MN Scheme

In this section, we analyze the average rate of the MN scheme in order to present some insights about its performance in LMS systems, as well as to provide a benchmark for the ACC scheme.

As  $B = 1$  in the ACC algorithm yields  $\Lambda$ -MN (cf. Chapter 2), the achievable rate in this case is (cf. Section 2.3.1)

$$R^{\text{MN}} = G \ln(1 + \min_{g \in \Psi} \{\text{SNR}_g\}) = G \ln(1 + \text{SNR}_{\text{MN}}) \quad \text{nats/s/Hz}, \quad (5.9)$$

where we simplify the subscript of SNR because  $B = 1$ , and  $\text{SNR}_{\text{MN}} \triangleq \min_{g \in \Psi} \{\text{SNR}_g\}$  reflects the worst-user effect in multi-user multicasting.

Before presenting the average rate of the MN scheme, let us use  $\Gamma(\cdot, \cdot)$  to denote the upper incomplete Gamma function [76] and  $\binom{G}{\hbar_1, \hbar_2, \dots, \hbar_{m_0}} \triangleq \frac{G!}{\hbar_1! \hbar_2! \dots \hbar_{m_0}!}$  to denote the multinomial coefficient for  $m_0$  non-negative integers  $\hbar_1, \hbar_2, \dots, \hbar_{m_0}$ .

**Lemma 5.1.** *The closed-form expression for the average rate of the MN scheme over Rician-Shadowed fading channels is*

$$\begin{aligned} \bar{R}^{\text{MN}} = & G \alpha_0^G \sum_{\hbar_1 + \dots + \hbar_{m_0} = G} \binom{G}{\hbar_1, \dots, \hbar_{m_0}} \left[ \prod_{t=0}^{m_0-1} \left( \sum_{\ell=t}^{m_0-1} \frac{\Xi(\ell)\ell!}{t!} \frac{\rho^{-t}}{(\varphi_0 - \delta_0)^{\ell-t+1}} \right)^{\hbar_{t+1}} \right] \\ & \times \exp\left(\frac{G(\varphi_0 - \delta_0)}{\rho}\right) \left( \sum_{t=0}^{m_0-1} t \hbar_{t+1} \right)! \cdot \Gamma\left(-\sum_{t=0}^{m_0-1} t \hbar_{t+1}, \frac{G(\varphi_0 - \delta_0)}{\rho}\right), \quad (5.10) \end{aligned}$$

where the summation is over all possible  $m$  non-negative integer combinations  $(\hbar_1, \dots, \hbar_m)$  that satisfies  $\hbar_1 + \dots + \hbar_m = G$ .

*Proof.* See Appendix D.2.  $\square$

Next in Corollary 5.1 we present the average rate of simple TDM transmission (no coded caching), corresponding to the average rate of the MN scheme when  $G = 1$  (cf. Section 2.3.1).

**Corollary 5.1.** *The average rate of the uncoded TDM scheme over Rician-Shadowed fading channels is*

$$\bar{R}^{\text{TDM}} = \alpha_0 \exp\left(\frac{\varphi_0 - \delta_0}{\rho}\right) \sum_{j=0}^{m_0-1} \left( \sum_{\ell=j}^{m_0-1} \frac{\Xi(\ell)\ell!\rho^{-j}}{(\varphi_0 - \delta_0)^{\ell-j+1}} \right) \Gamma\left(-j, \frac{\varphi_0 - \delta_0}{\rho}\right). \quad (5.11)$$

*Proof.* This result can be easily derived by setting  $G = 1$  in (5.10).  $\square$

To obtain insights on the performance of the MN scheme, we further simplify the average rate expression from Lemma 5.1 by exploring the limit of  $\rho \rightarrow 0$ .

**Corollary 5.2.** *The low SNR approximation of the average rate of the MN scheme takes the form*

$$\begin{aligned} \bar{R}^{\text{MN}} = \rho \frac{\alpha_0^G}{\varphi_0 - \delta_0} \sum_{\hbar_1 + \dots + \hbar_m = G} \binom{G}{\hbar_1, \dots, \hbar_m} & \left[ \prod_{t=0}^{m-1} \left( \sum_{\ell=t}^{m-1} \frac{\Xi(\ell)\ell!}{t!} \left( \frac{1}{\varphi_0 - \delta_0} \right)^{\ell+1} \right)^{\hbar_{t+1}} \right] \\ & \times \left( \sum_{t=0}^{m_0-1} t \hbar_{t+1} \right)! G^{-\sum_{t=0}^{m_0-1} t \hbar_{t+1}} + o(\rho). \end{aligned} \quad (5.12)$$

*Proof.* For  $\rho \rightarrow 0$ , by using  $\Gamma(s, x) \rightarrow x^{s-1} \exp(-x)$  as  $x \rightarrow \infty$  in (5.10), we can easily derive this result after some simple mathematical manipulations.  $\square$

For the low-SNR approximation of the uncoded TDM scheme, in view of  $\ln(1+x) = x + o(x)$  as  $x \rightarrow 0$  and given Proposition 5.1, we can have that

$$\bar{R}^{\text{TDM}} = \mathbb{E} \{ \ln(1 + \text{SNR}_g) \} = \mathbb{E} \{ \text{SNR}_g \} + o(\rho) = (2b_0 + \aleph)\rho + o(\rho), \text{ as } \rho \rightarrow 0. \quad (5.13)$$

Combining (5.12) and (5.13), we derive the effective gain of the MN scheme in the low-SNR limit as

$$\begin{aligned} \lim_{\rho \rightarrow 0} \mathcal{G}_{\text{MN}} &= \lim_{\rho \rightarrow 0} \frac{\bar{R}^{\text{MN}}}{\bar{R}^{\text{TDM}}} \\ &= \frac{\alpha_0^G}{(\varphi_0 - \delta_0)(2b_0 + \aleph)} \sum_{\hbar_1 + \dots + \hbar_{m_0} = G} \binom{G}{\hbar_1, \dots, \hbar_{m_0}} \\ & \quad \times \left[ \prod_{t=0}^{m_0-1} \left( \sum_{\ell=t}^{m-1} \frac{\Xi(\ell)\ell!}{t!(\varphi_0 - \delta_0)^{\ell+1}} \right)^{\hbar_{t+1}} \right] \left( \sum_{t=0}^{m_0-1} t \hbar_{t+1} \right)! G^{-\sum_{t=0}^{m_0-1} t \hbar_{t+1}}. \end{aligned} \quad (5.14)$$

Fig. 5.1 plots, in the limit of low SNR, the effective gain of the MN scheme (i.e., (5.14)) versus the nominal gain  $G$  (from 2 to 10) in four typical LMS fading scenarios listed in Table 5.4. We can see how the effective gain increases for every scenario except for the frequent heavy shadowing case. Let us recall that this is in contrast to the case with Rayleigh fading, where the effective gain of the MN scheme entirely vanished in the low-SNR limit (cf. Proposition 2.2). The steeper gain boost for the infrequent light

shadowing and the average shadowing cases shows that the MN scheme (and XOR-based coded caching) is advantageous mainly when the LOS component is not heavily obstructed. However, we can also observe that even for the best LMS channel statistics corresponding to infrequent light shadowing, the effective MN gain in the low-SNR limit is just close to 2 when  $G = 6$  and stays below 2.5 when for (the rather unrealistic case of)  $G = 10$ . We can readily conclude thsu, that even though MN remains pertinent in such lower SNR settings, indeed the MN scheme loses most of the (high-SNR) nominal gain in the low-SNR governed LMS systems.

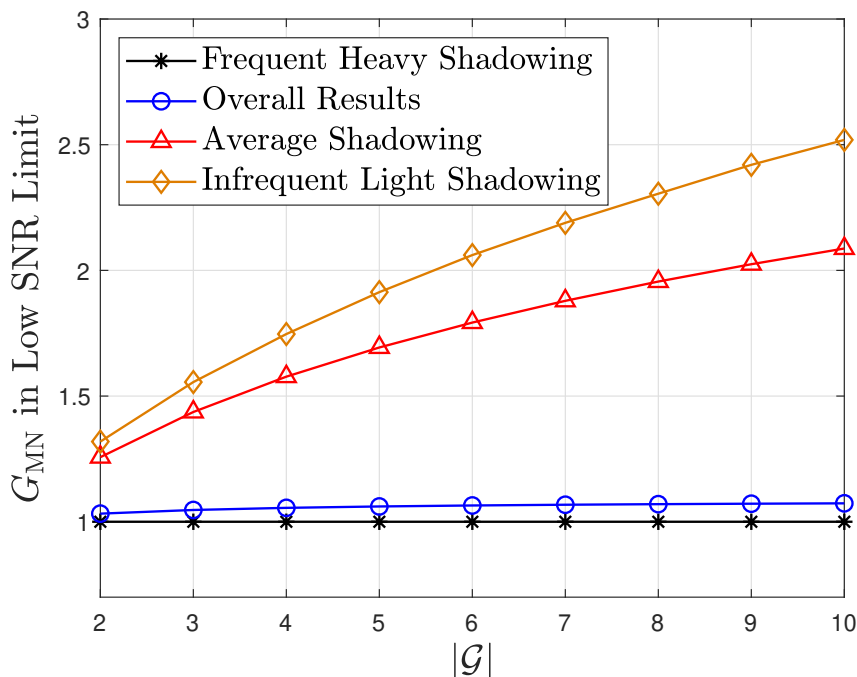


Figure 5.1: Effective gain of MN coded caching versus  $G$  in low-SNR limit.

## 5.4 Average Rate and Effective Gain of the ACC Scheme

In this section, we analyze the average rate and the effective gain of the ACC scheme in LMS systems. We first derive the analytical expression for the average rate, and then perform a large- $K$  approximation to facilitate the analysis on the effective gain. In the end, we extend the large- $K$  approximation to more general fading scenarios.

### 5.4.1 Exact Analytical Expression

In the following,  $j = \sqrt{-1}$  denotes the imaginary unit,  $\text{Im}\{\cdot\}$  is the operator of taking the imaginary part of a complex number, and  $\mathcal{U}(\cdot, \cdot, \cdot)$  represents the confluent hypergeometric

function of the second kind [76]. We have the following lemma for the average rate of the ACC scheme.

**Lemma 5.2.** *The exact average rate of the ACC scheme over Rician-Shadowed fading channels takes the form*

$$\begin{aligned} & \bar{R}^{\text{ACC}} \\ &= \frac{G}{B} \int_0^\infty \left( \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \text{Im} \left\{ \frac{\exp(-jxt)}{t} \left[ \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} i! \cdot \mathcal{U} \left( i+1, 2+i+jt, \frac{\varphi_0 - \delta_0}{\rho} \right) \right]^B \right\} dt \right)^G dx. \end{aligned} \quad (5.15)$$

*Proof.* See Appendix D.3.  $\square$

Lemma 5.2 presents a general result for the average rate of the ACC algorithm, but the result includes a double-integral and a confluent hypergeometric function, which is hard to evaluate numerically or to draw insights from. For this reason, in the following, we explore the limiting behavior of the above expression, in the limit of large  $K$ .

#### 5.4.2 Large-User Approximation

The large- $K$  approximation is well justified by the fact that a satellite usually serves a huge number of users due to its large coverage area [127]. When the number  $\Lambda$  of cache states is fixed, this large  $K$  assumption also implies a large number  $B$  of users sharing the same cache state. To imagine how big  $B$  is in practice, let us consider a cache-aided communication system with  $\gamma = 10\%$  and  $10^8$  bytes in each library file. When the sub-file size is no smaller than 1000 bytes, there are about 40 dedicated caches. In this system setting, we have  $B = 20$  for  $K = 800$  and  $B = 50$  for  $K = 2000$ .

The derived result in Lemma 5.3 for approximating the average rate of the ACC algorithm involves the well-known Q-function  $Q(\cdot)$  appearing in the CDF of the normal distribution, and the extended generalized bivariate Meijer's G-function (EGBMGF)  $G_{\nu, \nu}^{\nu, \nu}(\cdot, \cdot)$  defined in [128]. We also define  $\mathcal{H}_G$  as the expectation of the maximum of  $G$  i.i.d. standard normal random variables. We refer to Lemma 2.6 for more information about  $\mathcal{H}_G$ .

**Lemma 5.3.** *When  $B$  is sufficiently large, the average rate of the ACC scheme in this considered cached-aided LMS system can be approximated as*

$$\bar{R}^{\text{ACC}} \approx G \left( \varrho_l - \sqrt{\frac{\sigma_l^2}{B}} \times \mathcal{H}_G \right), \quad (5.16)$$

where  $\varrho_l$  and  $\sigma_l^2$  are the mean and the variance of  $\ln(1 + \text{SNR}_{g,b})$  respectively. Specifically, we can express  $\varrho_l$  and  $\sigma_l^2$  respectively as

$$\varrho_l = \alpha_0 \exp \left( \frac{\varphi_0 - \delta_0}{\rho} \right) \sum_{j=0}^{m_0-1} \left( \sum_{\ell=j}^{m_0-1} \frac{\Xi(\ell) \ell! \rho^{-j}}{(\varphi_0 - \delta_0)^{\ell-j+1}} \right) \Gamma \left( -j, \frac{\varphi_0 - \delta_0}{\rho} \right), \quad (5.17)$$

$$\sigma_l^2 = \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{(\varphi_0 - \delta_0)^{i+1}} G_{1,0;2,2;2,2}^{0,1;1,2;1,2} \left( \frac{\rho}{\varphi_0 - \delta_0}, \frac{\rho}{\varphi_0 - \delta_0} \middle| \begin{matrix} -i & 1,1 \\ - & 1,0 \end{matrix} \middle| \begin{matrix} 1,1 \\ 1,0 \end{matrix} \right) - \varrho_l^2. \quad (5.18)$$

*Proof.* See Appendix D.4.  $\square$

When computing the mean and the variance of  $\ln(1 + \text{SNR}_{g,b})$  in Lemma 5.3, we need to implement the incomplete Gamma function and the EGBMGF, which can be indeed a very time-consuming implementation. Here, with respect to the result in Lemma 5.4, we propose a *low-SNR approximation* for the mean and the variance of  $\ln(1 + \text{SNR}_{g,b})$  based on the method proposed in [129–131], which has been also considered in Chapter 2 as a robust approximation for the average rate of the MN scheme over Rayleigh fading channels (cf. Lemma 2.2).

**Lemma 5.4.** *In the low SNR limit of  $\rho \rightarrow 0$ , we can robustly approximate the mean and the variance of  $\ln(1 + \text{SNR}_{g,b})$  over Rician-Shadowed fading channels respectively as*

$$\varrho_l \approx \ln \left( 1 + (2b_0 + \aleph)\rho \right) - \frac{\left( 4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0} \right) \rho^2}{2(1 + (2b_0 + \aleph)\rho)^2}, \quad (5.19)$$

$$\sigma_l^2 \approx \frac{\left( 4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0} \right) \rho^2}{(1 + (2b_0 + \aleph)\rho)^2} \left( 1 - \frac{\left( 4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0} \right) \rho^2}{4(1 + (2b_0 + \aleph)\rho)^2} \right). \quad (5.20)$$

*Proof.* See Appendix D.5.  $\square$

**Remark 5.1.** *In fact, (5.19) presents a robust low-SNR approximation for the average rate of the uncoded TDM transmission over Rician-Shadowed fading channels. Compared to the exact result in (5.11), this low-SNR approximation in (5.19) is very concise and much easier to implement because it entails no special/advanced function. Numerical results in Fig. 5.5 show that the accuracy of (5.19) to the exact result is very high (with an almost non-discernible gap) even in the medium-to-high SNR region.*

With the help of Lemma 5.3, we are able to analyze the effective coded caching. In the large- $B$  case, the effective gain of the ACC scheme can be approximated by

$$\mathcal{G}_{\text{ACC}} = \frac{\bar{R}_{\text{ACC}}}{\bar{R}_{\text{TDM}}} \approx G - G \frac{\mathcal{H}_G \sqrt{\sigma_l^2/B}}{\varrho_l}, \quad (5.21)$$

where the second part represents the gain loss due to fading and shadowing. Moreover, for a sufficiently large  $B$ , we will always have that  $\frac{\mathcal{H}_G \sqrt{\sigma_l^2/B}}{\varrho_l} < 1$  in (5.21), which means that  $\mathcal{G}_{\text{ACC}}$  is increasing as the nominal gain  $G$  increases.

For low SNR, we have  $\varrho \approx \mathbb{E}\{\text{SNR}_g\} = (2b_0 + \aleph)\rho$  and  $\sigma_l^2 \approx \text{Var}\{\text{SNR}_g\} = \left( 4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0} \right) \rho^2$ . The effective ACC gain in the low SNR and large  $B$  regime will then take the form

$$\mathcal{G}_{\text{ACC}} \approx G - G \frac{\mathcal{H}_G \sqrt{\left( 4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0} \right) / B}}{2b_0 + \aleph}, \quad (5.22)$$

which shows that  $\mathcal{G}_{\text{ACC}}$  is an increasing function with  $m_0$ , which is a parameter that reflects how severe the obstruction of LOS components is in the LMS channel. Specifically,  $m_0 = 0$  stands for complete obstruction of LOS components (as this might occur in dense urban areas), while  $m_0 \rightarrow \infty$  corresponds to open areas with no obstruction of LOS components. Having an intermediate positive finite  $m_0 < \infty$  represents the partial obstruction that is usually encountered in suburban and rural areas.

### 5.4.3 Extension to General Fading Channels

Lemma 5.3 shows that  $\rho_l$  and  $\sigma_l^2$  depend on the fading scenarios. To make the large- $B$  approximation more general, let us extend the result in Lemma 5.3 to include most of common wireless channels. To facilitate the analysis, we consider the MG distribution to model the fading channel proposed in [72], which has a high accuracy for modeling most of fading channels, such as Nakagami-lognormal (NL) composite fading,  $\eta - \mu$  fading, Nakagami- $q$  (Hoyt) fading,  $\kappa - \mu$  fading, and Nakagami- $n$  (Rician) fading<sup>3</sup>. We outline these common fading models in Table 5.1. We also list the calibration parameters for those aforementioned fading models in Table 5.2. Note that there are two formats of the meaning of  $\eta$  in the  $\eta - \mu$  fading model [136], but we, for simplicity, only present Format 1 in Table 5.2. We refer to [72] and [50, Ch. 2] for more details about these fading models.

From [72, Eq. (1)] we know that the PDF of the instantaneous SNR at  $U_{g,b}$  over MG channels takes the form

$$f_{\text{SNR}_{g,b}}(x) = \sum_{v=1}^V \alpha_v x^{\varphi_v-1} \exp(-\xi_v x), \quad x \geq 0, \quad (5.23)$$

where  $\alpha_v$ ,  $\varphi_v$  and  $\xi_v$  are the parameters of MG distribution. It is easy to see that when  $V = 1$ , the MG distribution becomes the Nakagami- $m$  fading model. Further, for  $V = 1$  and  $\varphi_v = 1$ , we will obtain the Rayleigh fading model. In (5.23),  $V$  is the number of summation terms needed to reach a satisfied accuracy for modeling a specific wireless channel. In view of the PDF in (5.23), it is easy to derive the average SNR ( $\Upsilon$ ) and the variance over MG fading channels, which are respectively given by

$$\Upsilon = \mathbb{E}\{\text{SNR}_{g,b}\} = \sum_{v=1}^V \alpha_v \Gamma(\varphi_v + 1) \xi_v^{-(\varphi_v+1)}, \quad (5.24)$$

$$\text{Var}\{\text{SNR}_{g,b}\} = \sum_{v=1}^V \alpha_v \Gamma(\varphi_v + 2) \xi_v^{-(\varphi_v+2)} - \Upsilon^2. \quad (5.25)$$

As pointed out in [72], the definition of  $\alpha_v$  over NL composite fading,  $\eta - \mu$  fading, Nakagami- $q$  (Hoyt) fading,  $\kappa - \mu$  fading and Nakagami- $n$  (Rician) fading channels, is identical, and given by

$$\alpha_v = \frac{\theta_v}{\sum_{v'=1}^V \theta_{v'} \Gamma(\varphi_{v'}) \xi_{v'}^{-\varphi_{v'}}}. \quad (5.26)$$

<sup>3</sup>As pointed out in [132–134], the fluctuating two-ray (FTR) fading — which accurately models the small-scale fading in mmWave channels (especially in 28GHz outdoor mmWave channels) [135] — can be also expressed as a MG distribution.

Table 5.3 outlines how the remaining parameters of the general MG model are calibrated to fit each aforementioned channel. We here ignore Nakagami- $m$  and Rayleigh models, as this calibration is a straightforward reason. For the MG parameters calibrated to fit the NL composite channel in Table 5.3, we recall that  $\omega_v$  and  $x_v$  are the weights and sample points of GHQ, respectively.

Table 5.1: Common Fading Models in Wireless Channels

Channel Models	Fitting Scenarios
NL Composite Channel	Composite multipath/shadowing channels
$\eta - \mu$ Channel	Non-line of sight small-scale fading
Nakagami- $q$ (Hoyt) Channel	Satellite links with strong ionospheric scintillation
$\kappa - \mu$ Channel	Having line-of-sight components
Nakagami- $n$ (Rician) Channel	Having a strong line-of-sight component

Table 5.2: Parameters of Fading Models Listed in Table 5.1

Channel Models	Parameters
NL Composite	$m$ : fading parameter in Nakagami- $m$ fading; $\rho$ : unfaded SNR; $\mu$ and $\lambda$ : mean and standard deviation of the lognormal distribution
$\eta - \mu$	$\mu$ : number of multipath clusters; $\eta$ : power ratio of the in-phase component to the quadrature component; $h = \frac{2+\eta^{-1}+\eta}{4}$ ; $H = \frac{\eta^{-1}-\eta}{4}$
Nakagami- $q$ (Hoyt)	$q \in [0, 1]$ : ratio of standard deviations of real and imaginary components
$\kappa - \mu$	$\kappa$ : power ratio of dominant components to scattered components of signal; $\mu = (1 + 2\kappa)\mathbb{E}^2\{\text{SNR}\}/((1 + \kappa)^2\text{Var}\{\text{SNR}\})$
Nakagami- $n$ (Rician)	$n \in [0, +\infty)$ : arithmetic square root of Rician factor $K$

Let us use  $G_{\cdot}(\cdot)$  to denote the Meijer's G-function [76]. We now state the following result to make Lemma 5.3 valid for a general fading channel.

**Lemma 5.5.** *In the general MG fading model, when applying the large- $B$  approximation in Lemma 5.3, the mean and the variance of  $\ln(1 + \text{SNR}_{g,b})$  becomes respectively,*

$$\varrho_l = \sum_{v=1}^V \alpha_v \xi_v^{-\varphi_v} G_{3,2}^{1,3} \left( \frac{1}{\xi_v} \middle|_{1,0}^{1-\varphi_v, 1, 1} \right), \quad (5.27)$$

$$\sigma_l^2 = \sum_{v=1}^V \alpha_v \xi_v^{-\varphi_v} G_{1,0:2,2:2,2}^{0,1:1,2:1,2} \left( \frac{1}{\xi_v}, \frac{1}{\xi_v} \middle|_{1,0}^{1-\varphi_v} \middle|_{1,0}^{1,1} \middle|_{1,0}^{1,1} \right) - \varrho_l^2. \quad (5.28)$$

*Proof.* See Appendix D.6. □

To simplify Lemma 5.5, similar to Lemma 5.4, we have the following robust approximations in low SNR over MG fading channels.

**Lemma 5.6.** *In the MG fading model, when the variance of  $\text{SNR}_{g,b}$  is small, we can robustly approximate the mean and the variance of  $\ln(1 + \text{SNR}_{g,b})$  respectively as*

$$\varrho_l \approx \ln(1 + \Upsilon) - \frac{\sum_{v=1}^V \alpha_v \Gamma(\varphi_v + 2) \xi_v^{-(\varphi_v+2)} - \Upsilon^2}{2(1 + \Upsilon)^2}, \quad (5.29)$$

Table 5.3: Parameters in MG Distribution

Channel Models	Parameters in MG Distribution
NL Composite Channel	$\varphi_v = m, \xi_v = \frac{m}{\rho} \exp(-\sqrt{2}\lambda x_v + \mu), \theta_v = \left(\frac{m}{\rho}\right)^m \frac{\omega_v \exp(-m(\sqrt{2}\lambda x_v + \mu))}{\sqrt{\pi}\Gamma(m)}$
$\eta - \mu$ Channel	$\varphi_v = 2(\mu - 1 + v), \xi_v = \frac{2\mu h}{\Upsilon}, \theta_v = \frac{2\sqrt{\pi}\mu^{\mu+0.5}h^\mu}{\Gamma(\mu)H^{\mu-0.5}\Upsilon^{\mu+0.5}} \frac{(\mu H/\Upsilon)^{\mu+v-2.5}}{(v-1)!\Gamma(\mu+v-0.5)}$
Nakagami- $q$ (Hoyt) Channel	$\varphi_v = 2v - 1, \xi_v = \frac{(1+q^2)^2}{4q^2\Upsilon}, \theta_v = \frac{1+q^2}{2q\Upsilon\Gamma(v)(v-1)!} \left(\frac{1-q^4}{8q^2\Upsilon}\right)^{2v-2}$
$\kappa - \mu$ Channel	$\varphi_v = \mu + v - 1, \xi_v = \frac{\mu(1+\kappa)}{\Upsilon}, \theta_v = \frac{\mu(1+\kappa)^{(\mu+1)/2} \mu^{2v+\mu-3} \left(\frac{\kappa(1+\kappa)}{\Upsilon}\right)^{\frac{2v+\mu-3}{2}}}{\kappa^{(\mu-1)/2} \exp(\mu\kappa)\Upsilon^{(\mu+1)/2}\Gamma(\mu-1+v)(v-1)!}$
Nakagami- $n$ (Rician) Channel	$\varphi_v = v, \xi_v = \frac{(1+n^2)}{\Upsilon}, \theta_v = \frac{(1+n^2)}{\exp(n^2)[(v-1)!]^2\Upsilon} \left(\frac{n^2(1+n^2)}{\Upsilon}\right)^{v-1}$

$$\sigma_l^2 \approx \frac{\sum_{v=1}^V \alpha_v \Gamma(\varphi_v + 2) \xi_v^{-(\varphi_v+2)} - \Upsilon^2}{(1 + \Upsilon)^2} - \frac{\left(\sum_{v=1}^V \alpha_v \Gamma(\varphi_v + 2) \xi_v^{-(\varphi_v+2)} - \Upsilon^2\right)^2}{4(1 + \Upsilon)^4}. \quad (5.30)$$

*Proof.* This approximation for  $\varrho_l$  (or  $\sigma_l^2$ ) can be easily obtained by substituting the average and variance of  $\text{SNR}_{g,b}$  (from (5.24) and (5.25)), into (D.20) (or (D.23)).  $\square$

## 5.5 Numerical Results

We here provide numerical results to validate the correctness of derived expressions. In the numerical results, referring to [71, Table III], we consider four typical LMS channel fading scenarios listed in Table 5.4. When implementing the EGBMGF function, one can refer to [137] for MATLAB code or to [138] for MATHEMATICA code.

Table 5.4: Typical Fading Scenarios in LMS Channel

Fading Scenarios	$m_0$	$b_0$	$\aleph$
Frequent Heavy Shadowing	1	0.063	$8.97 \times 10^{-4}$
Overall Results	5	0.251	0.278
Average Shadowing	10	0.126	0.835
Infrequent Light Shadowing	20	0.158	1.29

In Fig. 5.2, for an average shadowing case we plot the effective coded caching gain of the ACC algorithm versus  $\rho$  for different values of  $B$ . We also plot the effective gain of the MN scheme as a performance benchmark. As expected, both the effective gains of the ACC and MN schemes converge to the nominal gain  $G$  as the SNR increases, where clearly the ACC converges much faster and thus enjoys a much better performance. Moreover, the convergence of the effective ACC gain is significantly accelerated with increasing  $B$ . As  $\rho$  decreases, both the effective gains of the ACC and of the MN schemes converge to their lower bounds, but the effective gain of the ACC scheme has a much larger lower-bound and approaches to the nominal gain very fast as  $B$  increases. For example, the ACC scheme with  $B = 20$  recovers over 80% of the nominal gain, even in this low SNR limit.

In Fig. 5.3, we present both the effective coded caching gains of the ACC and the MN schemes in four typical LMS channels fading scenarios listed in Table 5.4. We can easily see that the effective gain in the infrequent light shadowing case is the largest in both ACC and MN schemes. The average shadowing case has the second largest effective gain, while the effective gain in the frequent heavy shadowing case is the smallest (and is equal to 1 in the MN case). For a realistic  $G = 6$ , we observe that a) even in the ‘optimistic’ scenario of infrequent light shadowing, the low-SNR MN gain is approximately 2, and b) that for approximately  $B = 20$ , ACC boosts MN by a factor of approximately 2.5.

In Figs. 5.4–5.5, we present some numerical results to validate the correctness of the derived expressions for the average rates of the MN and the uncoded TDM schemes. In Figs. 5.4–5.5, the analytical result is derived based on Lemma 5.1 for the MN case,

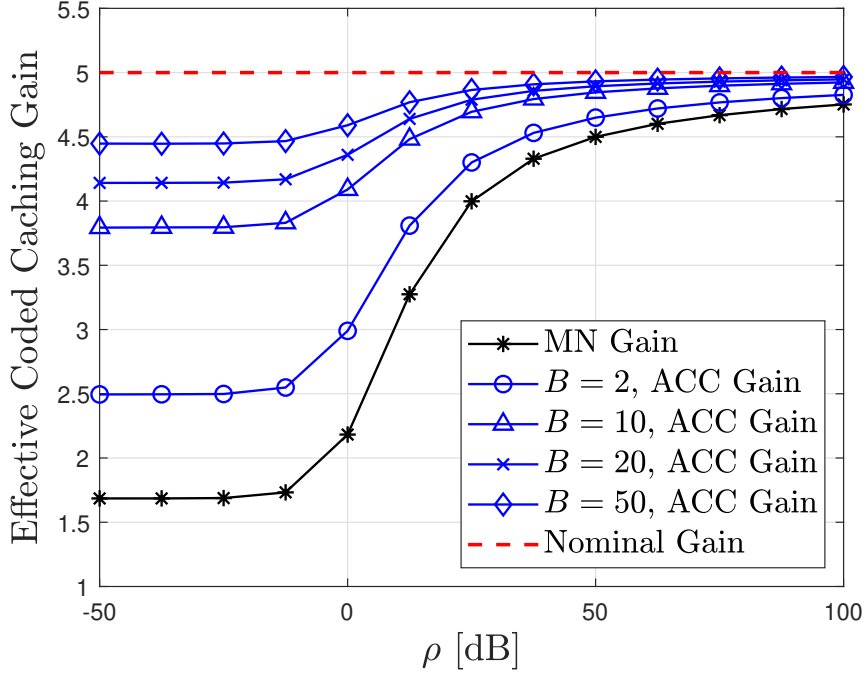
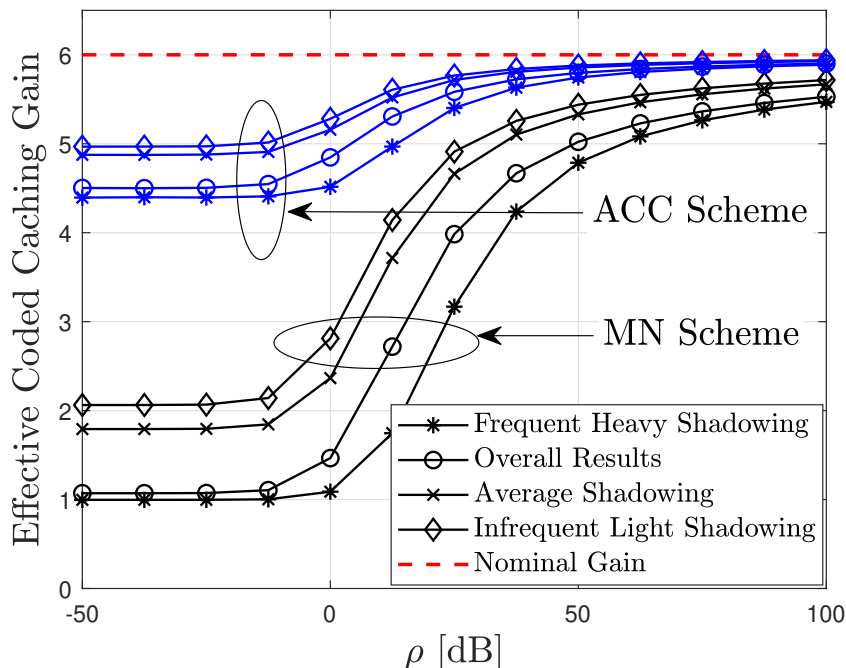


Figure 5.2: Effective coded caching gain versus  $\rho$  for  $G = 5$  in the average shadowing case.

and on Corollary 5.1 for the TDM case, while we use the low-SNR approximation from Corollary 5.2 for the MN and from Lemma 5.4 for the TDM. As expected, the average rate improves as the statistics of LMS channels becomes better. In Fig. 5.4, the value of  $\rho$  at which the nonlinear part of the average rate becomes significant as the LMS channel statistics improves<sup>4</sup>. This explains why the matching performance between the low-SNR approximation in Corollary 5.2 and the exact result becomes worse with improving the LMS channel statistics. In contrast, the low-SNR approximation of the TDM scheme in Lemma 5.4 not only matches the exact result very well in the low-to-medium  $\rho$  region, but also has a very high accuracy when  $\rho$  is as high as 30 dB, as shown in Fig. 5.5. This is because we actually consider a second-order approximation in Lemma 5.4, while the low-SNR approximation for the MN rate in Corollary 5.2 is based just on the first-order truncation from the expanded Taylor series.

In Figs. 5.6–5.8, we demonstrate the accuracy of the derived expressions for the average rate of the ACC algorithm, where the large- $B$  approximation refers to Lemma 5.3, while the large- $B$  and low-SNR approximation is derived by substituting the low-SNR approximations for  $\varrho_l$  and  $\sigma_l^2$  (as shown in Lemma 5.4) into Lemma 5.3. Fig. 5.6 shows the accuracy of the derived expressions in four typical fading scenarios listed in Table 5.4.

<sup>4</sup>We note that the infrequent light shadowing case has the best channel statistics, followed by the average shadowing and overall results cases, while the channel statistics is the worst in the frequent heavy shadowing case

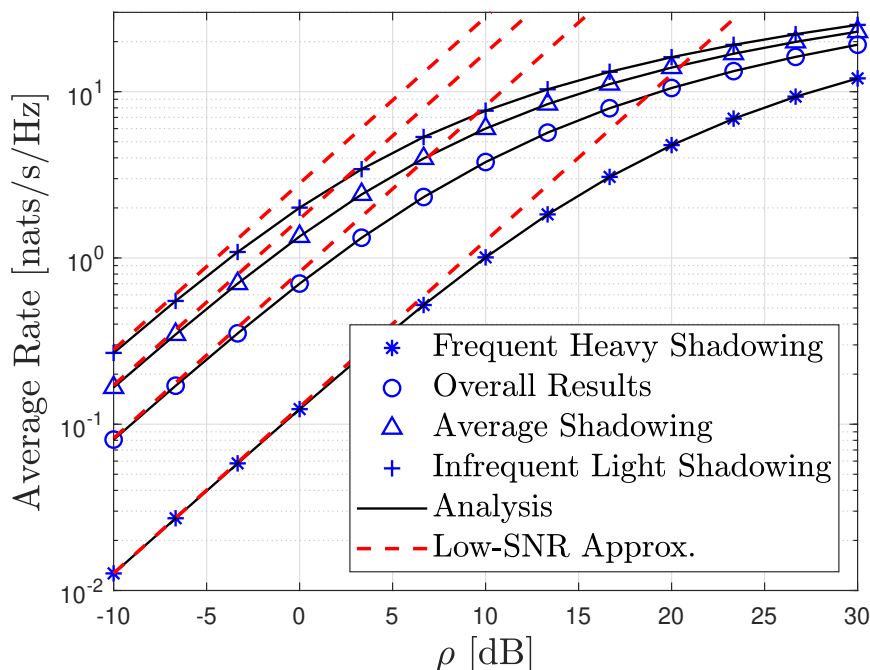
Figure 5.3: Effective coded caching gain versus  $\rho$  for  $G = 6$  and  $B = 20$ .

Obviously, there is no discernible gap between the large- $B$  approximation in Lemma 5.3 and the exact (simulated) result in the four typical fading scenarios even for a not-so-large  $B$  (e.g.,  $B = 20$ ). It is also worth noting that the large- $B$  and low-SNR approximation matches the exact result very well over the entire  $\rho$  region from -10 dB to 30 dB. In Figs. 5.7–5.8, we select the average shadowing case to represent the LMS channel statistics. Under this assumption, Fig. 5.7 plots the average rate of the ACC scheme versus  $B$  (from 2 to 50), revealing that the large- $B$  approximation in Lemma 5.3 is robust even for values of  $B$  as low as 2. In Fig. 5.7, the average rate is improved as  $B$  increases. In Fig. 5.8, we plot the average rate versus  $G$ .

In the end, in Fig. 5.9 we plot the effective coded caching gain of the ACC scheme versus  $\rho$  for different values of  $B$ . The result of the large- $B$  approximation is derived by using the average rate of the TDM scheme in Corollary 5.1 and dividing this by the approximate average rate in Lemma 5.3.

## 5.6 Conclusions

We have investigated the delivery performance of the ACC scheme in LMS systems by analyzing the average rate and the effective coded caching gain. Specifically, we have derived the closed-form expression for the exact average rate of the MN scheme, as well as derived a low-SNR approximation. The analytical expression for the average rate of the ACC algorithm has been also derived in terms of a double-integral form. To

Figure 5.4: Average rate of the MN scheme versus  $\rho$  for  $G = 4$ .

simplify the complicated expression for the ACC average rate, we have presented a tight approximation with the assumption of a large number of users. From numerical results, we can observe the significant improvement of the effective coded caching gain brought about from the ACC algorithm. Apart from providing some interesting comparisons, numerical results have validated the accuracy of our derived expressions. We can also see that the large- $B$  approximation has a robust accuracy even for a small value of  $B$ . Our analysis shows that the ACC scheme can recover most of the high-SNR nominal gain with a reasonable value of  $B$  (number of users per cache) to asymptotically overcome the worst-user bottleneck. This implies that coded caching has the ability to operate in such (low-SNR governed) LMS systems.

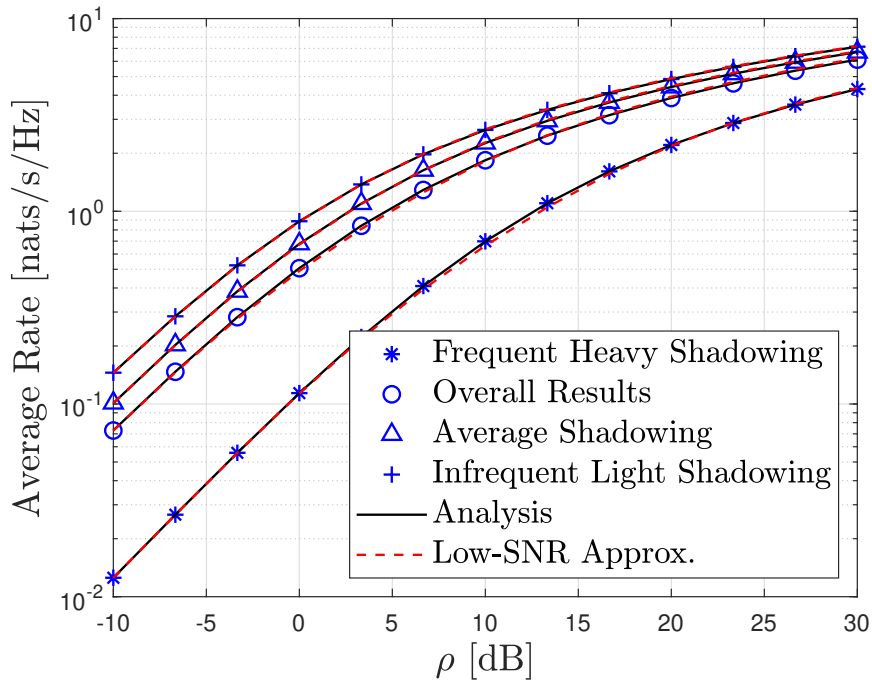


Figure 5.5: Average rate of the uncoded TDM scheme versus  $\rho$ .

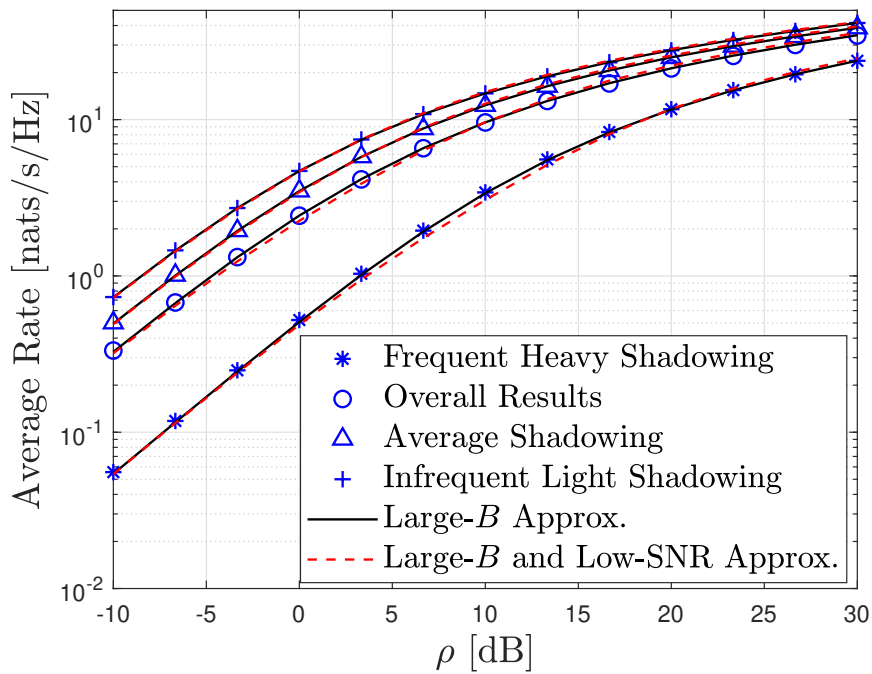


Figure 5.6: Average rate of the ACC scheme versus  $\rho$  for  $G = 6$  and  $B = 20$ .

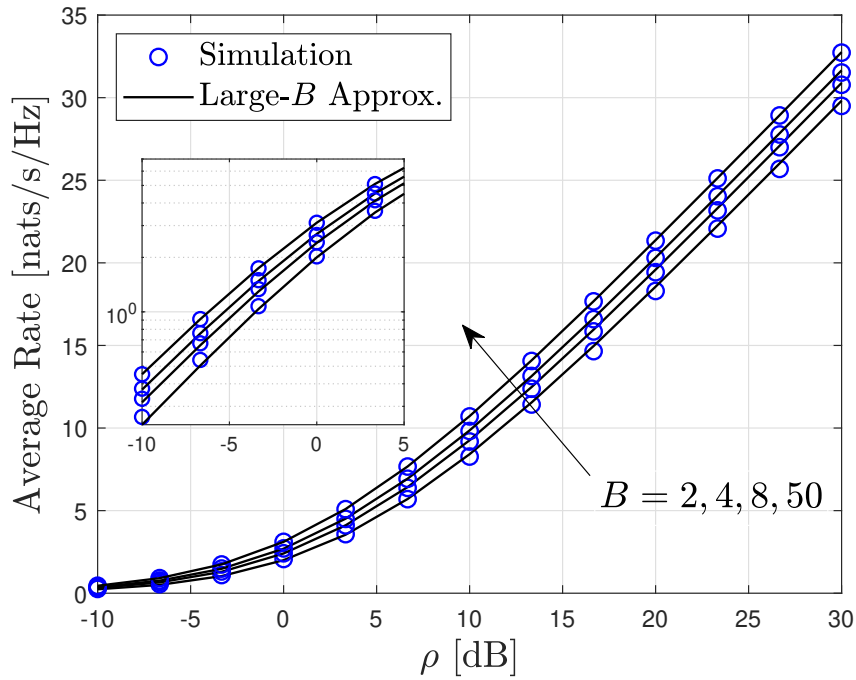


Figure 5.7: Average rate of the ACC versus  $\rho$  for  $G = 5$  in the average shadowing case.

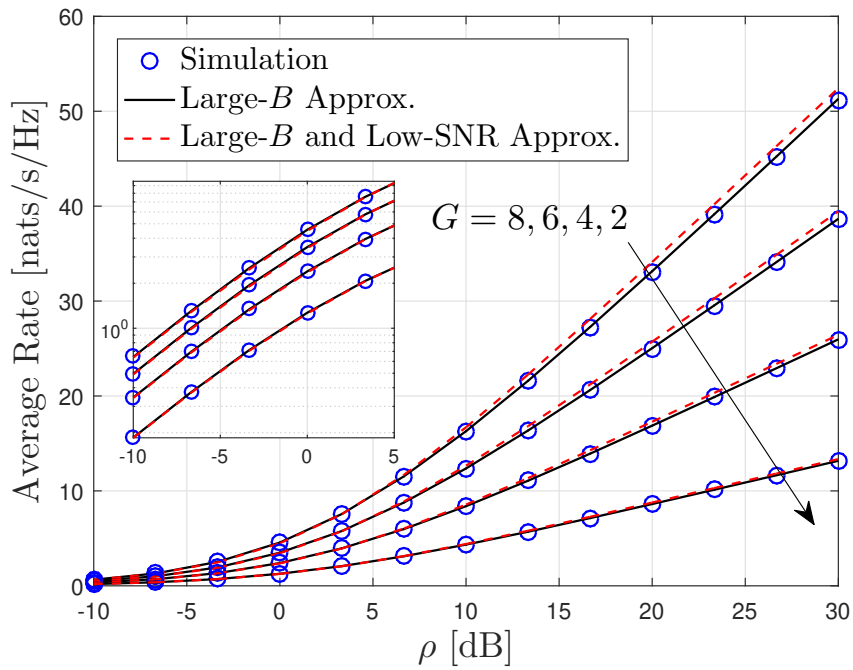


Figure 5.8: Average rate of the ACC versus  $\rho$  for  $B = 20$  in the average shadowing case.

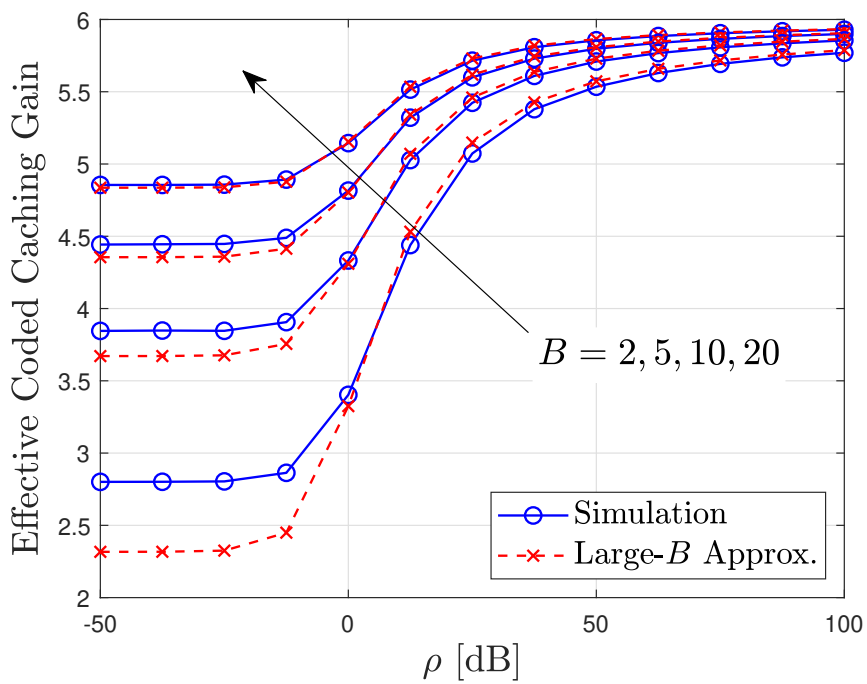


Figure 5.9: Effective gain of the ACC versus  $\rho$  for  $G = 6$  in the average shadowing case.



## Chapter 6

# Conclusions and Perspectives

This thesis explores the true effect of coded caching when it is applied in various realistic wireless communications settings. Our emphasis is on the use of caching both on single stream (rank 1) settings, as well as when caching is used to boost advanced state-of-art multi-antenna systems that are now at the cutting edge of technology. We believe that the thesis has made some considerable contributions in both cases.

In the single-stream scenario, we have first shown that traditional coded caching approaches provide massively diminished effective gains as the SNR becomes smaller. In this context, we have invented the novel ACC approach, which yields gains that are close to the nominal (infinite SNR) gains, and does so with very little overhead. ACC was explored in a variety of settings that include basic cellular settings with quasi-static or ergodic fading, satellite settings, and settings with reduced allowable hardware complexity, in particular with a reduced number of RF chains. What we show is that if the number of RF chains is modest, then a basic single-RF-chain ACC solution can compete with the state of art of solutions that require multiple transmit antennas, multiple RF chains and considerably larger CSI overheads. This naturally may also apply in mm-Wave scenarios, where maintaining a reduced number of RF chains is crucial.

Then the thesis explored the actual impact that coded caching can have in improving advanced MU-MIMO systems, where such systems are known to build heavily on multi-antenna arrays and on an optimized exploitation of beamforming and spatial-multiplexing techniques. As one can imagine, it is imperative that any impactful cache-aided technique must be able to work in tandem with these multi-antenna systems, and it would be inconceivable to expect an operator to substitute their multi-antenna multiplexing gains with cache-aided multicasting gains that will remain modest under most realistic assumptions on file sizes and cache sizes. While prior attempts to incorporate coded caching with multi-antenna arrays, had provided — under practical considerations — minimal, if any at all, improvements over multiplexing gain techniques, it is vector coded caching that managed to provide dramatic gains, at least in theory, over non-optimized MU-MISO systems. The work in this thesis focuses on optimizing vector coded caching and most importantly, focuses on answering a simple question: Under a fixed set of antenna and SNR resources, and under various practical considerations, what is the multiplicative throughput boost obtained from being able to add receiver-side caches to

downlink systems that would have otherwise been able to enjoy an optimized exploitation of multiplexing and beamforming gains. This is done in this thesis for a variety of settings, that include the symmetric MU-MISO setting, as well as the non symmetric (variable pathloss) MU-MISO and MU-MIMO settings, in the presence of various high-performance precoders. In various scenarios, such as for example in the symmetric MU-MISO case, or the small cell MU-MISO and MU-MIMO settings, the gains are indeed very notable and can exceed 300-400% boost in achievable rates (over the optimized cacheless counterparts) under a variety of practical considerations such as CSI costs, variable pathloss, various propagation models, etc.

The work has provided a resolution of the infamous worst-user bottleneck of wireless coded caching, which was thought to severely diminish cache-aided multicasting gains — which is irrespective of either the (high-SNR) nominal gain or the number of users — due to the fundamental worst-channel limitation of multicasting. The novel ACC algorithm solves this problem and we now know that this bottleneck is not a fundamental one. Key to the ACC algorithm is a dual idea: users share cache states (which is effectively unavoidable due to the finite file sizes), and multi-rate transmissions to match each single-link capacity in the presence of side information. Together these two aspects allow ACC to convert the worst-user effect to a dramatically ameliorated worst-group-of-users effect, allowing for a near-complete recovery of the nominal gains.

Another contribution of this thesis is the extensive analysis that it has provided. Our analysis includes channels with Rayleigh fading, Nakagami- $m$  fading, Shadowed-Rician fading (for LMS channels) and a more generalized MG fading, various ergodic settings with different pathloss among the served users, and incorporation of 3GPP-proposed urban macro/micro cell specifications. The thesis has also developed general but simple methods to analyze various schemes based on the central limit theorem, and various other techniques. The thesis has also provided substantial validation of the analytical results, which are almost always shown to hold very tight for a practical and not-so-large number of users.

Another aspect that comes out of this thesis, is the optimization of vector coded caching. This optimization was provided here analytically, and validated numerically. The optimized cache-aided linear precoding schemes are very simple to implement, as they simply exploit cached content in order to be able to simultaneously transmit carefully selected precoded vectors that would have otherwise been sent one after the other. Because of the simplicity of this idea, it is conceivable to expect the gains to persist for a broader class of precoders. Our performance analysis derives some simple expressions based on realistic assumptions that reveal significant multiplicative gains from applying caching over already optimized downlink systems, where these gains persist for various well-known precoding classes. This same analysis and optimization are here shown to hold very tight in realistic wireless network settings, while also incorporating the aforementioned variety of practical considerations such as power dissemination across signals, realistic SNR values, statistically asymmetric channels, MMF, as well as CSI costs. The comparisons of optimized cache-aided vs. optimized cacheless downlink systems reveal that vector coded caching can recover a sizeable portion of its theoretic (high-SNR) gain  $G = \Lambda\gamma + 1$ , even in realistic wireless settings operating at realistic SNR values. For example, for 32

transmit antennas and 2 receive antennas each user and for a typical BS transmit power of 33 dBm in a Micro-cell, under the assumption of uniformly distributed users, vector coded caching is shown to offer a 410% boost in overall throughput over the cacheless MU-MIMO system, where both their multiplexing gains are independently optimized for ZF precoding, and both them implement optimal power allocation for the MMF criterion.

In terms of challenges, indeed  $G$  remains, under current practices, bounded in the range of single digits. Any improvement beyond this range would require either a dramatic increase in the storage capability of nodes ( $\gamma$ ), or a research breakthrough in the area of subpacketization-constrained coded caching. Further improving the subpacketization-constrained performance of coded caching primitives (thus effectively allowing for a larger  $\Lambda$ ) remains to date the big challenge in coded caching, and any progress in that direction would undoubtedly have a profound impact on the performance of cache-aided multi-antenna systems.

Another practical problem is that this vector coded caching scheme requires very high hardware/software overheads and energy consumption especially on the fully connected RF chains. Although this implementation cost may be affordable in conventional sub-6GHz MU-MIMO communications, we still need to design a much more energy-efficient vector coded caching scheme which will rely on less RF chains and perform hybrid precoding at only a slight performance loss to the fully connected RF chains. Furthermore, vector coded caching with fully connected RF chains may not be possible in mmWave communications because of 1) the extremely high energy consumption on the RF chains, 2) very small SNR before beamforming during CSI feedback, and 3) a huge training overhead brought about by the large number of antennas. It is also worth noting that the spectrum bandwidth in mmWave is dozens of times that of a sub-6GHz system, which implies that a very high spectral efficiency may not be necessary. This in turn motivates the investigation of vector coded caching with much less RF chains and limited CSI feedback in multi-user mmWave systems.

The reported gains here will naturally come under pressure from additional realistic considerations such as having statistically asymmetric channels, although this problem can be partially ameliorated with power control, with rate-splitting approaches [139, 140], with smaller cell sizes (see Fig. 4.7), or with the novel ACC approach in Chapter 2. These same reported gains may also come under pressure from the additional CSI costs that would arise in the event where multi-antenna coded caching algorithms start serving more and more users. Remedies for this can be found in the novel clique structures recently reported in [105]. A big associated open problem is the simultaneous reduction of both the subpacketization and CSI costs (see [141] for some early efforts). Naturally the system performance also remains subject to the need for cacheable and live-streamed data to co-exist (cf. [140]), the need for cache-aided and cacheless users to coexist,<sup>1</sup> as well as will depend on the stochastic nature of the network topology and user behavior (for some early remedies, the reader can refer to [99, 142]).

The presented new results, as well as the aforementioned challenges, arrive at an instance when bandwidth and antenna resources are asked to handle an aggressively

---

<sup>1</sup>See [47], which reveals the surprising conclusion that cacheless users can benefit from full coded caching gains.

increasing volume of data. At the same time though, the new results come at a time when Moore's law on storage capabilities remains intact and the ever-increasing majority of communicated content is cacheable [143]. For these reasons, and given the powerful gains reported here, we believe that the aforementioned techniques can further help translate the abundance of Gbytes of storage space into much needed spectral efficiency in sub-6GHz bands.

# Appendices



# Appendix A

## Proofs in Chapter 2

### A.1 Capacity Region of Proposition 2.1

This appendix is meant to orient the reader as to how the existing results in [49] on multicasting with side information<sup>1</sup> can be applied to our setting.

Using the notation of [49] and following the same derivation as in [73], we recover Proposition 2.1 from [49, Thm. 6] by choosing  $X^n$  to be  $(X_1, X_2, \dots, X_t)^n$ , selecting  $m = n$  in [49, Thm. 6], setting the side information  $Y_i$  to be  $Y_i = \{X_\ell\}_{\ell \in [t] \setminus i}$ , and applying invertible mappings between  $X_i^n$  and  $W_i^n$  for any  $i \in t$ . From the maximum entropy theorem [145, Thm. 9.6.5], we obtain Proposition 2.1.

For the achievability part, we proceed as in [49] and consider a codebook of  $2^{n(\sum_{\ell=1}^t R_\ell)}$  codewords. The codewords are denoted by  $x^n(w_1, w_2, \dots, w_t)$ , with  $w_\ell \in [2^{nR_\ell}]$  for any  $\ell \in [t]$ . The letters of the codewords, denoted by  $x_j(w_1, w_2, \dots, w_t)$ ,  $j \in [n]$ , are i.i.d. distributed as  $\mathcal{N}(0, P)$ . Each user can decode its intended message from the received signal and from the (cached) side information using typical set decoding. The intuition behind the successful decoding at a certain user  $i$  is that, after receiving one of the  $2^{n(\sum_{\ell=1}^t R_\ell)}$  codewords and thanks to the cached information, user  $i$  applies typical decoding over only  $2^{nR_i}$  possible codewords.

### A.2 Proof of Lemma 2.1

Let us start by introducing the notation  $S_g \triangleq \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b})$ , for any group  $g \in [\Lambda]$  of users, such that we can write the average rate of the ACC scheme as

$$\bar{R}^{\text{ACC}} = \frac{G}{B \ln 2} \mathbb{E}_H \left\{ \min_{g \in \Psi} \{S_g\} \right\}. \quad (\text{A.1})$$

---

<sup>1</sup>Several works have considered this Gaussian setting after [49]. In [73], the capacity region was derived for the 2-user case, the 3-user case was studied in [74, Group 8, case  $\mathcal{G}_{18} \cup \mathcal{G}_{28}$ ], and the converse of Prop. 2.1 can be also found in [144, Thm. 4].

For  $t \in (-\infty, +\infty)$ , the characteristic function (CF) in probability [146, Ch. 5] of  $S_g$  is defined as

$$\text{CF}_{S_g}(t) = \mathbb{E} \{ \exp(jtS_g) \} = \mathbb{E} \left\{ \exp \left( jt \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right) \right\} = [\mathbb{E} \{ (1 + \text{SNR}_{g,b})^{jt} \}]^B. \quad (\text{A.2})$$

Substituting the PDF of  $\text{SNR}_{g,b}$  into (A.2) yields

$$\text{CF}_{S_g}(t) = \frac{1}{\rho^B} \left[ \int_0^\infty (1+x)^{jt} \exp\left(-\frac{x}{\rho}\right) dx \right]^B \stackrel{(a)}{=} \frac{1}{\rho^B} \exp\left(\frac{B}{\rho}\right) \text{E}_{-jt}^B\left(\frac{1}{\rho}\right), \quad (\text{A.3})$$

where (a) follows from [76, Eq. (3.382.4)]. By considering the Gil-Pelaez Theorem [147], the CDF of  $S_g$  is obtained as

$$F_{S_g}(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\text{Im} \left\{ \exp(-jxt) \frac{\exp(B/\rho)}{\rho^B} \text{E}_{-jt}^B\left(\frac{1}{\rho}\right) \right\}}{t} dt.$$

Define  $S_{\min} \triangleq \min_{g \in \Psi} \{S_g\} = \min_{g \in \Psi} \left\{ \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right\}$ . The CDF of  $S_{\min}$  can be expressed by

$$\begin{aligned} F_{S_{\min}}(y) &= \Pr \left\{ \min_{g \in \Psi} \{S_g\} \leq y \right\} = 1 - \Pr \left\{ \min_{g \in \Psi} \{S_g\} > y \right\} = 1 - (\Pr \{S_g > y\})^G \\ &= 1 - \left( \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\text{Im} \left\{ \exp(-jxt) \frac{\exp(B/\rho)}{\rho^B} \text{E}_{-jt}^B\left(\frac{1}{\rho}\right) \right\}}{t} dt \right)^G \end{aligned} \quad (\text{A.4})$$

As  $S_{\min}$  is a non-negative random variable, it holds that  $\mathbb{E} \{S_{\min}\} = \mathbb{E} \left\{ \int_0^{S_{\min}} dx \right\}$ , and furthermore,

$$\begin{aligned} \mathbb{E} \left\{ \int_0^{S_{\min}} dx \right\} &= \mathbb{E} \left\{ \int_0^\infty \mathbb{I} \{x \leq S_{\min}\} dx \right\} \\ &= \int_0^\infty \mathbb{E} \{ \mathbb{I} \{x \leq S_{\min}\} \} dx = \int_0^\infty [1 - F_{S_{\min}}(y)] dy, \end{aligned} \quad (\text{A.5})$$

where  $\mathbb{I}\{\cdot\}$  denotes the indicator function, which, for claim  $\mathcal{A}$ , takes the value  $\mathbb{I}\{\mathcal{A}\} = 1$  if  $\mathcal{A}$  is true and  $\mathbb{I}\{\mathcal{A}\} = 0$  otherwise. Combining (A.4) and (A.5) yields that the expectation of  $S_{\min}$  is

$$\mathbb{E} \{S_{\min}\} = \int_0^\infty \left( \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\text{Im} \left\{ \exp(-jxt) \frac{\exp(B/\rho)}{\rho^B} \text{E}_{-jt}^B\left(\frac{1}{\rho}\right) \right\}}{t} dt \right)^G dy.$$

It follows from (A.1) that  $\bar{R}^{\text{ACC}} = \frac{G}{B \ln 2} \mathbb{E} \{S_{\min}\}$ , which gives Lemma 2.1 by considering the integral form of  $\mathbb{E} \{S_{\min}\}$ , and therefore Lemma 2.1 is proven.  $\square$

### A.3 Proofs for Section 2.3.2 and Section 2.3.3

#### A.3.1 Proof of Lemma 2.2

The fact that  $\text{SNR}_g$  is distributed as  $\text{Exp}(G/\rho)$  implies that  $\text{Var}(\min_{g \in \Psi} \{\text{SNR}_g\}) = \rho^2/G^2 = o(\rho)$ . Thus, in a similar way as in [130, Eq. (4)], in the low-SNR region we can approximate  $\bar{R}^{\text{MN}}$  by its robust approximation based on the Taylor series: Let  $\mathcal{P}(X)$  be a real-valued function with respect to a random variable  $X$  with mean  $\mu_X$  and variance  $\sigma_X^2$ . The expectation of  $\mathcal{P}(X)$  can be tightly approximated in the low  $\sigma_X^2$  region as

$$\mathbb{E}\{\mathcal{P}(X)\} \approx \mathcal{P}(\mu_X) + \frac{\sigma_X^2}{2} \frac{\partial^2 \mathcal{P}(X)}{\partial X^2} \Big|_{X=\mu_X} \quad (\text{A.6})$$

where  $\frac{\partial^2 \mathcal{P}(X)}{\partial X^2}$  stands for the second derivative of  $\mathcal{P}(X)$  with respect to  $X$  (cf. [129]).

Considering that  $\mathcal{P}(X) = \frac{G}{\ln 2} \ln(1 + \min_{g \in \Psi} \{\text{SNR}_g\})$  and that  $X = \min_{g \in \Psi} \{\text{SNR}_g\}$  and adopting the robust approximation of (A.6) yields that  $\bar{R}^{\text{MN}}$  can be tightly approximated at low SNR by (2.9).  $\square$

#### A.3.2 Proof of Proposition 2.2

From the fact that  $\text{Ei}(-x)$  is bounded as (cf. [148])

$$-e^{-x} \ln\left(1 + \frac{1}{x}\right) < \text{Ei}(-x) < \frac{-e^{-x}}{2} \ln\left(1 + \frac{2}{x}\right), \quad (\text{A.7})$$

we can upper bound the numerator and lower bound the denominator of the exact expression of  $\frac{\bar{R}^{\text{MN}}}{\bar{R}^{\text{TDM}}}$  in (2.10) to obtain that

$$\lim_{\rho \rightarrow 0} \frac{\bar{R}^{\text{MN}}}{\bar{R}^{\text{TDM}}} \leq \lim_{\rho \rightarrow 0} \frac{G \ln\left(1 + \frac{2\rho}{G}\right)}{2 \ln(1 + \rho)} = 1. \quad (\text{A.8})$$

By interchanging the bounds to lower bound the ratio, we obtain that the limit is also lower bounded by 1, which concludes the proof of Proposition 2.2.  $\square$

#### A.3.3 Proof of Lemma 2.3

We start by proving that

$$\mathbb{E}\left\{\sum_{b=1}^B \ln(1 + \text{SNR}_{g,b})\right\} = \mathbb{E}\left\{\sum_{b=1}^B \text{SNR}_{g,b}\right\} + o(\rho), \quad (\text{A.9})$$

which is obtained from the fact that  $\mathbb{E}\left\{\sum_{b=1}^B \ln(1 + \text{SNR}_{g,b})\right\} = \mathbb{E}\left\{\sum_{b=1}^B (\ln(1 + \text{SNR}_{g,b}) - \text{SNR}_{g,b})\right\} + \mathbb{E}\left\{\sum_{b=1}^B \text{SNR}_{g,b}\right\}$ . In the above, we obtain (A.9) from the Lebesgue's Dominated Convergence Theorem [149, Thm. 16.4] as follows: First, we know that  $\lim_{x \rightarrow 0} (\ln(1 + x) - x)/x = 0$ , and hence  $\ln(1 + x) - x = o(x)$  as  $x \rightarrow 0$ . In order to prove that the expectation is also  $o(\rho)$  as  $\rho \rightarrow 0$ , we need to prove that  $|\ln(1 + x) - x|$  is bounded

by some integrable function. For that, since  $\ln(1+x) \leq x$  for any  $x > 0$ , it follows that  $|\ln(1 + \text{SNR}_{g,b}) - \text{SNR}_{g,b}| \leq |\text{SNR}_{g,b}|$ , which satisfies that  $\mathbb{E}\{|\text{SNR}_{g,b}|\} = \rho < \infty$ . Hence, we can apply the Dominated Convergence Theorem and obtain (A.9).

Since  $\text{SNR}_{g,b}$  is distributed as  $\text{Exp}(\frac{1}{\rho})$ ,  $\sum_{b=1}^B \text{SNR}_{g,b}$  follows a  $\text{Gamma}(B, \rho)$  distribution, with shape and scale parameters  $B$  and  $\rho$ . Then, the CDF of  $S' \triangleq \min_{g \in \Psi} \{\sum_{b=1}^B \text{SNR}_{g,b}\}$  is given by

$$F_{S'}(y) = 1 - \left( \frac{1}{\Gamma(B)} \Gamma\left(B, \frac{y}{\rho}\right) \right)^G \stackrel{(a)}{=} 1 - \left( \exp\left(-\frac{y}{\rho}\right) \sum_{t=0}^{B-1} \frac{y^t}{t! \rho^t} \right)^G, \quad (\text{A.10})$$

where  $\Gamma(\cdot, \cdot)$  denotes the upper incomplete Gamma function [76], and (a) follows from [76, Eq. (8.352.2)] since  $B$  is a positive integer. For  $\mathbf{b} \in \mathbb{Z}^B$ , let  $b_t \triangleq \mathbf{b}(t) \geq 0$ ,  $t \in [B]$ , denote its  $t$ -th element. Recalling that  $\binom{n}{\mathbf{b}} \triangleq \frac{n!}{b_1! b_2! \dots b_B!}$ , we apply the Multinomial theorem [150] to get that

$$F_{S'}(y) = 1 - \exp\left(-\frac{Gy}{\rho}\right) \sum_{\|\mathbf{b}\|_1=G} \binom{G}{\mathbf{b}} \frac{\rho^{-\sum_{t=1}^B (t-1)b_t}}{\prod_{t=1}^B ((t-1)!)^{b_t}} y^{\sum_{t=1}^B (t-1)b_t}.$$

In view of the relationship between the CDF and the expectation in (A.5), the average rate of the ACC scheme can be approximated in the low-SNR region by

$$\begin{aligned} \bar{R}^{\text{ACC}} &= \frac{G}{B \ln 2} (\mathbb{E}\{S'\} + o(\rho)) = \frac{G}{B \ln 2} \int_0^\infty [1 - F_{S'}(y)] dy + o(\rho) \\ &= \frac{G}{B \ln 2} \sum_{\|\mathbf{b}\|_1=G} \binom{G}{\mathbf{b}} \frac{\rho^{-\sum_{t=1}^B (t-1)b_t}}{\prod_{t=1}^B ((t-1)!)^{b_t}} \int_0^\infty \exp\left(-\frac{Gy}{\rho}\right) y^{\sum_{t=1}^B (t-1)b_t} dy + o(\rho), \end{aligned} \quad (\text{A.11})$$

which can be solved by using the definition of Gamma function [76, Eq. (8.312.2)].  $\square$

### A.3.4 Proof of Corollary 2.1

From Lemma 2.3 we have that  $\bar{R}^{\text{ACC}} = \frac{\rho G}{B \ln 2} \mathcal{L}_G + o(\rho)$  and also that  $\bar{R}^{\text{TDM}} = \bar{R}^{\text{ACC}}|_{B=G=1} = \frac{\rho}{\ln 2} + o(\rho)$ , whereas from Proposition 2.2 it follows that  $\lim_{\rho \rightarrow 0} \frac{\bar{R}^{\text{MN}}}{\bar{R}^{\text{TDM}}} = 1$ . These results yield the desired  $\bar{R}^{\text{MN}} = \frac{\rho}{\ln 2} + o(\rho)$  and

$$\lim_{\rho \rightarrow 0} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} = \lim_{\rho \rightarrow 0} \frac{\frac{\rho G}{B \ln 2} \mathcal{L}_G + o(\rho)}{\frac{\rho}{\ln 2} + o(\rho)} = \frac{G}{B} \mathcal{L}_G. \quad \square$$

### A.3.5 Proof of Lemma 2.4

We want to prove that  $\lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{TDM}}} = \Lambda \gamma + 1$  for a fixed number of caches  $\Lambda$  and for any  $\rho$ . Since  $\mathbb{E}\{|\ln(1 + \text{SNR}_{g,b})|\} < \infty$ , the Strong Law of Large Numbers implies that

$$\frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \xrightarrow{a.s.} \mathbb{E}\{\ln(1 + \text{SNR}_{g,b})\} \quad (\text{A.12})$$

as  $B \rightarrow \infty$ , which implies that

$$\lim_{B \rightarrow \infty} \frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) = \mathbb{E} \{ \ln(1 + \text{SNR}_{g,b}) \}, \quad (\text{A.13})$$

except for zero-probability events. Then since  $\ln(1+x) \leq x \forall x > 0$ , we get that

$$\mathbb{E}_H \left\{ \min_{g \in \mathcal{G}} \frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right\} \leq \mathbb{E}_H \left\{ \frac{1}{B} \sum_{b=1}^B \text{SNR}_{g,b} \right\} \stackrel{(a)}{=} \rho < \infty, \quad (\text{A.14})$$

where (a) comes from the fact that  $\text{SNR}_{g,b} \sim \text{Exp}(\frac{1}{\rho})$  and thus  $\sum_{b=1}^B \text{SNR}_{g,b} \sim \text{Gamma}(B, \rho)$ .

From (A.13) and (A.14), we can apply Lebesgue's Dominated Convergence Theorem [149, Thm. 16.4] to interchange the order of expectation and limit and show that

$$\begin{aligned} & \lim_{B \rightarrow \infty} \bar{R}^{\text{ACC}} / \bar{R}^{\text{TDM}} \\ & \stackrel{(a)}{=} \frac{\lim_{B \rightarrow \infty} \frac{G}{\ln 2} \mathbb{E}_H \left\{ \min_{g \in \mathcal{G}} \frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right\}}{\frac{1}{\ln 2} \mathbb{E}_H \{ \ln(1 + \text{SNR}_{g,b}) \}} \end{aligned} \quad (\text{A.15})$$

$$\begin{aligned} & \stackrel{(b)}{=} G \frac{\mathbb{E}_H \left\{ \min_{g \in \mathcal{G}} \lim_{B \rightarrow \infty} \frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right\}}{\mathbb{E}_H \{ \ln(1 + \text{SNR}_{g,b}) \}} \\ & \stackrel{(c)}{=} G = \Lambda\gamma + 1, \end{aligned} \quad (\text{A.16})$$

where (a) follows from substituting  $\bar{R}^{\text{ACC}}$  and  $\bar{R}^{\text{TDM}}$  by their respective expressions, (b) comes from the Dominated Convergence Theorem and the fact that the minimum of several continuous functions is a continuous function, and (c) is due to (A.13).  $\square$

### A.3.6 Proof of Lemma 2.5

From (A.12), and by applying the same steps as in (A.15)–(A.16), we obtain (2.16) as

$$\lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} = \frac{\frac{G}{\ln 2} \mathbb{E}_H \{ \ln(1 + \text{SNR}_{g,b}) \}}{\frac{G}{\ln 2} \mathbb{E}_H \{ \ln(1 + \min_{g \in \Psi} \{ \text{SNR}_{g,b} \}) \}} \stackrel{(a)}{=} \exp \left( \frac{1}{\rho} - \frac{G}{\rho} \right) \frac{\text{Ei} \left( -\frac{1}{\rho} \right)}{\text{Ei} \left( -\frac{G}{\rho} \right)}, \quad (\text{A.17})$$

where (a) follows from (2.5). To prove (2.17), we first obtain from (A.17) that

$$\lim_{\rho \rightarrow 0} \lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} = \lim_{\rho \rightarrow 0} \exp \left( \frac{1}{\rho} - \frac{G}{\rho} \right) \frac{\text{Ei} \left( -\frac{1}{\rho} \right)}{\text{Ei} \left( -\frac{G}{\rho} \right)}. \quad (\text{A.18})$$

Then, in a similar manner as for the proof of Proposition 2.2 in Appendix A.3.2, we can apply the relations  $-e^{-x} \ln(1 + \frac{1}{x}) < \text{Ei}(-x) < \frac{-e^{-x}}{2} \ln(1 + \frac{2}{x})$  [148] in (A.18) to obtain that

$$\lim_{\rho \rightarrow 0} \lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} \leq \lim_{\rho \rightarrow 0} \exp \left( \frac{1-G}{\rho} \right) \frac{\frac{1}{2} \exp \left( \frac{-1}{\rho} \right) \ln(1 + 2\rho)}{\exp \left( \frac{-G}{\rho} \right) \ln(1 + \frac{\rho}{G})} = G, \quad (\text{A.19})$$

$$\lim_{\rho \rightarrow 0} \lim_{B \rightarrow \infty} \frac{\bar{R}^{\text{ACC}}}{\bar{R}^{\text{MN}}} \geq \lim_{\rho \rightarrow 0} \exp\left(\frac{1-G}{\rho}\right) \frac{2 \exp\left(\frac{-1}{\rho}\right) \ln(1+\rho)}{\exp\left(\frac{-G}{\rho}\right) \ln\left(1+\frac{2\rho}{G}\right)} = G, \quad (\text{A.20})$$

which concludes the proof of Lemma 2.5.  $\square$

#### A.4 Proof of Lemma 2.6

To prove Lemma 2.6, we first derive the approximation in (2.18). Afterward, we obtain the values of  $\mu$  and  $\sigma$  in (2.19) and (2.20), and finally we derive the integral expression of  $\mathcal{H}_G$  in (2.21).

Let  $A_g \triangleq \frac{1}{B} \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b})$ , for any  $g \in [A]$ , represent the arithmetic mean of the user capacity over the set of  $B$  users of group  $g$ , normalized by  $\ln(2)$ . Let us consider the Central Limit Theorem (CLT) in the large  $B$  case. According to the Lindeberg-Lévy CLT [151], we have that  $A_g \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{B}\right)$  as  $B \rightarrow \infty$ , where  $d$ . stands for *convergence in distribution*, and where  $\mu = \mathbb{E}\{\ln(1 + \text{SNR}_{g,b})\}$  and  $\sigma^2 = \text{Var}\{\ln(1 + \text{SNR}_{g,b})\}$ <sup>2</sup>. We consider now the average rate for the ACC scheme when  $B \rightarrow \infty$ . Recall that  $A_1, \dots, A_G$  are i.i.d. normal random variables with mean  $\mu$  and variance  $\sigma^2/B$ . Although convergence in distribution does not generally imply convergence in mean, it was shown in [152] that this indeed holds in the specific case of extreme values of i.i.d. random variables. Consequently,  $\bar{R}^{\text{ACC}}$  is given by

$$\lim_{B \rightarrow \infty} \bar{R}^{\text{ACC}} = \frac{G}{\ln 2} \mathbb{E}\{\min\{A_1, \dots, A_G\}\}. \quad (\text{A.21})$$

Deriving a simple closed-form expression for (A.21) is challenging. Consequently, we propose a simple method to obtain an approximation to this expectation. Since  $B \rightarrow \infty$  and  $A_1, \dots, A_G$  are i.i.d. normal random variables, we can write each  $A_i$ ,  $i \in [G]$ , as  $A_i = \mu + \frac{\sigma}{\sqrt{B}} A'_i$ , where  $A'_i \sim \mathcal{N}(0, 1)$ . Then, the minimum of  $A_1, \dots, A_G$  is re-written as

$$\min_{i \in [G]} \{A_i\} = \mu + \frac{\sigma}{\sqrt{B}} \min_{i \in [G]} \{A'_i\}. \quad (\text{A.22})$$

Then (2.18) is obtained by taking the expectation of both sides, multiplying (A.22) by  $\frac{G}{\ln 2}$ , and recalling that  $\mathcal{H}_G \triangleq -\mathbb{E}\{\min_{i \in [G]} \{A'_i\}\}$ , as defined in Section 2.3.4.

We derive now the expressions for  $\mu$  in (2.19) and  $\sigma$  in (2.20). Note that  $\frac{\mu}{\ln(2)} = \mathbb{E}\{\log_2(1 + \text{SNR}_{g,b})\}$  is exactly  $\bar{R}^{\text{TDM}}$ , so that we have (2.19) by considering (2.5) with  $G = 1$ . Moreover, we have that

$$\mathbb{E}\left\{(\ln(1 + \text{SNR}_{g,b}))^2\right\} = \frac{1}{\rho} \int_0^\infty (\ln(1+x))^2 \exp\left(-\frac{x}{\rho}\right) dx.$$

<sup>2</sup>Note that, if we focused on the low-SNR region, we could apply the approximations  $\mu \approx \mathbb{E}\{\text{SNR}_{g,b}\}$  and  $\sigma^2 \approx \text{Var}\{\text{SNR}_{g,b}\}$ . We do not consider them here for sake of generality, and our approximation holds for any value of SNR.

To obtain a closed-form expression for the previous integral, we re-write both the logarithmic function and the exponential function into their Meijer's G-function forms [76, Eq. (9.301)], given by  $\ln(1+x) = G_{2,2}^{1,2}\left(x \left| \begin{smallmatrix} 1,1 \\ 1,0 \end{smallmatrix} \right.\right)$  and  $\exp\left(-\frac{x}{\rho}\right) = G_{0,1}^{1,0}\left(\frac{x}{\rho} \left| \begin{smallmatrix} - \\ 0 \end{smallmatrix} \right.\right)$ , respectively. Then, the previous integral becomes

$$\begin{aligned} & \mathbb{E} \left\{ (\ln(1 + \text{SNR}_{g,b}))^2 \right\} \\ &= \frac{1}{\rho} \int_0^\infty G_{2,2}^{1,2}\left(x \left| \begin{smallmatrix} 1,1 \\ 1,0 \end{smallmatrix} \right.\right) G_{2,2}^{1,2}\left(x \left| \begin{smallmatrix} 1,1 \\ 1,0 \end{smallmatrix} \right.\right) G_{0,1}^{1,0}\left(\frac{x}{\rho} \left| \begin{smallmatrix} - \\ 0 \end{smallmatrix} \right.\right) dx \\ &\stackrel{(a)}{=} 2 \exp\left(\frac{1}{\rho}\right) G_{2,3}^{3,0}\left(\frac{1}{\rho} \left| \begin{smallmatrix} 1,1 \\ 0,0,0 \end{smallmatrix} \right.\right) \end{aligned}$$

where (a) follows from [153, Eq. (07.34.21.0081.01)] after basic simplifications. By combining this expression with the relationship  $\sigma^2 = \mathbb{E}\{(\ln(1 + \text{SNR}_{g,b}))^2\} - (\mathbb{E}\{\ln(1 + \text{SNR}_{g,b})\})^2$ , we obtain (2.20).

To derive the integral form of  $\mathcal{H}_G$ , we calculate the CDF of  $A'_{\min} \triangleq \min\{A'_1, \dots, A'_G\}$  to be

$$\begin{aligned} F_{A'_{\min}}(y) &= 1 - \Pr\{\min\{A'_1, \dots, A'_G\} > y\} \\ &= 1 - (\Pr\{A'_1 > y\})^G \stackrel{(a)}{=} 1 - (\mathcal{Q}(y))^G, \end{aligned} \quad (\text{A.23})$$

where (a) holds because the CDF of the standard normal distribution is  $F_{A'_i}(x) = 1 - \mathcal{Q}(x)$ . The corresponding PDF is then derived by

$$f_{A'_{\min}}(y) = \frac{\partial F_{A'_{\min}}(y)}{\partial y} = -G(\mathcal{Q}(y))^{G-1} \frac{\partial \mathcal{Q}(y)}{\partial y} \stackrel{(a)}{=} \frac{1}{\sqrt{2\pi}} G(\mathcal{Q}(y))^{G-1} \exp\left(-\frac{y^2}{2}\right),$$

where (a) follows from the integral form of the Q-function and by applying the Leibniz's Rule for differentiation under the integral sign. The value of  $\mathcal{H}_G$  in (2.21) is then obtained by writing the expectation of  $A'_{\min}$  as an integral form by using the above PDF of  $A'_{\min}$ .  $\square$

## A.5 Proof of Lemma 2.9

In the following, we prove Lemma 2.9, i.e., we obtain the exact expression

$$\mathring{T}_{\text{ACC}} \triangleq \rho^{-1} \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)} \mathbb{E}\{T_\Psi\}. \quad (\text{A.24})$$

For that, we need to obtain  $\mathbb{E}\{T_\Psi\}$ , and, since  $T_\Psi$  has non-negative support, it follows that  $\mathbb{E}\{T_\Psi\} = \int_0^\infty (1 - F_{T_\Psi}(z)) dz$ . Consequently, we have to obtain  $F_{T_\Psi}(z)$ .

Let us define  $\tau_g \triangleq \sum_{b=1}^B |h_{g,b}|^{-2}$  for the sake of readability, such that we can write  $T_{\mathcal{G}}$  as  $T_\Psi = \max_{g \in \Psi} \{\tau_g\}$ . Then, the CDF of  $T_\Psi$  can be obtained as

$$F_{T_\Psi}(z) = \Pr\left\{\max_{g \in \Psi} \{\tau_g\} \leq z\right\} = [F_{\tau_g}(y)]^G. \quad (\text{A.25})$$

Since  $|h_{g,b}|^2 \sim \text{Gamma}(m, 1)$ ,  $1/|h_{g,b}|^2$  follows an inverse Gamma distribution, it follows that (cf. [82])

$$\text{CF}_{1/|h_{g,b}|^2}(t) = \frac{2(-jt)^{\frac{m}{2}}}{\Gamma(m)} \text{K}_m(\sqrt{-4jt}). \quad (\text{A.26})$$

From (A.26), the CF of  $\tau_g$  can be expressed as

$$\begin{aligned} \text{CF}_{\tau_g}(t) &= \mathbb{E} \left\{ \exp \left( jt \sum_{b=1}^B \frac{1}{|h_{g,b}|^2} \right) \right\} \\ &= \mathbb{E} \left\{ \prod_{b=1}^B \exp \left( \frac{jt}{|h_{g,b}|^2} \right) \right\} = \left( \frac{2(-jt)^{\frac{m}{2}} \text{K}_m(\sqrt{-4jt})}{\Gamma(m)} \right)^B. \end{aligned} \quad (\text{A.27})$$

We can apply the Gil-Pelaez Theorem [147] to obtain the CDF of  $\tau_g$  from its CF, which yields

$$\begin{aligned} F_{\tau_g}(y) &= \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\text{Im} \left\{ \exp(-jty) \text{CF}_{\tau_g}(t) \right\}}{t} dt \\ &= \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \text{Im} \left\{ \frac{\exp(-jty)}{t} \left( \frac{2(-jt)^{\frac{m}{2}} \text{K}_m(\sqrt{-4jt})}{\Gamma(m)} \right)^B \right\} dt. \end{aligned} \quad (\text{A.28})$$

By plugging this expression in (A.25), we obtain  $F_{T_\Psi}(z)$ . Next, we apply the facts that  $\mathbb{E}\{T_\Psi\} = \int_0^\infty (1 - F_{T_\Psi}(z)) dz$  and that  $\hat{T}_{\text{ACC}} \triangleq \rho^{-1} \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)} \mathbb{E}\{T_\Psi\}$  to obtain (2.36).  $\square$

## A.6 Proof of Lemma 2.14

From (2.47), we can write that

$$\tilde{R}^{\text{ACC}} = G \ln \rho + \frac{G}{B} \mathbb{E}_{h,r} \left\{ \min_{g \in \Psi} \left\{ \sum_{b=1}^B \ln(|h_{g,b}|^2 r_{g,b}^{-\eta_0}) \right\} \right\}. \quad (\text{A.29})$$

For  $X_{g,b} = \ln(|h_{g,b}|^2 r_{g,b}^{-\eta_0})$ , the characteristic function (CF)  $\text{CF}_{X_{g,b}}(t) = \mathbb{E}_{h,r} \{ \exp(jt X_{g,b}) \}$  can be shown to be

$$\text{CF}_{X_{g,b}}(t) = \mathbb{E}_h \left\{ (|h_{g,b}|^2)^{jt} \right\} \mathbb{E}_r \left\{ r_{g,b}^{-j\eta_0 t} \right\}. \quad (\text{A.30})$$

By substituting the PDFs of  $|h_{g,b}|^2$  and  $r_{g,b}$  into (A.30), we have

$$\text{CF}_{X_{g,b}}(t) = \Gamma(1+jt) \frac{j2(D_2^{2-j\eta_0 t} - D_1^{2-j\eta_0 t})}{(\eta_0 t + j2)(D_2^2 - D_1^2)}. \quad (\text{A.31})$$

Let us define  $X_g \triangleq \sum_{b=1}^B X_{g,b}$ . The CF of  $X_g$  takes the form  $\text{CF}_{X_g}(t) = \prod_{b=1}^B \text{CF}_{X_{g,b}}(t)$ . By using the Gil-Pelaez Theorem [147], the CDF of  $X_g$  can be obtained as

$$\begin{aligned}
 F_{X_g}(x) &= \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\text{Im}\{\exp(-jtx)\text{CF}_{X_g}(t)\}}{t} dt \\
 &= \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \Im \left\{ \frac{\exp(-jtx)}{t} \left[ \Gamma(1+jt) \frac{j2(D_2^{2-\eta_0 t} - D_1^{2-\eta_0 t})}{(\eta_0 t + j2)(D_2^2 - D_1^2)} \right]^B \right\} dt. \quad (\text{A.32})
 \end{aligned}$$

Let  $X \triangleq \min_{g \in \Psi} \{X_g\}$ . The CDF of  $X$  is

$$F_X(x) = 1 - \Pr \left\{ \min_{g \in \mathcal{G}} \{X_g\} > x \right\} = 1 - (1 - F_{X_g}(x))^G. \quad (\text{A.33})$$

As  $D_1^{\eta_0} \gg 1$  and  $|h_{g,b}|^2 \sim \text{Exp}(1)$ , the probability that  $\ln(|h_{g,b}|^2 r_{g,b}^{-\eta_0})$  is bigger than zero is negligible,<sup>3</sup> so we can consider the variable  $X = \min_{g \in \mathcal{G}} \left\{ \sum_{b=1}^B \ln(|h_{g,b}|^2 r_{g,b}^{-\eta_0}) \right\}$  to be a non-positive random variable. Hence  $X' = -X$  satisfies

$$\mathbb{E}\{X'\} = \int_0^\infty F_X(-x) dx = \int_0^\infty 1 - (1 - F_{X_g}(-x))^G dx. \quad (\text{A.34})$$

Combining (A.29), (A.32) and (A.34), we obtain (2.56).  $\square$

## A.7 Proof of Lemma 2.15

It is clear that the variables  $\{S_{g,b}\}_{g \in \Psi, b \in [B]}$  are i.i.d. variables, and thus we can apply the CLT to get

$$\frac{1}{B} \sum_{b=1}^B \ln \left( \rho |h_{g,b}|^2 r_{g,b}^{-\eta_0} \right) \xrightarrow{d_i} \mathcal{N} \left( \varrho_s, \frac{\sigma_s^2}{B} \right), \text{ as } B \rightarrow \infty, \quad (\text{A.35})$$

where  $d$ . stands for the convergence in distribution. Let  $Y_g, g \in \Psi$ , be i.i.d. Gaussian random variables with zero mean and unit variance. It then follows from (A.35) that (cf. [51, App. IV])

$$\min_{g \in \Psi} \frac{1}{B} \sum_{b=1}^B \ln \left( \rho |h_{g,b}|^2 r_{g,b}^{-\eta_0} \right) \xrightarrow{d_i} \varrho_s + \sqrt{\frac{\sigma_s^2}{B}} \min_{g \in \Psi} \{Y_g\}. \quad (\text{A.36})$$

By considering the definition of  $\tilde{R}^{\text{ACC}}$ , (A.36) implies that

$$\tilde{R}^{\text{ACC}} = \mathbb{E}_{h,r} \left\{ |\mathcal{G}| \min_{g \in \mathcal{G}} \left\{ \frac{1}{B} \sum_{b=1}^B \ln(\text{SNR}_{g,b}) \right\} \right\} \quad (\text{A.37})$$

$$\rightarrow G \left( \varrho_s + \sqrt{\sigma_s^2/B} \mathbb{E}_{h,r} \left\{ \min_{g \in \Psi} \{Y_g\} \right\} \right) \quad (\text{A.38})$$

<sup>3</sup>The minimum distance between the base station and a user is generally considered to be at least 10 meters [88, 90]. Even considering that all the users were located in the inner border of the ring, the probability  $\Pr \{|h_{g,b}|^2 r_{g,b}^{-\eta_0} > 1\}$  is as small as  $\exp(-100) = 3.72 * 10^{-44}$  (for  $\eta_0 = 2$ ).

as  $B \rightarrow \infty$ , which together with the definition of  $-\mathcal{H}_G$  yields (2.57). Note that  $\varrho_s \triangleq \mathbb{E}_{h,r} \{\ln(\rho |h_{g,b}|^2 r_{g,b}^{-\eta_0})\}$  is equivalent to  $\tilde{R}_{\text{TDM}}$  (cf. (2.49)) and thus (2.58) follows from (2.48). For  $\sigma_s^2$ , it holds that

$$\begin{aligned} \sigma_s^2 &= \text{Var} \left\{ \ln \rho + \ln \left( |h_{g,b}|^2 r_{g,b}^{-\eta_0} \right) \right\} = \text{Var} \left\{ \ln \left( |h_{g,b}|^2 r_{g,b}^{-\eta_0} \right) \right\} \\ &= \text{Var} \left\{ \ln \left( |h_{g,b}|^2 \right) \right\} + \eta_0^2 \text{Var} \left\{ \ln r_{g,b} \right\}. \end{aligned} \quad (\text{A.39})$$

We can then derive  $\text{Var} \left\{ \ln \left( |h_{g,b}|^2 \right) \right\}$  and  $\text{Var} \left\{ \ln r_{g,b} \right\}$  in (A.39) from the PDFs of  $|h_{g,b}|^2$  and  $r_{g,b}$ , which yields (2.59).  $\square$

## Appendix B

# Proofs in Chapter 3

### B.1 Proof of Theorem 3.1

Similar to the proof of [65, Lemma 1], we define  $X \triangleq \frac{P_t}{GcL^2} |\mathbf{h}_{\psi,k}^T \mathbf{h}_{\psi,k}^*|^2$  and  $Y \triangleq 1 + \frac{1}{Q} \sum_{\vartheta=1, \vartheta \neq k}^Q Y_\vartheta$ , where  $Y_\vartheta \triangleq \frac{P_t}{GL} |\mathbf{h}_{\psi,k}^T \mathbf{h}_{\psi,\vartheta}^*|^2$ . From [63, Lemma 1], we know that  $\mathbb{E}\{X\} = \frac{P_t}{cG}(1 + 1/L)$ ,  $\text{Var}\{X\} = \frac{P_t^2}{G^2c^2} \left( \frac{4}{L} + \frac{10}{L^2} + \frac{6}{L^3} \right) < \infty$ ,  $\mathbb{E}\{Y_\vartheta\} = P_t/G$  and  $\text{Var}\{Y_\vartheta\} = \frac{P_t^2}{G^2}(1 + 2/L) < \infty$ . We want to prove that

$$\frac{\bar{R}^{\text{MF}}(G, cL)}{cGL} = \mathbb{E} \left\{ \ln \left( 1 + \frac{X}{Y} \right) \right\} = \ln \left( 1 + \frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}} \right) + o(1), \text{ as } Q = cL \rightarrow \infty. \quad (\text{B.1})$$

By applying Jensen's inequality on  $\mathbb{E}\{\ln(X+Y)\}$  and  $\mathbb{E}\{\ln(Y)\}$  separately, we can get

$$\ln \left( \frac{1}{\mathbb{E}\{(X+Y)^{-1}\}} \right) \leq \mathbb{E}\{\ln(X+Y)\} \leq \ln(\mathbb{E}\{X+Y\}) \quad (\text{B.2})$$

$$-\ln(\mathbb{E}\{Y\}) \leq -\mathbb{E}\{\ln(Y)\} \leq -\ln \left( \frac{1}{\mathbb{E}\{Y^{-1}\}} \right), \quad (\text{B.3})$$

and after combining these two bounds, we get

$$\ln \left( \frac{1}{\mathbb{E}\{(X+Y)^{-1}\}} \right) - \ln(\mathbb{E}\{Y\}) \leq \mathbb{E}\{\ln(1 + \frac{X}{Y})\} \leq \ln(\mathbb{E}\{X+Y\}) - \ln \left( \frac{1}{\mathbb{E}\{Y^{-1}\}} \right). \quad (\text{B.4})$$

On the other hand, Jensen's inequality says that  $\mathbb{E}\{Y^{-1}\} \geq 1/\mathbb{E}\{Y\}$  and  $\mathbb{E}\{(X+Y)^{-1}\} \geq 1/\mathbb{E}\{X+Y\}$ , which yields

$$\ln \left( 1 + \frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}} \right) = \ln(\mathbb{E}\{X+Y\}) - \ln(\mathbb{E}\{Y\}) \leq \ln(\mathbb{E}\{X+Y\}) - \ln \left( \frac{1}{\mathbb{E}\{Y^{-1}\}} \right), \quad (\text{B.5})$$

$$\ln \left( 1 + \frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}} \right) = \ln(\mathbb{E}\{X+Y\}) - \ln(\mathbb{E}\{Y\}) \geq \ln \left( \frac{1}{\mathbb{E}\{(X+Y)^{-1}\}} \right) - \ln(\mathbb{E}\{Y\}). \quad (\text{B.6})$$

At this point, both  $\mathbb{E}\{\ln(1 + \frac{X}{Y})\}$  and  $\ln\left(1 + \frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}\right)$  are bounded above and below by the same bounds (B.4)–(B.6). The gap between these bounds takes the form

$$\begin{aligned} \Delta &\triangleq \left\{ \ln(\mathbb{E}\{X + Y\}) - \ln\left(\frac{1}{\mathbb{E}\{Y^{-1}\}}\right) \right\} - \left\{ \ln\left(\frac{1}{\mathbb{E}\{(X + Y)^{-1}\}}\right) - \ln(\mathbb{E}\{Y\}) \right\} \\ &= \ln\left[\left(\mathbb{E}\{X + Y\} \mathbb{E}\{(X + Y)^{-1}\}\right) \left(\mathbb{E}\{Y\} \mathbb{E}\{Y^{-1}\}\right)\right]. \end{aligned} \quad (\text{B.7})$$

We want to show that this gap vanishes as  $Q = cL \rightarrow \infty$ . By expanding the Taylor series of  $Y^{-1}$  at  $\mathbb{E}\{Y\}$ , we can have that

$$\begin{aligned} \lim_{Q \rightarrow \infty} \mathbb{E}\{Y\} \mathbb{E}\{Y^{-1}\} &= \lim_{Q \rightarrow \infty} \mathbb{E}\{Y\} \mathbb{E}\left\{ \frac{1}{\mathbb{E}\{Y\}} - \frac{(Y - \mathbb{E}\{Y\})}{\mathbb{E}^2\{Y\}} + \frac{(Y - \mathbb{E}\{Y\})^2}{\mathbb{E}^3\{Y\}} + \dots \right\} \\ &= 1 + \lim_{Q \rightarrow \infty} \mathbb{E}\{g(Y)\} \stackrel{(a)}{=} 1 + \mathbb{E}\left\{ \lim_{Q \rightarrow \infty} g(Y) \right\} \stackrel{(b)}{=} 1, \end{aligned} \quad (\text{B.8})$$

where  $g(Y) \triangleq \sum_{n=2}^{\infty} (-1)^n \frac{(Y - \mathbb{E}\{Y\})^n}{\mathbb{E}^n\{Y\}}$ , where (a) follows from exchanging the order of the limitation and expectation operators (validated via the Dominated Convergence Theorem (DCT))<sup>1</sup>, and where (b) follows from using the DCT to exchange the limitation and infinite summation operators in  $\lim_{Q \rightarrow \infty} g(Y)$  (similar to the step (a)) and then by considering that  $Y - \mathbb{E}\{Y\} \rightarrow 0$  as  $Q \rightarrow \infty$  (due to the law of large numbers). By using similar mathematical manipulations, we have that

$$\lim_{Q=cL \rightarrow \infty} \mathbb{E}\{X + Y\} \mathbb{E}\{(X + Y)^{-1}\} = 1. \quad (\text{B.9})$$

Considering the two limits (B.8) and (B.9), we can directly conclude that  $\lim_{Q=cL \rightarrow \infty} \Delta = 0$ , and therefore prove (B.1).

Finally, substituting  $\mathbb{E}\{X\} = \frac{P_t}{cG}(1 + \frac{1}{L})$  and  $\mathbb{E}\{Y\} = 1 + \frac{P_t}{G} \frac{Q-1}{Q}$  into (B.1) and considering  $Q = cL \rightarrow \infty$ , completes the proof of Theorem 3.1.

## B.2 Proof of Theorem 3.3

We split the proof in three parts. First, we present the proof of (3.19). Then, we provide two useful lemmas, and we conclude by deriving the asymptotic deterministic equivalent of the SINR.

We provide here the proof of the expression of  $\text{SINR}_{\psi,k}^{\text{RZF}}$  in (3.19). Let us recall that  $\mathbf{H}_{\psi,-k}$  represents the matrix  $\mathbf{H}_{\psi}$  after removing its  $k$ -th row. The useful signal contribution to the received signal in (3.16) (omitting the term  $\rho_{\psi}/\sqrt{G}$  for the sake of conciseness) can be written as

$$\mathbf{h}_{\psi,k}^T (\alpha \mathbf{I}_L + \mathbf{H}_{\psi}^H \mathbf{H}_{\psi})^{-1} \mathbf{h}_{\psi,k}^* s_{\psi,k} = \mathbf{h}_{\psi,k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} + \mathbf{h}_{\psi,k}^* \mathbf{h}_{\psi,k}^T \right)^{-1} \mathbf{h}_{\psi,k}^* s_{\psi,k}$$

<sup>1</sup>To see this, first define  $Z \triangleq |Y - \mathbb{E}\{Y\}| \geq 0$ . As  $Q \rightarrow \infty$ ,  $Z \rightarrow 0$  (due to the law of large numbers), there always exists a constant  $Q_0$  and  $\varepsilon < 1$  such that  $Z < \varepsilon$  for any  $Q > Q_0$ . For  $Z < \varepsilon$ , we have that  $\sum_{n=2}^{\infty} Z^n = \frac{Z^2}{1-Z} < \frac{\varepsilon^2}{1-\varepsilon}$ . Considering  $g(Y) \leq \sum_{n=2}^{\infty} Z^n$  and  $\mathbb{E}\{\sum_{n=2}^{\infty} Z^n\} < \frac{\varepsilon^2}{1-\varepsilon} < \infty$ , which satisfies the DCT condition, yields that  $\lim_{Q \rightarrow \infty} \mathbb{E}\{g(Y)\} = \mathbb{E}\{\lim_{Q \rightarrow \infty} g(Y)\}$ .

$$\stackrel{(a)}{=} \frac{\mathbf{h}_{\psi,k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-1} \mathbf{h}_{\psi,k}^*}{1 + \mathbf{h}_{\psi,k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-1} \mathbf{h}_{\psi,k}^*} s_{\psi,k} \stackrel{(b)}{=} \frac{A_{\psi,k}}{1 + A_{\psi,k}} s_{\psi,k}, \quad (\text{B.10})$$

where (a) follows from the relation

$$\left( \mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C} \right)^{-1} \mathbf{B} \mathbf{D}^{-1} = \mathbf{A}^{-1} \mathbf{B} \left( \mathbf{D} - \mathbf{C} \mathbf{A}^{-1} \mathbf{B} \right)^{-1}, \quad (\text{B.11})$$

and where (b) follows after applying the definition of  $A_{\psi,k}$  from (3.17).

On the other hand, the power of the interference averaged over data signals in (3.16) is given by

$$\begin{aligned} |I_{\psi,k}|^2 &= \frac{\rho_\psi^2}{G} \sum_{\substack{\vartheta=1 \\ \vartheta \neq k}}^L \sum_{\substack{\vartheta'=1 \\ \vartheta' \neq k}}^L \mathbf{h}_{\psi,\vartheta}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_\psi^H \mathbf{H}_\psi \right)^{-1} \mathbf{h}_{\psi,k}^* \mathbf{h}_{\psi,k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_\psi^H \mathbf{H}_\psi \right)^{-1} \mathbf{h}_{\psi,\vartheta'}^* \mathbb{E} \{ s_{\psi,\vartheta}^* s_{\psi,\vartheta'} \} \\ &= \frac{\rho_\psi^2}{G} \mathbf{h}_{\psi,k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_\psi^H \mathbf{H}_\psi \right)^{-1} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \left( \alpha \mathbf{I}_L + \mathbf{H}_\psi^H \mathbf{H}_\psi \right)^{-1} \mathbf{h}_{\psi,k}^*. \end{aligned} \quad (\text{B.12})$$

By applying again the matrix identity in (B.11) and by considering the definitions of  $A_{\psi,k}$  and  $B_{\psi,k}$  in (3.17)-(3.18), we can obtain

$$\begin{aligned} |I_{\psi,k}|^2 &= \frac{\rho_\psi^2 \mathbf{h}_{\psi,k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-1} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-1} \mathbf{h}_{\psi,k}^*}{G \left( 1 + \mathbf{h}_{\psi,k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-1} \mathbf{h}_{\psi,k}^* \right)^2} \\ &= \frac{B_{\psi,k} \rho_\psi^2 / G}{(1 + A_{\psi,k})^2} \end{aligned}$$

which combined with (B.10) yields the expression of  $\text{SINR}_{\psi,k}^{\text{RZF}}$  in (3.19). This concludes the proof.

### B.2.1 Two Useful Lemmas

In the following, we present two lemmas that are instrumental in the derivation of Lemma 3.3.

**Lemma B.1.** *For any fixed  $c$ ,  $0 < c < \infty$ , the trace of  $\frac{1}{L} \left( z \mathbf{I}_L + \frac{1}{L} \mathbf{H}_\psi^H \mathbf{H}_\psi \right)^{-1}$  converges to  $S_c(z)$  almost surely as  $L \rightarrow \infty$ , where  $S_c(z)$  is defined as*

$$S_c(z) \triangleq \frac{1}{2} \left( \sqrt{\frac{(1-c)^2}{z^2} + \frac{2(1+c)}{z}} + 1 + \frac{1-c}{z} - 1 \right). \quad (\text{B.13})$$

*Proof.* This lemma can be obtained as a direct application of a known result from [154, Ch. 3] for the Stieltjes transform [155]. Hence, we omit the proof due to the page limitation and refer the reader to [154, Ch. 3] for more details.  $\square$

**Lemma B.2.** For any fixed  $0 < c < \infty$  and arbitrary  $0 < \theta < \infty$ , we have that, as  $L \rightarrow \infty$ ,

$$\mathrm{Tr} \left\{ \frac{1}{L} \left( \theta \mathbf{I} + \frac{1}{L} \mathbf{H}_{\psi, -k}^H \mathbf{H}_{\psi, -k} \right)^{-2} \right\} \xrightarrow{a.s.} \mathrm{Tr} \left\{ \frac{1}{L} \left( \theta \mathbf{I} + \frac{1}{L} \mathbf{H}_{\psi}^H \mathbf{H}_{\psi} \right)^{-2} \right\}. \quad (\text{B.14})$$

*Proof.* Let us first define  $\mathbf{A} \triangleq \theta \mathbf{I} + \frac{1}{L} \mathbf{H}_{\psi, -k}^H \mathbf{H}_{\psi, -k}$ , and let us also define

$$\delta \triangleq \left| \mathrm{Tr} \left\{ \frac{1}{L} \left( \theta \mathbf{I} + \frac{1}{L} \mathbf{H}_{\psi, -k}^H \mathbf{H}_{\psi, -k} \right)^{-2} \right\} - \mathrm{Tr} \left\{ \frac{1}{L} \left( \theta \mathbf{I} + \frac{1}{L} \mathbf{H}_{\psi}^H \mathbf{H}_{\psi} \right)^{-2} \right\} \right|. \quad (\text{B.15})$$

By applying the Woodbury matrix identity [156], we can rewrite  $\delta$  as

$$\delta = \left| \frac{1}{L} \mathrm{Tr} \left\{ \frac{2}{L} \frac{\mathbf{h}_k^T \mathbf{A}^{-3} \mathbf{h}_k^*}{1 + \frac{1}{L} \mathbf{h}_k^T \mathbf{A}^{-1} \mathbf{h}_k^*} - \frac{1}{L^2} \frac{(\mathbf{h}_k^T \mathbf{A}^{-2} \mathbf{h}_k^*)(\mathbf{h}_k^T \mathbf{A}^{-2} \mathbf{h}_k^*)}{\left(1 + \frac{1}{L} \mathbf{h}_k^T \mathbf{A}^{-1} \mathbf{h}_k^*\right)^2} \right\} \right|, \quad (\text{B.16})$$

which can be further rewritten as  $\delta = |\Theta_1 - \Theta_2|$ , where  $\Theta_1 \triangleq \frac{2}{L^2} \frac{\mathbf{h}_k^T \mathbf{A}^{-3} \mathbf{h}_k^*}{1 + \frac{1}{L} \mathbf{h}_k^T \mathbf{A}^{-1} \mathbf{h}_k^*}$ , and  $\Theta_2 \triangleq \frac{1}{L^3} \left( \frac{\mathbf{h}_k^T \mathbf{A}^{-2} \mathbf{h}_k^*}{1 + \frac{1}{L} \mathbf{h}_k^T \mathbf{A}^{-1} \mathbf{h}_k^*} \right)^2$ . Furthermore, we apply eigenvalue decomposition by factorizing  $\frac{1}{L} \mathbf{H}_{\psi, -k}^H \mathbf{H}_{\psi, -k}$  as  $\frac{1}{L} \mathbf{H}_{\psi, -k}^H \mathbf{H}_{\psi, -k} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^H$ , which yields  $\mathbf{A}^{-1} = \mathbf{Q} (\theta \mathbf{I} + \mathbf{\Lambda})^{-1} \mathbf{Q}^H$ , and  $\mathbf{A}^{-3} = \mathbf{Q} (\theta \mathbf{I} + \mathbf{\Lambda})^{-3} \mathbf{Q}^H$ . Thus, upon defining  $\mathbf{g} \triangleq \mathbf{Q} \mathbf{h}_k^* / \sqrt{L}$ , the term  $\Theta_1$  can be rewritten as

$$\begin{aligned} \Theta_1 &= \frac{2}{L^2} \frac{\mathbf{h}_k^T \mathbf{Q} (\theta \mathbf{I} + \mathbf{\Lambda})^{-3} \mathbf{Q}^H \mathbf{h}_k^*}{1 + \frac{1}{L} \mathbf{h}_k^T \mathbf{Q} (\theta \mathbf{I} + \mathbf{\Lambda})^{-1} \mathbf{Q}^H \mathbf{h}_k^*} = \frac{2}{L} \frac{\mathbf{g}^H (\theta \mathbf{I} + \mathbf{\Lambda})^{-3} \mathbf{g}}{1 + \mathbf{g}^H (\theta \mathbf{I} + \mathbf{\Lambda})^{-1} \mathbf{g}} \\ &= \frac{2}{L} \frac{\sum_{\ell=1}^L |g_{\ell}|^2 \frac{1}{(\theta + \lambda_{\ell})^3}}{1 + \sum_{\ell=1}^L |g_{\ell}|^2 \frac{1}{\theta + \lambda_{\ell}}} \leq \frac{2}{\theta^2 L} \frac{\sum_{\ell=1}^L |g_{\ell}|^2 \frac{1}{\theta + \lambda_{\ell}}}{1 + \sum_{\ell=1}^L |g_{\ell}|^2 \frac{1}{\theta + \lambda_{\ell}}} \leq \frac{2}{\theta^2 L} \rightarrow 0, \text{ as } L \rightarrow \infty, \end{aligned} \quad (\text{B.17})$$

where  $g_{\ell}$  and  $\lambda_{\ell}$  are the  $\ell$ -th element of  $\mathbf{g}$  and the  $\ell$ -th eigenvalue of  $\frac{1}{L} \mathbf{H}_{\psi, k}^H \mathbf{H}_{\psi, k}$ , respectively. Similarly, we have that

$$\Theta_2 = \frac{1}{L} \left( \frac{\mathbf{g}^H (\theta \mathbf{I} + \mathbf{\Lambda})^{-2} \mathbf{g}}{1 + \mathbf{g}^H (\theta \mathbf{I} + \mathbf{\Lambda})^{-1} \mathbf{g}} \right)^2 \leq \frac{1}{\theta^2 L} \left( \frac{\sum_{\ell=1}^L |g_{\ell}|^2 \frac{1}{\theta + \lambda_{\ell}}}{1 + \sum_{\ell=1}^L |g_{\ell}|^2 \frac{1}{\theta + \lambda_{\ell}}} \right)^2 \leq \frac{1}{\theta^2 L} \rightarrow 0, \text{ as } L \rightarrow \infty. \quad (\text{B.18})$$

Finally, from (B.17), (B.18), and from the fact that  $\delta \leq |\Theta_1| + |\Theta_2|$ , the difference  $\delta$  approaches zero almost surely as  $L \rightarrow \infty$ . This concludes the proof of Lemma B.2.  $\square$

### B.2.2 Proof of Theorem 3.3

We obtain Theorem 3.3 by deriving the asymptotic deterministic equivalent of  $\mathrm{SINR}_{\psi, k}$  in (3.19). For that, we first derive the asymptotic deterministic equivalent of  $A_{\psi, k}$  and  $\rho_{\psi}^2$ .

Let us start by considering  $A_{\psi, k}$ , defined in (3.17). By means of the Trace Lemma and the Rank-1 Perturbation Lemma from [154], we can obtain that

$$A_{\psi, k} = \mathbf{h}_{\psi, k}^T \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi, -k}^H \mathbf{H}_{\psi, -k} \right)^{-1} \mathbf{h}_{\psi, k}^* \xrightarrow{a.s.} \mathrm{Tr} \left\{ \left( \alpha \mathbf{I}_L + \mathbf{H}_{\psi}^H \mathbf{H}_{\psi} \right)^{-1} \right\} \quad (\text{B.19})$$

as  $L \rightarrow \infty$ . From this, we can apply Lemma B.1 and the fact that  $\alpha = L/P_t$  to obtain the deterministic equivalent of  $A_{\psi,k}$ , which we denote as  $a_{\psi,k}$ , and which is given by  $a_{\psi,k} = S_c\left(\frac{1}{P_t}\right)$ , where  $S_c(z) = \frac{1}{2} \left[ \sqrt{\frac{(1-c)^2}{z^2} + \frac{2(1+c)}{z}} + 1 + \frac{1-c}{z} - 1 \right]$  as defined in (B.13). This yields the expression of  $a_{\psi,k}$  in (3.21).

Next, we focus on  $B_{\psi,k}$ , introduced in (3.18), and we again apply the Trace Lemma and the Rank-1 Perturbation Lemma from [154] in the limit of  $L \rightarrow \infty$  to obtain that

$$\begin{aligned} B_{\psi,k} &\xrightarrow{a.s.} \frac{1}{L} \text{Tr} \left\{ \frac{1}{L} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \left( \frac{1}{P_t} \mathbf{I}_L + \frac{1}{L} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-2} \right\} \\ &= \frac{1}{L} \text{Tr} \left\{ \left( \frac{1}{P_t} \mathbf{I}_L + \frac{1}{L} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-1} \right\} - \frac{1}{P_t L} \text{Tr} \left\{ \left( \frac{1}{P_t} \mathbf{I}_L + \frac{1}{L} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-2} \right\}. \end{aligned} \quad (\text{B.20})$$

The first trace term of the R.H.S. of (B.20) matches (B.19), and thus its deterministic equivalent is  $a_{\psi,k}$ . With respect to the second term of the R.H.S. of (B.20), applying Lemmas B.1 and B.2 yields

$$\begin{aligned} &\frac{1}{P_t L} \text{Tr} \left\{ \left( \frac{1}{P_t} \mathbf{I}_L + \frac{1}{L} \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k} \right)^{-2} \right\} \xrightarrow{a.s.} \frac{1}{P_t L} \text{Tr} \left\{ \left( \frac{1}{P_t} \mathbf{I}_L + \frac{1}{L} \mathbf{H}_{\psi}^H \mathbf{H}_{\psi} \right)^{-2} \right\} \\ &= \frac{1}{P_t L} \sum_{\ell=1}^L \frac{1}{(\lambda_{\ell} + 1/P_t)^2} = -\frac{1}{P_t} \frac{\partial}{\partial z} \left( \frac{1}{L} \sum_{\ell=1}^L \frac{1}{\lambda_{\ell} + z} \right) \Big|_{z=1/P_t} \\ &= -\frac{1}{P_t} \frac{\partial}{\partial z} \left( \text{Tr} \left\{ \frac{1}{L} \left( z \mathbf{I}_L + \frac{1}{L} \mathbf{H}_{\psi}^H \mathbf{H}_{\psi} \right)^{-1} \right\} \right) \Big|_{z=1/P_t} \xrightarrow{a.s.} -\frac{1}{P_t} \frac{\partial S_c(z)}{\partial z} \Big|_{z=1/P_t}, \end{aligned} \quad (\text{B.21})$$

as  $L \rightarrow \infty$ , where  $\{\lambda_{\ell}\}_{\ell=1}^L$  are the eigenvalues of  $\frac{1}{L} \mathbf{H}_{\psi}^H \mathbf{H}_{\psi}$  and where  $\frac{\partial S_c(z)}{\partial z}$  is the derivative of  $S_c(z)$  with respect to  $z$ , which is given by

$$\frac{\partial S_c(z)}{\partial z} = \frac{1}{2} \left[ \frac{-c^2 - c(z-2) - z - 1}{z^2 \sqrt{c^2 + 2c(z-1) + (z+1)^2}} - \frac{1-c}{z^2} \right]. \quad (\text{B.22})$$

From (B.20) and (B.21) it holds that  $B_{\psi,k} \xrightarrow{a.s.} b_{\psi,k} \triangleq a_{\psi,k} + \frac{1}{P_t} \frac{\partial S_c(z)}{\partial z} \Big|_{z=1/P_t}$  as  $L \rightarrow \infty$ .

Finally, we focus on the power control factor for the RZF precoder, which was given by  $\rho_{\psi}^2 = \frac{P_t}{\frac{1}{L} \text{Tr} \left\{ \frac{1}{L} \mathbf{H}_{\psi}^H \mathbf{H}_{\psi} \left( \frac{1}{L} \mathbf{H}_{\psi}^H \mathbf{H}_{\psi} + \frac{1}{P_t} \mathbf{I}_L \right)^{-2} \right\}}$ . From the derivation of  $B_{\psi,k}$  and (B.20)–(B.21), it follows that  $\rho_{\psi}^2 \xrightarrow{a.s.} \frac{P_t}{b_{\psi,k}}$ . Thus, the asymptotic deterministic equivalent of  $\rho_{\psi}^2$ , denoted by  $p_{\psi}^2$ , takes the form  $p_{\psi}^2 = \frac{P_t}{b_{\psi,k}}$ , which, upon substituting (B.22) in  $b_{\psi,k}$ , yields the expression of  $p_{\psi}^2$  in (3.22).

Next, we obtain the asymptotic deterministic equivalent of  $\text{SINR}_{\psi,k}$  by substituting the asymptotic deterministic equivalent of  $A_{\psi,k}$ ,  $B_{\psi,k}$  and  $\rho_{\psi}^2$  into (3.19), which yields

$$\text{SINR}_{\psi,k}^{\text{RZF}} \xrightarrow{a.s.} \frac{a_{\psi,k}^2 p_{\psi}^2 / G}{(1 + a_{\psi,k})^2 + \frac{P_t}{G}}. \quad (\text{B.23})$$

Finally, a direct application of the Continuous Mapping Theorem [157] yields (3.20), which concludes the proof of Theorem 3.3.  $\square$

### B.3 Proof of Theorem 3.4

We prove the theorem by demonstrating that  $\bar{\mathcal{R}}^{\text{MF}}$  as derived in Corollary 3.2 is concave over  $c \in (0, \infty)$ . Let us first note that the first derivative of  $\bar{R}^{\text{MF}}$  in (3.10) is given by

$$\frac{\partial \bar{R}^{\text{MF}}}{\partial c} = GL \left[ \ln \left( \frac{\Omega + c}{c} \right) + \frac{c}{\Omega + c} - 1 \right], \quad (\text{B.24})$$

whereas the second derivative is then given by

$$\frac{\partial^2 \bar{R}^{\text{MF}}}{\partial c^2} = -GL \frac{\Omega^2}{c(\Omega + c)^2} < 0. \quad (\text{B.25})$$

By differentiating  $\bar{\mathcal{R}}^{\text{MF}}$  in (3.24) with respect to  $c$ , we have that

$$\frac{\partial \bar{\mathcal{R}}^{\text{MF}}}{\partial c} = (1 - \zeta_{G,Q}c) \frac{\partial \bar{R}^{\text{MF}}}{\partial c} - \zeta_{G,Q} \bar{R}^{\text{MF}}, \quad \frac{\partial^2 \bar{\mathcal{R}}^{\text{MF}}}{\partial c^2} = (1 - \zeta_{G,Q}c) \frac{\partial^2 \bar{R}^{\text{MF}}}{\partial c^2} - 2\zeta_{G,Q} \frac{\partial \bar{R}^{\text{MF}}}{\partial c}. \quad (\text{B.26})$$

Let us now inspect the signs of these derivatives. First, note that  $\frac{\Omega+c}{c} \geq 1$  for any feasible  $\Omega, c$ , simply because  $\Omega = \frac{P_t}{P_t+G} \geq 0$ . Let us also note that the function  $\ln(x) + 1/x$  is decreasing when  $x \in (0, 1)$  and is increasing when  $x \in [1, \infty)$ , and also that its minimum value (attained at  $x = 1$ ) is equal to 1. Consequently, it follows that

$$\frac{\partial \bar{R}^{\text{MF}}}{\partial c} = GL \left[ \ln \left( \frac{\Omega + c}{c} \right) + \frac{c}{\Omega + c} - 1 \right] \geq 0, \quad (\text{B.27})$$

where the inequality is strict unless  $\frac{\Omega+c}{c} = 1$  corresponding to  $c \rightarrow \infty$ . Therefore, we conclude that  $\bar{R}^{\text{MF}}$  is monotonically increasing over  $c \in (0, \infty)$ .

From the fact that  $\frac{\partial \bar{R}^{\text{MF}}}{\partial c} \geq 0$  (cf. (B.27)), the fact that  $\frac{\partial^2 \bar{R}^{\text{MF}}}{\partial c^2} < 0$  (cf. (B.25)), and the fact that  $1 - \zeta_{G,Q}c \geq 0$ , we can conclude that  $\frac{\partial^2 \bar{\mathcal{R}}^{\text{MF}}}{\partial c^2} < 0$  in (B.26). Therefore,  $\bar{\mathcal{R}}^{\text{MF}}$  is concave over  $c \in (0, \infty)$ , and thus the global maximum point of  $\bar{\mathcal{R}}^{\text{MF}}$  is at the root  $c^*$  of  $\frac{\partial \bar{\mathcal{R}}^{\text{MF}}}{\partial c}$ .  $\square$

### B.4 Proof of Theorem 3.5

The proof builds on the properties of the first and second derivatives of  $\bar{\mathcal{R}}^{\text{ZF}}$ , in a similar manner as in the proof of Theorem 3.4. These derivatives now take the form  $\frac{\partial \bar{\mathcal{R}}^{\text{ZF}}}{\partial c} = (1 - \zeta_{G,Q}c) \frac{\partial \bar{R}^{\text{ZF}}}{\partial c} - \zeta_{G,Q} \bar{R}^{\text{ZF}}$ , and  $\frac{\partial^2 \bar{\mathcal{R}}^{\text{ZF}}}{\partial c^2} = (1 - \zeta_{G,Q}c) \frac{\partial^2 \bar{R}^{\text{ZF}}}{\partial c^2} - 2\zeta_{G,Q} \frac{\partial \bar{R}^{\text{ZF}}}{\partial c}$ . After applying (3.15), these derivatives take the form

$$\frac{\partial \bar{R}^{\text{ZF}}}{\partial c} = GL \left[ \ln \left( 1 + \frac{P_t}{G} \left( \frac{1}{c} - 1 \right) \right) - \frac{P_t/G}{(1 - P_t/G)c + P_t/G} \right], \quad (\text{B.28})$$

$$\frac{\partial^2 \bar{R}^{\text{ZF}}}{\partial c^2} = -GL \frac{(P_t/G)^2}{c((1 - P_t/G)c + P_t/G)^2} < 0. \quad (\text{B.29})$$

Since the second derivative  $\frac{\partial^2 \bar{R}^{ZF}}{\partial c^2}$  in (B.29) is always negative,  $\bar{R}^{ZF}$  is a concave function with respect to  $c$ . Therefore, the root of  $\frac{\partial \bar{R}^{ZF}}{\partial c} = 0$ , which we denote by  $c_R^*$ , is the global maximum of  $\bar{R}^{ZF}$  over  $c \in (0, \infty)$ . Moreover, it follows from (3.15) that  $\bar{R}^{ZF} = 0$  for  $c = 1$  and that  $\bar{R}^{ZF} > 0$  for  $0 < c < 1$ , which implies that  $c_R^*$  belongs in the interval  $(0, 1)$ .

Since  $\frac{\partial \bar{R}^{ZF}}{\partial c} \Big|_{c=c_R^*} = 0$  and since  $\frac{\partial^2 \bar{R}^{ZF}}{\partial c^2}$  is always negative, we know that  $\frac{\partial \bar{R}^{ZF}}{\partial c}$  is monotonically decreasing and that this same  $\frac{\partial \bar{R}^{ZF}}{\partial c}$  is negative for all  $c \in (c_R^*, 1)$ .

Consequently,  $\bar{\mathcal{R}}^{ZF}$  is monotonically decreasing in the interval  $c \in (c_R^*, 1)$ . Thus the maximum point of  $\bar{\mathcal{R}}^{ZF}$  must belong in the interval  $(0, c_R^*)$  where we can see that  $\frac{\partial \bar{\mathcal{R}}^{ZF}}{\partial c} > 0$  and  $\frac{\partial^2 \bar{\mathcal{R}}^{ZF}}{\partial c^2} < 0$ . Hence,  $\bar{\mathcal{R}}^{ZF}$  is concave throughout  $c \in (0, c_R^*)$ , and thus the root of  $\frac{\partial \bar{\mathcal{R}}^{ZF}}{\partial c}$  is the global maximum point of  $\bar{\mathcal{R}}^{ZF}$ , where this point  $c^*$  must belong in  $(0, c_R^*)$ . Finally, substituting (3.15) and (B.28) into  $\frac{\partial \bar{\mathcal{R}}^{ZF}}{\partial c}$  yields (3.31) and proves the theorem.  $\square$



## Appendix C

# Proofs in Chapter 4

### C.1 Proof of Lemma 4.1

Let  $\mathbf{\Omega}_{\psi,k} \triangleq \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \in \mathbb{C}^{L \times L}$ , which can be easily seen to be Hermitian semi-definite. Now let us consider the problem of maximizing  $\text{SINR}_{\psi,k,q}$  as seen below:

$$\begin{aligned} \mathbf{m}_{\psi,k,q}^* &= \arg \max_{\mathbf{m}_{\psi,k,q} \in \mathbb{C}^L} \mathbf{m}_{\psi,k,q}^H \mathbf{\Omega}_{\psi,k} \mathbf{m}_{\psi,k,q} \\ \text{s. t. } & \|\mathbf{T}_{\psi,-k} \mathbf{m}_{\psi,k,q}\|^2 = \mathbf{m}_{\psi,k,q}^H \mathbf{T}_{\psi,-k} \mathbf{T}_{\psi,-k} \mathbf{m}_{\psi,k,q} = 1, \end{aligned} \quad (\text{C.1})$$

for which we can easily derive that  $\mathbf{m}_{\psi,k,q}^* = \mathbf{u}_{\psi,k,q} / \|\mathbf{T}_{\psi,-k} \mathbf{u}_{\psi,k,q}\|$ , where  $\mathbf{u}_{\psi,k,q}$  is the unit-norm eigenvector corresponding to the  $q$ -th largest eigenvalue  $\lambda_{\psi,k,q}$  of  $\mathbf{\Omega}_{\psi,k}$ . We can now also see that the maximum of  $\mathbf{m}_{\psi,k,q}^H \mathbf{\Omega}_{\psi,k} \mathbf{m}_{\psi,k,q}$  is  $\frac{\lambda_{\psi,k,q}}{\|\mathbf{T}_{\psi,-k} \mathbf{u}_{\psi,k,q}\|^2}$ , and thus the optimized SINR takes the form

$$\text{SINR}_{\psi,k,q} = \frac{\lambda_{\psi,k,q} P_{\psi,k,q} / N_0}{\|\mathbf{T}_{\psi,-k} \mathbf{u}_{\psi,k,q}\|^2}. \quad (\text{C.2})$$

To ease the problem of computing  $\lambda_{\psi,k,q}$  and  $\mathbf{u}_{\psi,k,q}$ , we can alternatively decompose  $\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \in \mathbb{C}^{M_{\psi,k} \times M_{\psi,k}}$  and then find the  $q$ -th largest eigenvalue and the corresponding eigenvector  $\mathbf{t}_{\psi,k,q}$ ; this is directly from the property that for any two matrices  $\mathbf{A}, \mathbf{B}$ , we have that  $\mathbf{AB}$  and  $\mathbf{BA}$  share the same non-zero eigenvalues. After deriving  $\lambda_{\psi,k,q}$  and  $\mathbf{t}_{\psi,k,q}$ , we can have that

$$\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q} = \lambda_{\psi,k,q} \mathbf{t}_{\psi,k,q}$$

which gives us

$$\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \left( \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q} \right) = \lambda_{\psi,k,q} \left( \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q} \right) \quad (\text{C.3})$$

and we can have

$$\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{u}_{\psi,k,q} = \lambda_{\psi,k,q} \mathbf{u}_{\psi,k,q}$$

which gives us

$$\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} (\mathbf{T}_{\psi,-k} \mathbf{u}_{\psi,k,q}) = \lambda_{\psi,k,q} (\mathbf{T}_{\psi,-k} \mathbf{u}_{\psi,k,q}),$$

where we use that  $\mathbf{T}_{\psi,-k}^H = \mathbf{T}_{\psi,-k} = \mathbf{T}_{\psi,-k}^2$ . This shows that both  $\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}$  and  $\mathbf{u}_{\psi,k,q}$  are the eigenvectors associated to the  $q$ -th largest eigenvalue  $\lambda_{\psi,k,q}$  of  $\mathbf{\Omega}_{\psi,k} = \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k}$ . We can also see that both  $\mathbf{T}_{\psi,-k} \mathbf{u}_{\psi,k,q}$  and  $\mathbf{u}_{\psi,k,q}$  are the eigenvectors associated to  $\lambda_{\psi,k,q}$  of  $\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k}$ . Therefore, the optimal precoding vector  $\mathbf{v}_{\psi,k,q}$  for  $\mathbf{U}_{\psi,k}$  can be rewritten as

$$\mathbf{v}_{\psi,k}^* = \frac{\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}}{\|\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}\|}. \quad (\text{C.4})$$

Under the optimal precoding vector  $\mathbf{v}_{\psi,k,q}^*$  and given an MRC receiver,  $\text{SINR}_{\psi,k,q}$  is derived as

$$\begin{aligned} \text{SINR}_{\psi,k,q} &= \frac{P_{\psi,k,q} \mathbf{t}_{\psi,k,q}^H \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}}{N_0 \|\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}\|^2} \\ &= \frac{P_{\psi,k,q} (\mathbf{t}_{\psi,k,q}^H \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k}) \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} (\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q})}{N_0 \|\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{t}_{\psi,k,q}\|^2} \stackrel{(a)}{=} \frac{P_{\psi,k,q}}{N_0} \lambda_{\psi,k,q}, \end{aligned} \quad (\text{C.5})$$

where (a) follows from (C.3), and where  $\lambda_{\psi,k,q}$  is the  $q$ -th largest eigenvalue of both  $\mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \in \mathbb{C}^{L \times L}$  and  $\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \in \mathbb{C}^{M_{\psi,k} \times M_{\psi,k}}$ .  $\square$

## C.2 Proof of Lemma 4.2

For  $L \rightarrow \infty$  and finite  $M_{\psi,k}$ , we can use the Trace Lemma to derive that

$$\frac{1}{L} \frac{1}{\beta_{\psi,k}} (\mathbf{h}_{\psi,k}^{(\ell)})^T \mathbf{T}_{\psi,-k} (\mathbf{h}_{\psi,k}^{(\ell)})^* \xrightarrow{a.s.} \frac{1}{L} \text{Tr} \{ \mathbf{T}_{\psi,-k} \} = 1 - \frac{1}{L} \sum_{k' \in [Q] \setminus k} M_{\psi,k'} \quad (\text{C.6})$$

$$\frac{1}{L} (\mathbf{h}_{\psi,k}^{(\ell)})^T \mathbf{T}_{\psi,-k} (\mathbf{h}_{\psi,k}^{(\ell')})^* \xrightarrow{a.s.} 0, \text{ for } \ell' \neq \ell, \quad (\text{C.7})$$

where  $\mathbf{h}_{\psi,k}^{(\ell)}$  denotes the  $\ell$ -th column of  $\mathbf{H}_{\psi,k}$ . Therefore, we can derive that

$$\frac{1}{L} \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^* \xrightarrow{a.s.} \beta_{\psi,k} \left( 1 - \frac{1}{L} \sum_{k' \in [Q] \setminus k} M_{\psi,k'} \right) \mathbf{I}_{M_{\psi,k}}, \text{ as } L \rightarrow \infty, \quad (\text{C.8})$$

which reveals that all eigenvalues of  $\frac{1}{L} \mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^*$  become identical as  $L \rightarrow \infty$ . This in turn allows us to employ the following approximation

$$\lambda_{\psi,k,q} \approx \beta_{\psi,k} \left( L - \sum_{k' \in [Q] \setminus k} M_{\psi,k'} \right) \quad (\text{C.9})$$

which thus allows us to say that the effective rate in (4.11) can be approximated as

$$R_{\psi,k} \approx \xi_{G,Q} \sum_{q=1}^{J_{\psi,k}} \ln \left( 1 + \frac{P_{\psi,k,q}}{N_0} \beta_{\psi,k} \left( L - \sum_{k' \in [Q] \setminus k} M_{\psi,k'} \right) \right). \quad (\text{C.10})$$

According to the properties of the water-filling algorithm, we know that equal-power allocation among  $\{s_{\psi,k,q} : q \in [J_{\psi,k}]\}$  is optimal, which in turn tells us that

$$R_{\psi,k}^*(P_{\psi,k}) \approx \xi_{G,Q} J_{\psi,k} \ln \left( 1 + \frac{P_{\psi,k}}{N_0 J_{\psi,k}} \beta_{\psi,k} \left( L - \sum_{k' \in [Q] \setminus k} M_{\psi,k'} \right) \right). \quad (\text{C.11})$$

At this point we approximate the MMF optimization in (4.16) as

$$\mathcal{S}_5 : \begin{cases} \max_{\mathbf{P}_\Psi} \min_{\psi \in \Psi} \min_{k \in [Q]} \xi_{G,Q} J_{\psi,k} \ln \left( 1 + \frac{P_{\psi,k}}{N_0 J_{\psi,k}} \beta_{\psi,k} \left( L - \sum_{k' \in [Q] \setminus k} M_{\psi,k'} \right) \right) \\ \text{s. t. } P = \sum_{\psi \in \Psi} \sum_{k \in [Q]} P_{\psi,k} = P_{\text{tot}} \end{cases} \quad (\text{C.12})$$

Finally, as all users have the same effective rate under optimal power allocation for (C.12), we can use the total power constraint to derive (4.24) and (4.26) respectively.  $\square$

### C.3 Proof of Lemma 4.3

When considering the case of single-antenna users under BD precoding, the matrix form  $\mathbf{H}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{H}_{\psi,k}^*$  becomes the scalar  $\lambda_{\psi,k} \triangleq \mathbf{h}_{\psi,k}^T \mathbf{T}_{\psi,-k} \mathbf{h}_{\psi,k}^* \in \mathbb{C}$ . The effective average sum-rate under the BD precoding and the optimal MMF power allocation is

$$\bar{R}_{\text{BD-MRC}}^*(G, Q) = GQ \xi_{G,Q} \mathbb{E} \left\{ \ln \left( 1 + \frac{\rho}{\sum_{\psi \in \Psi} \sum_{k \in [Q]} \lambda_{\psi,k}^{-1}} \right) \right\}, \quad (\text{C.13})$$

where the expectation is over channel fading. By using the eigendecomposition of  $\mathbf{T}_{\psi,-k} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H$ , where  $\mathbf{\Lambda} \in \mathbb{C}^{(L-Q+1) \times (L-Q+1)}$  is an identity matrix and the columns of  $\mathbf{U} \in \mathbb{C}^{L \times (L-Q+1)}$  constitutes the associated eigenvectors, we can rewrite  $\lambda_{\psi,k}$  as  $\lambda_{\psi,k} = \mathbf{h}_{\psi,k}^T \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H \mathbf{h}_{\psi,k}^* = \|\mathbf{U}^H \mathbf{h}_{\psi,k}^*\|^2$ . As each column of  $\mathbf{U}$  has unit-norm, we have that  $\mathbf{U}^H \mathbf{h}_{\psi,k}^* \sim \mathcal{CN}(\mathbf{0}_{L-Q+1}, \beta_{\psi,k} \mathbf{I}_{L-Q+1})$ . We can conclude that  $\lambda_{\psi,k} = \|\mathbf{U}^H \mathbf{h}_{\psi,k}^*\|^2 \sim \text{Gamma}(L-Q+1, \beta_{\psi,k})$ . Define  $X \triangleq \sum_{\psi \in \Psi, k \in [Q]} \lambda_{\psi,k}$ . The characteristic function (CF) of  $X$  is

$$\begin{aligned} \text{CF}_X(t) &\triangleq \mathbb{E} \{ \exp(jtX) \} = \mathbb{E} \left\{ \exp \left( jt \sum_{\psi \in \Psi, k \in [Q]} \lambda_{\psi,k} \right) \right\} = \prod_{\psi \in \Psi, k \in [Q]} \mathbb{E} \{ \exp(jt \lambda_{\psi,k}) \} \\ &\stackrel{(a)}{=} \prod_{\psi \in \Psi, k \in [Q]} \text{CF}_{\lambda_{\psi,k}}(t) = \prod_{\psi \in \Psi, k \in [Q]} (1 - jt \beta_{\psi,k})^{-J}, \end{aligned} \quad (\text{C.14})$$

where (a) follows from the well-known CF of Gamma distribution. Referring to [158, Thm. 1], we can rewrite (C.14) in terms of Gamma functions, and then take the inverse CF transform, which is actually a kind of Mellin-Barnes contour integral. With the aid of

Meijer's G-function, we are able to obtain a closed-form expression for the PDF of  $X$ , as shown below

$$f_X(x) = \left[ \prod_{\psi \in \Psi, k \in [Q]} \left( \frac{J}{\beta_{\psi, k}} \right)^J \right] G_{\varpi, \varpi}^{\varpi, 0} \left( \exp(-x) \middle| \begin{matrix} \Delta_{\varpi}^{(1)} \\ \Delta_{\varpi}^{(2)} \end{matrix} \right) \quad (\text{C.15})$$

Substituting the PDF of  $X$  into (C.13), we obtain  $\bar{R}_{\text{BD}}^*(G, Q)$  in (4.29).

When  $\beta_{\psi, k} = \beta$  for any  $\psi \in \Psi$  and  $k \in [Q]$ , i.e., same path-loss in all users,  $X = \sum_{\psi \in \Psi, k \in [Q]} \lambda_{\psi, k} \sim \text{Gamma}(\varpi, \beta)$ . The effective sum-rate can be simplified as

$$\begin{aligned} \bar{R}_{\text{sum}}^*(G, Q) &= GQ\xi_{G, Q} \int_0^\infty \ln \left( 1 + \frac{\rho}{x} \right) f_X(x) dx \\ &\stackrel{(a)}{=} \frac{GQ\xi_{G, Q}}{(\varpi - 1)! \beta^\varpi} \int_0^\infty \ln \left( 1 + \frac{\rho}{x} \right) x^{\varpi-1} \exp \left( -\frac{x}{\beta} \right) dx, \end{aligned} \quad (\text{C.16})$$

where (a) follows from the PDF of  $\text{Gamma}(GQJ, \beta)$ . Considering that  $\exp(x) = G_{0,1}^{1,0}(-x | \bar{\phantom{x}})$  and  $\ln(1+x) = G_{2,2}^{1,2}(x | \begin{matrix} 1,1 \\ 1,0 \end{matrix})$ , we can rewrite (C.16) as

$$\begin{aligned} \bar{R}_{\text{sum}}^*(G, Q) &= \frac{GQ\xi_{G, Q}}{(\varpi - 1)! \beta} \int_0^\infty G_{2,2}^{1,2} \left( \frac{\rho}{x} \middle| \begin{matrix} 1,1 \\ 1,0 \end{matrix} \right) \frac{x^{\varpi-1}}{\beta^{\varpi-1}} G_{0,1}^{1,0} \left( \frac{x}{\beta} \middle| \bar{\phantom{x}} \right) dx \\ &\stackrel{(a)}{=} \frac{GQ\xi_{G, Q}}{(\varpi - 1)! \beta} \int_0^\infty G_{2,2}^{2,1} \left( \frac{x}{\rho} \middle| \begin{matrix} 0,1 \\ 0,0 \end{matrix} \right) G_{0,1}^{1,0} \left( \frac{x}{\beta} \middle| \bar{\phantom{x}} \right) dx, \end{aligned} \quad (\text{C.17})$$

where (a) follows from [76, Eq. (9.31.2)] and [76, Eq. (9.31.5)]. By using [76, Eq. (7.811.1)] in (C.17), we derive (4.30).

As  $\rho \rightarrow \infty$ , we can rewrite  $\bar{R}_{\text{sum}}^*(G, Q)$  in (C.13) as

$$\begin{aligned} \bar{R}_{\text{sum}}^*(G, Q) &= GQ\xi_{G, Q} \mathbb{E} \left\{ \ln \left( \frac{\rho}{X} \right) \right\} + o(1) \\ &= GQ\xi_{G, Q} \ln \rho - \frac{GQ\xi_{G, Q}}{(\varpi - 1)! \beta^\varpi} \int_0^\infty \ln(x) x^{\varpi-1} \exp \left( -\frac{x}{\beta} \right) dx + o(1), \end{aligned} \quad (\text{C.18})$$

which leads to (4.31) by using [76, Eq. (4.352.2)].  $\square$

## C.4 Proof of Proposition 4.1

Let  $\mathbf{e}_{\psi, k, q} \in \mathbb{C}^{\sum_{k \in [Q]} M_{\psi, k}}$  denote a vector with all zero elements except the  $(\sum_{k'' \in [k-1]} M_{\psi, k''} + q)$ -th element equalling 1. After removing the inter-group interference, the received signal for decoding  $s_{\psi, k, q}$  in (4.3) under ZF precoding designed in (4.32) is

$$\begin{aligned} y'_{\psi, k, q} &= \mathbf{e}_{\psi, k, q}^T \mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^* (\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*)^{-1} \mathbf{e}_{\psi, k, q} \sqrt{([\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}]_{k(q), k(q)}^{-1}} \sqrt{P_{\psi, k, q}} s_{\psi, k, q} + z'_{\psi, k, q} \\ &\quad + \sum_{q' \in [M_{\psi, k}] \setminus q} \underbrace{\mathbf{e}_{\psi, k, q}^T \mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^* (\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*)^{-1} \mathbf{e}_{\psi, k, q'}}_{=0} \sqrt{([\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}]_{k(q'), k(q')}^{-1}} \sqrt{P_{\psi, k, q'}} s_{\psi, k, q'} \end{aligned}$$

$$\begin{aligned}
 & + \sum_{k' \in [Q] \setminus k} \sum_{p \in [M_{\psi, k'}]} \underbrace{\mathbf{e}_{\psi, k, q}^T \mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^* (\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*)^{-1} \mathbf{e}_{\psi, k', p}}_{=\mathbf{0}} \sqrt{([\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}]_{k'(p), k'(p)}^{-1}} \sqrt{P_{\psi, k', p}} \mathbf{s}_{\psi, k', p} \\
 & = \sqrt{([\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}]_{k(q), k(q)}^{-1}} \sqrt{P_{\psi, k, q}} \mathbf{s}_{\psi, k, q} + z'_{\psi, k, q}, \tag{C.19}
 \end{aligned}$$

where  $\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^* (\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*)^{-1} = \mathbf{I}$  is considered, and the precoding vector for symbol  $s_{\psi, k, q}$  is  $\mathbf{H}_{\psi}^* (\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*)^{-1} \mathbf{e}_{\psi, k, q} \sqrt{([\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}]_{k(q), k(q)}^{-1}}$  for any  $\psi \in \Psi$  and  $k \in [Q]$  (cf.  $\mathbf{V}_{\Psi}$  in (4.32)). Then, the effective average rate at  $U_{\psi, k}$  is

$$\bar{R}_{\psi, k}^{\text{ZF}} = \xi_{G, Q} \mathbb{E} \left\{ \sum_{q=1}^{M_{\psi, k}} \ln \left( 1 + \frac{P_{\psi, k, q}}{N_0 [\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}_{k(q), k(q)}} \right) \right\} \tag{C.20}$$

$$\stackrel{(a)}{\geq} \xi_{G, Q} \sum_{q=1}^{M_{\psi, k}} \ln \left( 1 + \frac{P_{\psi, k, q}}{N_0} \left( \mathbb{E} \left\{ [\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}_{k(q), k(q)} \right\} \right)^{-1} \right), \tag{C.21}$$

where (a) follows from Jensen's inequality on the convex function  $\ln(1+x^{-1})$ . Considering that  $\mathbb{E} \left\{ [\mathbf{H}_{\psi}^T \mathbf{H}_{\psi}^*]^{-1}_{k(q), k(q)} \right\} = \frac{1}{\beta_{\psi, k}(L - M_{\psi})}$  (cf. [104]), we obtain the lower-bound in (4.34).

To obtain the upper-bound of  $\bar{R}_{\psi, k}^{\text{ZF}}$ , we use Jensen's inequality for the convex function  $\ln(1+x)$  in (C.20), which yields that

$$\bar{R}_{\psi, k}^{\text{ZF}} \leq \xi_{G, Q} \sum_{q=1}^{M_{\psi, k}} \ln \left( 1 + \frac{P_{\psi, k, q}}{N_0} \mathbb{E} \left\{ \frac{1}{[\mathbf{H}_{\psi} \mathbf{H}_{\psi}^H]^{-1}_{k(q), k(q)}} \right\} \right), \tag{C.22}$$

which induces the upper-bound in (4.35) by considering that  $\mathbb{E} \left\{ \left( [\mathbf{H}_{\psi} \mathbf{H}_{\psi}^H]^{-1}_{k(q), k(q)} \right)^{-1} \right\} = \beta_{\psi, k}(L - M_{\psi} + 1)$  (cf. [159]).  $\square$



## Appendix D

### Proofs in Chapter 5

#### D.1 Proof of Proposition 5.1

In view of the expression of  $\mathcal{S}$  in (5.2), it is easy to derive the average of  $|\mathcal{S}|^2$ , i.e., the channel power gain, as

$$\begin{aligned}\mathbb{E}\{|\mathcal{S}|^2\} &= \mathbb{E}\{\mathcal{A}^2 + \mathcal{A}\mathcal{W}\exp(j(\varsigma - \varsigma_0)) + \mathcal{A}\mathcal{W}\exp(j(\varsigma_0 - \varsigma)) + \mathcal{W}^2\} \\ &\stackrel{(a)}{=} \mathbb{E}\{\mathcal{A}^2\} + \mathbb{E}\{\mathcal{W}^2\} = 2b_0 + \aleph,\end{aligned}\tag{D.1}$$

where (a) follows from  $\mathbb{E}\{\exp(j\varsigma)\} = \mathbb{E}\{\exp(-j\varsigma)\} = 0$  because of  $\varsigma$  uniformly distributed over  $[0, 2\pi)$ . The average of  $|\mathcal{S}|^4$  is

$$\begin{aligned}\mathbb{E}\{|\mathcal{S}|^4\} &= \mathbb{E}\{|\mathcal{S}|^2|\mathcal{S}|^2\} \\ &= \mathbb{E}\left\{(\mathcal{A}^2 + \mathcal{A}\mathcal{V}\exp(j(\varsigma - \varsigma_0)) + \mathcal{A}\mathcal{V}\exp(j(\varsigma_0 - \varsigma)) + \mathcal{V}^2)\right. \\ &\quad \left.\times (\mathcal{A}^2 + \mathcal{A}\mathcal{V}\exp(j(\varsigma - \varsigma_0)) + \mathcal{A}\mathcal{V}\exp(j(\varsigma_0 - \varsigma)) + \mathcal{V}^2)\right\} \\ &\stackrel{(a)}{=} \mathbb{E}\{\mathcal{A}^4 + 4\mathcal{A}^2\mathcal{V}^2 + \mathcal{V}^4\},\end{aligned}\tag{D.2}$$

where (a) follows from  $\mathbb{E}\{\exp(j\varsigma)\} = \mathbb{E}\{\exp(-j\varsigma)\} = \mathbb{E}\{\exp(j2\varsigma)\} = \mathbb{E}\{\exp(-j2\varsigma)\} = 0$ . As  $\mathcal{A}^2 \sim \text{Exp}\left(\frac{1}{2b_0}\right)$  and  $\mathcal{V}^2 \sim \text{Gamma}\left(m_0, \frac{\aleph}{m_0}\right)$ , we have

$$\mathbb{E}\{\mathcal{A}^4\} = 8b_0^2, \quad \mathbb{E}\{\mathcal{V}^4\} = \left(1 + \frac{1}{m_0}\right)\aleph^2.$$

Therefore, the average of  $|\mathcal{S}|^4$  is derived as

$$\mathbb{E}\{|\mathcal{S}|^4\} = 8b_0^2 + 8b_0\aleph + \left(1 + \frac{1}{m_0}\right)\aleph^2\tag{D.3}$$

Combining (D.1) and (D.3), we can easily obtain the variance of  $|\mathcal{S}|^2$  as

$$\text{Var}\{|\mathcal{S}|^2\} = \mathbb{E}\{|\mathcal{S}|^4\} - \mathbb{E}^2\{|\mathcal{S}|^2\} = 4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0},\tag{D.4}$$

which concludes the proof.  $\square$

## D.2 Proof of Lemma 5.1

The CDF of  $\text{SNR}_{\text{MN}}$  in the MN scheme is

$$\begin{aligned} F_{\text{SNR}_{\text{MN}}}(x) &= \Pr \left\{ \min_{g \in \Psi} \{ \text{SNR}_g \} \leq x \right\} \\ &= 1 - \Pr \left\{ \min_{g \in \Psi} \{ \text{SNR}_g \} > x \right\} = 1 - \left[ \Pr \{ \text{SNR}_g > x \} \right]^G \end{aligned} \quad (\text{D.5})$$

By using the simplified CDF in Proposition 5.3, the CDF of  $\text{SNR}_{\text{MN}}$  can be re-written as

$$F_{\text{SNR}_{\text{MN}}}(x) = 1 - \alpha_0^G \exp \left( -\frac{G(\varphi_0 - \delta_0)}{\rho} x \right) \left[ \sum_{j=0}^{m_0-1} \left( \sum_{\ell=j}^{m_0-1} \frac{\Xi(\ell)\ell!}{j!} \frac{\rho^{-j}}{(\varphi_0 - \delta_0)^{\ell-j+1}} \right) x^j \right]^G. \quad (\text{D.6})$$

Using Multinomial theorem [150], we can further write the CDF of  $\text{SNR}_{\text{MN}}$  as

$$\begin{aligned} F_{\text{SNR}_{\text{MN}}}(x) &= 1 - \alpha_0^G \exp \left( -\frac{G(\varphi_0 - \delta_0)}{\rho} x \right) \sum_{\tilde{h}_1 + \dots + \tilde{h}_{m_0} = G} \binom{G}{\tilde{h}_1, \dots, \tilde{h}_{m_0}} \\ &\quad \times \left[ \prod_{t=0}^{m_0-1} \left( \sum_{\ell=t}^{m_0-1} \frac{\Xi(\ell)\ell!}{t!} \frac{\rho^{-t}}{(\varphi_0 - \delta_0)^{\ell-t+1}} \right)^{\tilde{h}_{t+1}} \right] x^{\sum_{t=0}^{m_0-1} t\tilde{h}_{t+1}}, \end{aligned} \quad (\text{D.7})$$

where  $\tilde{h}_1, \dots, \tilde{h}_m$  are non-negative integers.

The average rate in MN coded caching is

$$\begin{aligned} \bar{R}^{\text{MN}} &= G \mathbb{E} \{ \ln(1 + \text{SNR}_{\text{MN}}) \} \stackrel{(a)}{=} G \int_0^\infty \frac{1 - F_{\text{SNR}_{\text{MN}}}(x)}{1+x} dx \\ &= G \alpha_0^G \sum_{\tilde{h}_1 + \dots + \tilde{h}_{m_0} = G} \binom{G}{\tilde{h}_1, \dots, \tilde{h}_{m_0}} \left[ \prod_{t=0}^{m_0-1} \left( \sum_{\ell=t}^{m_0-1} \frac{\Xi(\ell)\ell!}{t!} \frac{\rho^{-t}}{(\varphi_0 - \delta_0)^{\ell-t+1}} \right)^{\tilde{h}_{t+1}} \right] \\ &\quad \times \int_0^\infty \frac{x^{\sum_{t=0}^{m_0-1} t\tilde{h}_{t+1}}}{1+x} \exp \left( -\frac{G(\varphi_0 - \delta_0)}{\rho} x \right) dx, \end{aligned} \quad (\text{D.8})$$

where (a) follows from [160, Eq. (48)]. By applying [76, Eq. (3.383.10)] in (D.8), we can easily obtain (5.10).  $\square$

## D.3 Proof of Lemma 5.2

The characteristic function (CF) of  $S_g = \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b})$  is defined as

$$\text{CF}_{S_g}(t) = \mathbb{E} \{ \exp(jtS_g) \} = \mathbb{E} \left\{ \exp \left( jt \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right) \right\} = \left[ \mathbb{E} \{ (1 + \text{SNR}_{g,b})^{jt} \} \right]^B. \quad (\text{D.9})$$

Substituting the PDF of  $\text{SNR}_{g,b}$  into (D.9) yields

$$\begin{aligned} \text{CF}_{S_g}(t) &= \left[ \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} \int_0^\infty (1+x)^{jt} x^i \exp\left(-\frac{\varphi_0 - \delta_0}{\rho} x\right) dx \right]^B \\ &= \left[ \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} i! \cdot \mathcal{U}\left(i+1, 2+i+jt, \frac{\varphi_0 - \delta_0}{\rho}\right) \right]^B, \end{aligned} \quad (\text{D.10})$$

By considering Gil-Pelaez Theorem, the CDF of  $S_g$  is obtained as

$$F_{S_g}(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \text{Im} \left\{ \frac{\exp(-jxt)}{t} \left[ \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} i! \cdot \mathcal{U}\left(i+1, 2+i+jt, \frac{\varphi_0 - \delta_0}{\rho}\right) \right]^B \right\} dt. \quad (\text{D.11})$$

Define  $S = \min_{g \in \Psi} \{S_g\} = \min_{g \in \Psi} \left\{ \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right\}$ . The CDF of  $S$  can be expressed by

$$\begin{aligned} F_S(x) &= \Pr \left\{ \min_{g \in \Psi} \{S_g\} \leq y \right\} = 1 - \Pr \left\{ \min_{g \in \Psi} \{S_g\} > y \right\} = 1 - (\Pr \{S_g > y\})^G \\ &= 1 - \left( \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \text{Im} \left\{ \frac{\exp(-jxt)}{t} \left[ \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} i! \cdot \mathcal{U}\left(i+1, 2+i+jt, \frac{\varphi_0 - \delta_0}{\rho}\right) \right]^B \right\} dt \right)^G. \end{aligned} \quad (\text{D.12})$$

The average rate is finally derived as

$$\bar{R}_{\text{ACC}} = \frac{G}{B \ln 2} \mathbb{E}\{S\} = \frac{G}{B \ln 2} \int_0^\infty [1 - F_S(x)] dx, \quad (\text{D.13})$$

which yields Lemma 5.2 by substituting (D.12), and which concludes the proof.  $\square$

## D.4 Proof of Lemma 5.3

It is easy to see that  $\varrho_l = \mathbb{E}\{\ln(1 + \text{SNR}_{g,b})\}$  is actually the average rate in TDM. In view of Corollary 5.1, the closed-form expression for  $\varrho_l$  is derived in (5.17). To derive the closed-form expression for  $\sigma_l^2$ , we first consider the second moment of  $\ln(1 + \text{SNR}_{g,b})$ ,

$$\begin{aligned} \mathbb{E}\{(\ln(1 + \text{SNR}_{g,b}))^2\} &= \int_0^\infty (\ln(1+x))^2 f_{\text{SNR}_{g,b}}(x) dx \\ &= \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{\rho^{i+1}} \int_0^\infty (\ln(1+x))^2 x^i \exp\left(-\frac{\varphi_0 - \delta_0}{\rho} x\right) dx. \end{aligned} \quad (\text{D.14})$$

To derive a closed-form expression for (D.14), we transfer  $\ln(1+x)$  and  $x^i \exp\left(-\frac{\varphi_0 - \delta_0}{\rho} x\right)$  into their Meijer's G-function forms,

$$\ln(1+x) = G_{2,2}^{1,2} \left( x \middle|_{1,0}^{1,1} \right), \quad (\text{D.15})$$

$$x^i \exp\left(-\frac{\varphi_0 - \delta_0}{\rho} x\right) = \left(\frac{\rho}{\varphi_0 - \delta_0}\right)^i G_{0,1}^{1,0}\left(\frac{\varphi_0 - \delta_0}{\rho} x \Big|_i^-\right) \quad (\text{D.16})$$

We can rewrite (D.14) as

$$\begin{aligned} & \mathbb{E}\{(\ln(1 + \text{SNR}_{g,b}))^2\} \\ &= \frac{\alpha_0}{\rho} \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{(\varphi_0 - \delta_0)^i} \int_0^\infty G_{2,2}^{1,2}\left(x \Big|_{1,0}^{1,1}\right) G_{2,2}^{1,2}\left(x \Big|_{1,0}^{1,1}\right) G_{0,1}^{1,0}\left(\frac{\varphi_0 - \delta_0}{\rho} x \Big|_i^-\right) dx \\ &\stackrel{(a)}{=} \alpha_0 \sum_{i=0}^{m_0-1} \frac{\Xi(i)}{(\varphi_0 - \delta_0)^{i+1}} G_{1,0:2,2:2,2}^{0,1:1,2:1,2}\left(\frac{\rho}{\varphi_0 - \delta_0}, \frac{\rho}{\varphi_0 - \delta_0} \Big|_{1,0}^{-i} \Big|_{1,0}^{1,1} \Big|_{1,0}^{1,1}\right), \end{aligned} \quad (\text{D.17})$$

where (a) follows from [153, Eq. (07.34.21.0081.01)]. Considering

$$\sigma_l^2 = \mathbb{E}\{(\ln(1 + \text{SNR}_{g,b}))^2\} - (\mathbb{E}\{\ln(1 + \text{SNR}_{g,b})\})^2, \quad (\text{D.18})$$

we can derive the closed-form expression for  $\sigma_l^2$  in (5.18), which concludes the proof.  $\square$

## D.5 Proof of Lemma 5.4

In [130], a real-function  $P(X)$  is defined by a positive random variable  $X$  with mean  $\mu_X$  and variance  $\sigma_X^2$ . The expectation of  $P(X)$  can be tightly approximated in the low- $\sigma_X^2$  region as

$$\mathbb{E}\{P(X)\} \approx P(\mu_X) + \frac{\sigma_X^2}{2} \left( \frac{\partial^2 P(X)}{\partial X^2} \Big|_{X=\mu_X} \right), \quad (\text{D.19})$$

where  $\frac{\partial^2 P(X)}{\partial X^2}$  represents the second derivative of  $P(X)$  with respect to  $X$ . We refer to [129] for the theoretical proof of this approximation method. By setting  $P(X) = \ln(1 + X)$  and  $P(X) = (\ln(1 + X))^2$ , where  $X = \text{SNR}_{g,b}$ , we can derive the tight approximations for the first and second moments of  $\ln(1 + \text{SNR}_{g,b})$  respectively,

$$\mathbb{E}\{\ln(1 + \text{SNR}_{g,b})\} \approx \ln(1 + \mathbb{E}\{\text{SNR}_{g,b}\}) - \frac{\text{Var}\{\text{SNR}_{g,b}\}}{2} \frac{1}{(1 + \mathbb{E}\{\text{SNR}_{g,b}\})^2}, \quad (\text{D.20})$$

$$\begin{aligned} \mathbb{E}\left\{\left(\ln(1 + \text{SNR}_{g,b})\right)^2\right\} &\approx \left(\ln(1 + \mathbb{E}\{\text{SNR}_{g,b}\})\right)^2 \\ &\quad + \text{Var}\{\text{SNR}_{g,b}\} \frac{1 - \ln(1 + \mathbb{E}\{\text{SNR}_{g,b}\})}{(1 + \mathbb{E}\{\text{SNR}_{g,b}\})^2}, \end{aligned} \quad (\text{D.21})$$

Considering the closed-form expression for the average and variance of  $\text{SNR}_{g,b}$  in Proposition 5.1, we can further have the closed-form expression for  $\varrho_l$  in (5.19) and the second moment of  $\ln(1 + \text{SNR}_{g,b})$  as

$$\mathbb{E}\left\{\left(\ln(1 + \text{SNR}_{g,b})\right)^2\right\} \approx \left(\ln(1 + (2b_0 + \aleph)\rho)\right)^2$$

$$+ \frac{1 - \ln(1 + (2b_0 + \aleph)\rho)}{(1 + (2b_0 + \aleph)\rho)^2} \left(4b_0^2 + 4b_0\aleph + \frac{\aleph^2}{m_0}\right) \rho^2. \quad (\text{D.22})$$

By combining (D.20) and (D.21), the variance of  $\ln(1 + \text{SNR}_{g,b})$  is derived as

$$\begin{aligned} \sigma_l^2 &= \text{Var} \{ \ln(1 + \text{SNR}_{g,b}) \} = \mathbb{E} \left\{ \left( \ln(1 + \text{SNR}_{g,b}) \right)^2 \right\} - \mathbb{E}^2 \{ \ln(1 + \text{SNR}_{g,b}) \} \\ &\approx \frac{\text{Var} \{ \text{SNR}_{g,b} \}}{(1 + \mathbb{E} \{ \text{SNR}_{g,b} \})^2} - \frac{(\text{Var} \{ \text{SNR}_{g,b} \})^2}{4(1 + \mathbb{E} \{ \text{SNR}_{g,b} \})^4}, \end{aligned} \quad (\text{D.23})$$

which is derived as (5.20) by using Proposition 5.1. We complete the proof.  $\square$

## D.6 Proof of Lemma 5.5

By using the PDF of MG distribution in (5.23), we can obtain the average of  $\ln(1 + \text{SNR}_{g,b})$  as

$$\varrho_l = \sum_{v=1}^V \alpha_v \int_0^\infty \ln(1+x) x^{\varphi_v-1} \exp(-\xi_v x) dx \stackrel{(a)}{=} \sum_{v=1}^V \alpha_v \xi_v^{-\varphi_v} \text{G}_{3,2}^{1,3} \left( \frac{1}{\xi_v} \middle|_{1,0}^{1-\varphi_v, 1, 1} \right), \quad (\text{D.24})$$

where (a) follows from [72, Eq. (24)].

The second moment of  $\ln(1 + \text{SNR}_{g,b})$  in the MG fading model is

$$\begin{aligned} \mathbb{E} \left\{ \left( \ln(1 + \text{SNR}_{g,b}) \right)^2 \right\} &= \sum_{v=1}^V \alpha_v \int_0^\infty (\ln(1+x))^2 x^{\varphi_v-1} \exp(-\xi_v x) dx \\ &\stackrel{(a)}{=} \sum_{v=1}^V \alpha_v \xi_v^{-(\varphi_v-1)} \int_0^\infty \text{G}_{2,2}^{1,2} \left( x \middle|_{1,0}^{1,1} \right) \text{G}_{2,2}^{1,2} \left( x \middle|_{1,0}^{1,1} \right) \text{G}_{0,1}^{1,0} \left( \xi_v x \middle|_{\varphi_v-1}^- \right) dx \\ &\stackrel{(b)}{=} \sum_{v=1}^V \alpha_v \xi_v^{-\varphi_v} \text{G}_{1,0:2,2:2,2}^{0,1:1,2:1,2} \left( \frac{1}{\xi_v}, \frac{1}{\xi_v} \middle|_{-}^{1-\varphi_v} \middle|_{1,0}^{1,1} \middle|_{1,0}^{1,1} \right), \end{aligned} \quad (\text{D.25})$$

where (a) follows by transferring the exponential function and the logarithm function into their Meijer's G-function forms, and (b) follows from [153, Eq. (07.34.21.0081.01)].

Finally, combining (D.24) and (D.25), we obtain the variance of  $\ln(1 + \text{SNR}_{g,b})$  in (5.28), which concludes the proof.  $\square$



# Bibliography

- [1] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
- [2] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, “Wireless caching: technical misconceptions and business barriers,” *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.
- [3] A. Tang, S. Roy, and X. Wang, “Coded caching for wireless backhaul networks with unequal link rates,” *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 1–13, Jan. 2018.
- [4] S. Saeedi Bidokhti, M. Wigger, and A. Yener, “Benefits of cache assignment on degraded broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 6999–7019, Nov. 2019.
- [5] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, and D. Gündüz, “Coded caching with asymmetric cache sizes and link qualities: The two-user case,” *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6112–6126, Sep. 2019.
- [6] E. Lampiris *et al.*, “Fundamental limits of wireless caching under uneven-capacity channels,” in *Int. Zurich Seminar*, Feb. 2020.
- [7] J. Zhang and P. Elia, “Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback,” *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May 2017.
- [8] E. Piovano, H. Joudeh, and B. Clerckx, “Generalized degrees of freedom of the symmetric cache-aided MISO broadcast channel with partial CSIT,” *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5799–5815, Sep. 2019.
- [9] E. Lampiris, A. Bazco-Nogueras, and P. Elia, “Resolving the feedback bottleneck of multi-antenna coded caching,” *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2331–2348, Apr. 2022.
- [10] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, “Cooperative caching and transmission design in cluster-centric small cell networks,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

- 
- [11] M. Bayat, R. K. Mungara, and G. Caire, "Achieving spatial scalability for coded caching via coded multipoint multicasting," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 227–240, Jan. 2019.
- [12] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.
- [13] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.
- [14] S. Zhong and X. Wang, "Joint multicast and unicast beamforming for coded caching," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3354–3367, Aug. 2018.
- [15] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.
- [16] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.
- [17] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2091–2106, Mar. 2020.
- [18] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5524–5537, Oct. 2016.
- [19] K. Wan, D. Tuninetti, and P. Piantanida, "Novel delivery schemes for decentralized coded caching in the finite file size regime," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2017, pp. 1–6.
- [20] S. Jin *et al.*, "A new order-optimal decentralized coded caching scheme with good performance in the finite file size regime," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5297–5310, Aug. 2019.
- [21] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5821–5833, Sep. 2017.
- [22] W. Song, K. Cai, and L. Shi, "Some new constructions of coded caching schemes with reduced subpacketization," 2019. [Online]. Available: <https://arxiv.org/abs/1908.06570>
- [23] N. Jindal and Z. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2006, pp. 1841–1845.

- [24] M. A. Maddah-Ali and U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.
- [25] C. Shangguan, Y. Zhang, and G. Ge, “Centralized coded caching schemes: A hypergraph theoretical approach,” *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5755–5766, Aug. 2018.
- [26] E. Parrinello, A. Ünsal, and P. Elia, “Fundamental limits of coded caching with multiple antennas, shared caches and uncoded prefetching,” *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2252–2268, Apr. 2020.
- [27] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, “Femto-caching: Wireless video content delivery through distributed caching helpers,” in *Proc. IEEE Int. Conf. on Comput. Commun. (INFOCOM)*, 2012, pp. 1107–1115.
- [28] M. Jia *et al.*, “Broadband hybrid satellite-terrestrial communication systems based on cognitive radio toward 5G,” *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 96–106, Dec. 2016.
- [29] C. Bockelmann *et al.*, “Massive machine-type communications in 5G: Physical and MAC-layer solutions,” *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
- [30] Teltonika-Networks. Mobile signal strength recommendations. [Online]. Available: [https://wiki.teltonika-networks.com/view/Mobile\\_Signal\\_Strength\\_Recommendations](https://wiki.teltonika-networks.com/view/Mobile_Signal_Strength_Recommendations)
- [31] K.-H. Ngo, S. Yang, and M. Kobayashi, “Scalable content delivery with coded caching in multi-antenna fading channels,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan. 2018.
- [32] A. M. Daniel and W. Yu, “Optimization of heterogeneous coded caching,” *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1893–1919, Mar. 2020.
- [33] B. Tegin and T. M. Duman, “Coded caching with user grouping over wireless channels,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 920–923, Jun. 2020.
- [34] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, “Multi-server coded caching,” *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.
- [35] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “Fundamental limits of cache-aided interference management,” *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.
- [36] I. Bergel and S. Mohajer, “Cache-aided communications with multiple antennas at finite SNR,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1682–1691, Aug. 2018.

- 
- [37] F. Sofrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.
- [38] R. Méndez-Rial, C. Rusu, N. González-Prelcic, and A. Alkhateeb, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.
- [39] X. Xu and M. Tao, "Modeling, analysis, and optimization of caching in multi-antenna small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5454–5469, Nov. 2019.
- [40] M. J. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tölli, "Low-complexity high-performance cyclic caching for large MISO systems," *IEEE Trans. Wireless Commun.*, to be published, doi: 10.1109/TWC.2021.3119772.
- [41] S. Mohajer and I. Bergel, "MISO Cache-Aided Communication with Reduced Subpacketization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020.
- [42] N. S. Karat, S. Dey, A. Thomas, and B. S. Rajan, "An optimal linear error correcting delivery scheme for coded caching with shared caches," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 1217–1221.
- [43] M. Salehi, A. Tolli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization–rate trade-off in multi-antenna coded caching," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.
- [44] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-caching wireless networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827–2839, Apr. 2018.
- [45] M. Salehi and A. Tölli, "Diagonal multi-antenna coded caching for reduced subpacketization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020.
- [46] M. Salehi, A. Tölli, and S. P. Shariatpanahi, "Subpacketization - beamformer interaction in multi-antenna coded caching," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.
- [47] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7635–7651, Dec. 2020.
- [48] "The industry's first independent benchmark study of 5G NR MU-MIMO," Signals Research Group, Tech. Rep., Sept. 2020.
- [49] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1469–1482, Apr. 2006.
- [50] M. K. Simon and M.-S. Alouini, *Digital communication over fading channels*. John Wiley & Sons, 2005.

- [51] H. Zhao, A. Bazco-Nogueras and P. Elia, “Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5450–5466, Jul. 2022.
- [52] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Resolving the worst-user bottleneck of coded caching: Exploiting finite file sizes,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 1–5.
- [53] —, “Wireless coded caching with shared caches can overcome the near-far bottleneck,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 350–355.
- [54] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Coded caching gains at low SNR over Nakagami fading channels,” in *Proc. Asilomar Conf. Signals, Syst., and Comput. (ACSSC)*, Nov. 2021.
- [55] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up MIMO: Opportunities and challenges with very large arrays,” *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
- [56] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An overview of massive MIMO: Benefits and challenges,” *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
- [57] E. Björnson, J. Hoydis, and L. Sanguinetti, “Massive MIMO has unlimited capacity,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.
- [58] N. Rajatheva *et al.*, “White paper on broadband connectivity in 6G,” 2020. [Online]. Available: [arxiv.org/abs/2004.14247](https://arxiv.org/abs/2004.14247)
- [59] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [60] H. Q. Ngo and E. G. Larsson, “No downlink pilots are needed in TDD massive mimo,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2921–2935, May 2017.
- [61] 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.
- [62] J. Hoydis, S. Brink, and M. Debbah, “Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
- [63] Y.-G. Lim, C.-B. Chae, and G. Caire, “Performance analysis of massive MIMO for cell-boundary users,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6827–6842, Dec. 2015.

- 
- [64] C. Feng, Y. Jing, and S. Jin, “Interference and outage probability analysis for massive MIMO downlink with MF precoding,” *IEEE Signal Process. Lett.*, vol. 23, no. 3, pp. 366–370, Mar. 2016.
- [65] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, “Power scaling of uplink massive MIMO systems with arbitrary-rank channel means,” *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 966–981, Oct. 2014.
- [66] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Vector coded caching multiplicatively increases the throughput of realistic downlink systems,” *IEEE Trans. Wireless Commun.*, 2022, accepted for publication, doi: 10.1109/TWC.2022.3213475.
- [67] H. Zhao, A. Bazco-Nogueras and P. Elia, “Vector coded caching greatly enhances massive MIMO,” in *Proc. 23rd IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2022, pp. 1–5.
- [68] H. Zhao, E. Lampiris, G. Caire, and P. Elia, “Multi-antenna coded caching analysis in finite SNR and finite subpacketization,” in *Proc. 25th Int. ITG Workshop on Smart Antennas (WSA)*, Nov. 2021, pp. 433–438.
- [69] H. Zhao and P. Elia, “Vector coded caching substantially boosts MU-MIMO: Pathloss, CSI and power-allocation considerations,” in *Proc. 26th Int. ITG Workshop on Smart Antennas (WSA) and 13th Conf on Systems, Commun., and Coding (SCC)*, Feb 2023.
- [70] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Coded caching in land mobile satellite systems,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 4571–4576.
- [71] A. Abdi, W. Lau, M.-S. Alouini, and M. Kaveh, “A new simple model for land mobile satellite channels: first- and second-order statistics,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 519–528, May 2003.
- [72] S. Atapattu, C. Tellambura, and H. Jiang, “A mixture Gamma distribution to model the SNR of wireless channels,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4193–4203, Dec. 2011.
- [73] G. Kramer and S. Shamai, “Capacity for classes of Broadcast Channels with receiver side information,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2007, pp. 313–318.
- [74] B. Asadi, L. Ong, and S. J. Johnson, “Optimal coding schemes for the three-receiver AWGN BC with receiver message side information,” *IEEE Trans. Inf. Theory*, vol. 61, no. 10, pp. 5490–5503, Oct. 2015.
- [75] F. Xue and S. Sandhu, “PHY-layer network coding for broadcast channel with side information,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2007, pp. 108–113.
- [76] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. Academic press, 2007.

- [77] S. R. Finch, *Mathematical constants*. Cambridge university press, 2003.
- [78] G. Kamath, “Bounds on the expectation of the maximum of samples from a Gaussian,” 2015. [Online]. Available: <http://www.gautamkamath.com/writings/gaussianmax.pdf>
- [79] S. Venkateshan and P. Swaminathan, *Computational Methods in Engineering*. Academic Press, 2014.
- [80] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, “Exploiting caching and multicast for 5G wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.
- [81] M. Z. Shafiq *et al.*, “A first look at cellular network performance during crowded events,” *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 1, pp. 17–28, 2013.
- [82] F. J. Girón and C. del Castell, “A note on the convolution of inverted-gamma distributions with applications to the behrens-fisher distribution,” *RACSAM*, vol. 95, no. 1, pp. 39–44, 2001.
- [83] S. Venkateshan and P. Swaminathan, *Computational Methods in Engineering*. Academic Press, 2014.
- [84] “Study on channel model for frequencies from 0.5 to 100 ghz,” 3GPP, Tech. Rep. 38.901, version 16.1.0, Release 16, Dec. 2019, accessed on: 22/12/2020. [Online]. Available: [https://www.etsi.org/deliver/etsi\\_tr/138900\\_138999/138901/16.01.00\\_60/tr\\_138901v160100p.pdf](https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/16.01.00_60/tr_138901v160100p.pdf)
- [85] C. Zhang, J. Ye, G. Pan, and Z. Ding, “Cooperative hybrid VLC-RF systems with spatially random terminals,” *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6396–6408, Dec. 2018.
- [86] A. Lozano, A. Tulino, and S. Verdú, “High-SNR power offset in multiantenna communication,” *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4134–4151, Dec. 2005.
- [87] A. Bazco-Nogueras, P. de Kerret, D. Gesbert, and N. Gresset, “Asymptotically achieving centralized rate on the  $M \times K$  decentralized network MISO,” *IEEE Trans. Inf. Theory*, vol. 68, no. 1, pp. 248–271, Jan. 2022.
- [88] “5G Implementation Guidelines,” GSMA, Tech. Rep. version 2.0, Jul. 2019, accessed on: 20/01/2021. [Online]. Available: <https://www.gsma.com/futurenetworks/wp-content/uploads/2019/03/5G-Implementation-Guideline-v2.0-July-2019.pdf>
- [89] P. Popovski *et al.*, “Scenarios, requirements and KPIs for 5G mobile and wireless system,” Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS), Tech. Rep. ICT-317669-METIS/D1.1, Apr. 2013,

- accessed on: 01/12/2020. [Online]. Available: <https://cordis.europa.eu/docs/projects/cnect/9/317669/080/deliverables/001-METISD11v1pdf.pdf>
- [90] P. von Butovitsch, D. Astely, C. Friberg, A. Furuskar, B. Goransson, B. Hogan, J. Karlsson, and E. Larsson, “Advanced antenna systems for 5G networks,” Ericsson, Tech. Rep. GFMC-18:000530, Nov. 2018, accessed on: 22/01/2021. [Online]. Available: [https://www.ericsson.com/4a8a87/assets/local/reports-papers/white-papers/10201407\\_wp\\_advanced\\_antenna\\_system\\_nov18\\_181115.pdf](https://www.ericsson.com/4a8a87/assets/local/reports-papers/white-papers/10201407_wp_advanced_antenna_system_nov18_181115.pdf)
- [91] Z. Wang, Q. Liu, M. Li, and W. Kellerer, “Energy efficient analog beamformer design for mmWave multicast transmission,” *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 2, pp. 552–564, Jun. 2019.
- [92] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, “Transmit beamforming for physical-layer multicasting,” *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.
- [93] L. Liang, W. Xu, and X. Dong, “Low-complexity hybrid precoding in massive multiuser MIMO systems,” *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.
- [94] S. Tang, H. Yomo, T. Ueda, R. Miura, and S. Obana, “Full rate network coding via nesting modulation constellations,” *EURASIP J. Wireless Commun. Netw.*, p. 780632, 2011.
- [95] (2010, Mar.) Further advancements for E-UTRA physical layer aspects (release 9), tech. rep. 36.814. 3GPP, Sophia-Antipolis, France. [Online]. Available: <http://www.qtc.jp/3GPP/Specs/36814-900.pdf>
- [96] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, “On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users,” *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
- [97] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., “Spatially sparse precoding in millimeter wave MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
- [98] E. Lampaert and P. Elia, “Achieving full multiplexing and unbounded caching gains with bounded feedback resources,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2018, pp. 1440–1444.
- [99] A. Malik, B. Serbetci, E. Parrinello, and P. Elia, “Fundamental limits of stochastic shared-cache networks,” *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4433–4447, Jul. 2021.
- [100] C. B. Peel *et al.*, “A vector-perturbation technique for near-capacity multi-antenna multiuser communication-part I: Channel inversion and regularization,” *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

- [101] M. Vu and A. Paulraj, “MIMO wireless linear precoding,” *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 86–105, Sep. 2007.
- [102] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, “Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback,” *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, Jul. 2012.
- [103] M. Matthaiou, M. R. McKay, P. J. Smith, and J. A. Nossek, “On the condition number distribution of complex wishart matrices,” *IEEE Trans. Commun.*, vol. 58, no. 6, pp. 1705–1717, Jun. 2010.
- [104] A. M. Tulino and S. Verdú, “Random matrix theory and wireless communications,” *Foundations Trends Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, Jun. 2004.
- [105] E. Lampiris, A. Bazco-Nogueras, and P. Elia, “Resolving the feedback bottleneck of multi-antenna coded caching,” *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2331–2348, Apr. 2022.
- [106] M. Kobayashi and G. Caire, “On the net DoF comparison between ZF and MAT over time-varying MISO broadcast channels,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2012, pp. 2286–2290.
- [107] M. Sadeghi, E. Björnson, E. G. Larsson, C. Yuen, and T. L. Marzetta, “Max–min fair transmit precoding for multi-group multicasting in massive MIMO,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1358–1373, Feb. 2018.
- [108] M. Kobayashi, G. Caire, and N. Jindal, “How much training and feedback are needed in MIMO broadcast channels?” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2008, pp. 2663–2667.
- [109] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, “Multiuser MIMO achievable rates with downlink training and channel state feedback,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.
- [110] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Vector coded caching multiplicatively boosts the throughput of realistic downlink systems,” 2022. [Online]. Available: <https://arxiv.org/abs/2202.07047>
- [111] S. Gupta and S. Moharir, “Request patterns and caching for VoD services with recommendation systems,” in *Proc. Int. Conf. on Commun. Syst. and Netw. (COM-SNETS)*, Jan. 2017, pp. 31–38.
- [112] Q. Spencer, A. Swindlehurst, and M. Haardt, “Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels,” *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.
- [113] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

- 
- [114] G. Caire and S. Shamai, “On the achievable throughput of a multi-antenna Gaussian broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 43, no. 7, pp. 1691–1706, Jul. 2003.
- [115] E. Björnson, M. Kountouris, and M. Debbah, “Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination,” in *Proc. Int. Conf. Telecommun. (ICT)*, May 2013, pp. 1–5.
- [116] P. Wang *et al.*, “Convergence of satellite and terrestrial networks: A comprehensive survey,” *IEEE Access*, vol. 8, pp. 5550–5588, 2020.
- [117] H. Wu, J. Li, H. Lu, and P. Hong, “A two-layer caching model for content delivery services in satellite-terrestrial networks,” in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.
- [118] A. Kalantari *et al.*, “Cache-assisted hybrid satellite-terrestrial backhauling for 5G cellular networks,” in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.
- [119] T. X. Vu *et al.*, “Efficient 5G edge caching over satellite,” in *Proc. 36th Int. Commun. Satellite Syst. Conf. (ICSSC)*, Oct. 2018, pp. 1–5.
- [120] K. An, Y. Li, X. Yan, and T. Liang, “On the performance of cache-enabled hybrid satellite-terrestrial relay networks,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1506–1509, Oct. 2019.
- [121] X. Zhang *et al.*, “On the performance of hybrid satellite-terrestrial content delivery networks with non-orthogonal multiple access,” *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 454–458, Mar. 2021.
- [122] X. Wang, H. Li, T. Lan, and Q. Wu, “Overlay coded multicast for edge caching in 5G-satellite integrated networks,” in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, May 2020, pp. 1–7.
- [123] “Space communication calculations,” Australian Space Academy, 2019, accessed on: 21/10/2021. [Online]. Available: <http://www.spaceacademy.net.au/spacelink/spcomcalc.htm>
- [124] G. Pan, J. Ye, Y. Tian, and M.-S. Alouini, “On harq schemes in satelliteterrestrial transmissions,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, p. 7998–8010, Dec. 2020.
- [125] P. K. Sharma *et al.*, “Performance analysis of overlay spectrum sharing in hybrid satellite-terrestrial systems with secondary network selection,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, p. 6586–6601, Oct. 2017.
- [126] A. Erdelyi, “Transformation of a certain series of products of confluent hypergeometric functions. applications to laguerre and charlier polynomials,” *Compos. Math.*, vol. 7, pp. 340–352, 1940.

- [127] S. Ohmori, H. Wakana, and S. Kawase, *Mobile Satellite Communications*. Artech House, 1997.
- [128] M. Shah, “On generalization of some results and their applications,” *Collectanea Mathematica*, vol. 24, no. 3, pp. 249–266, 1973.
- [129] J. Holtzman, “A simple, accurate method to calculate spread-spectrum multiple-access error probabilities,” *IEEE Trans. Commun.*, vol. 40, no. 3, pp. 461–464, Mar. 1992.
- [130] H. Zhao, Y. Liu, A. Sultan-Salem, and M.-S. Alouini, “A simple evaluation for the secrecy outage probability over generalized- $K$  fading channels,” *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1479–1483, Sep. 2019.
- [131] H. Zhao and M.-S. Alouini, “On the transmission probabilities in quantum key distribution systems over FSO links,” *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 429–442, Jan. 2021.
- [132] J. Zhang, W. Zeng, X. Li, Q. Sun, and K. P. Peppas, “New results on the fluctuating two-ray model with arbitrary fading parameters and its applications,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2766–2770, Mar. 2018.
- [133] H. Zhao, Z. Liu, and M.-S. Alouini, “Different power adaption methods on fluctuating two-ray fading channels,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 592–595, Apr. 2019.
- [134] H. Zhao, J. Zhang, L. Yang, G. Pan, and M.-S. Alouini, “Secure mmWave communications in cognitive radio networks,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1171–1174, Aug. 2019.
- [135] J. M. Romero-Jerez, F. J. Lopez-Martinez, J. F. Paris, and A. J. Goldsmith, “The fluctuating two-ray fading model: Statistical characterization and performance analysis,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4420–4432, Jul. 2017.
- [136] M. D. Yacoub, “The  $\kappa - \mu$  distribution and the  $\eta - \mu$  distribution,” *IEEE Antennas Propagat. Mag.*, vol. 49, no. 1, pp. 68–81, Feb. 2007.
- [137] H. Chergui, M. Benjillali, and S. Saoudi, “Performance analysis of project-and-forward relaying in mixed MIMO-Pinhole and Rayleigh dual-hop channel,” *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 610–613, Mar. 2016.
- [138] I. S. Ansari, S. Al-Ahmadi, F. Yilmaz, M.-S. Alouini, and H. Yanikomeroglu, “A new formula for the BER of binary modulations with dual-branch selection over generalized- $K$  composite fading channels,” *IEEE Trans. Commun.*, vol. 59, no. 10, pp. 2654–2658, Oct. 2011.
- [139] E. Lampiris, J. Zhang, O. Simeone, and P. Elia, “Fundamental limits of wireless caching under uneven-capacity channels,” in *Proc. Int. Zurich Seminar on Inf. and Commun. (IZS)*, Feb. 2020, pp. 120–124.

- 
- [140] H. Joudeh, E. Lampaeri, P. Elia, and G. Caire, “Fundamental limits of wireless caching under mixed cacheable and uncacheable traffic,” *IEEE Trans. Inf. Theory*, vol. 67, no. 7, pp. 4747–4767, Jul. 2021.
- [141] E. Lampaeri and P. Elia, “Bridging two extremes: Multi-antenna coded caching with reduced subpacketization and CSIT,” in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019.
- [142] A. Malik, B. Serbetci, and P. Elia, “Coded caching in networks with heterogeneous user activity,” Jan. 2022. [Online]. Available: <https://arxiv.org/abs/2103.09156>
- [143] “Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022,” White Paper, Cisco, San Jose, CA, USA, Feb. 2019.
- [144] J. W. Yoo, T. Liu, and F. Xue, “Gaussian broadcast channels with receiver message side information,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2009, pp. 2472–2476.
- [145] T. Cover and A. Thomas, *Elements of information theory*. Wiley-Interscience, Jul. 1991.
- [146] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*, 4th ed. Tata McGraw-Hill Ed., 2001.
- [147] J. Gil-Pelaez, “Note on the inversion theorem,” *Biometrika*, vol. 38, no. 3-4, pp. 481–482, 1951.
- [148] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1970, vol. 55.
- [149] P. Billingsley, *Probability and Measure*, ser. Wiley Series in Probability and Statistics. Wiley, 1995.
- [150] K. Kataria, “A probabilistic proof of the multinomial theorem,” *Amer. Math. Monthly*, vol. 123, no. 1, pp. 94–96, Jan. 2016.
- [151] M. Inlow, “A moment generating function proof of the Lindeberg-Lévy central limit theorem,” *Amer. Statist.*, vol. 64, no. 3, pp. 228–230, Aug. 2010.
- [152] J. Pickands III, “Moment convergence of sample extremes,” *Ann. of Math. Statist.*, vol. 39, no. 3, pp. 881–889, Jun. 1968.
- [153] Wolfram Functions. [Online]. Available: <http://functions.wolfram.com/07.34.21.0081.01>
- [154] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.
- [155] D. V. Widder, “The Stieltjes transform,” *Trans. American Mathematical Society*, vol. 43, no. 1, pp. 7–60, 1938.

- [156] M. A. Woodbury, "Inverting modified matrices," *Statistical Research Group Memo. Reports, Princeton Univ. (42)*, 1950.
- [157] A. W. van der Vaart, *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 2000.
- [158] I. S. Ansari, F. Yilmaz, M.-S. Alouini, and O. Kucur, "New results on the sum of Gamma random variates with application to the performance of wireless communication systems over Nakagami- $m$  fading channels," 2012. [Online]. Available: <https://arxiv.org/abs/1202.2576>
- [159] K. K. Wong and Z. Pan, "Array gain and diversity order of multiuser MISO antenna systems," *Int. J. Wireless Inf. Netw.*, vol. 15, no. 2, pp. 82–89, May 2008.
- [160] H. Zhao, Z. Liu, L. Yang, and M.-S. Alouini, "Secrecy analysis in DF relay over generalized- $K$  fading channels," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7168–7182, Oct. 2019.