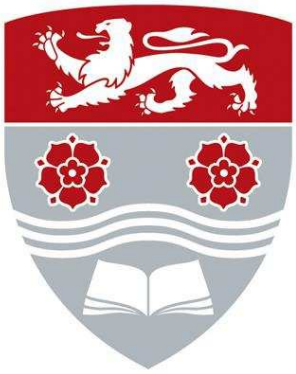
Statistical Delay QoS Driven Resource Allocation and Performance Analysis for Wireless Communication Networks



Wenjuan Yu
School of Computing and Communications
Lancaster University

A thesis submitted for the degree of
Doctor of Philosophy

October 2017

To my family

Abstract

Delay quality-of-service (QoS) guarantees play a critical role in enabling delay-sensitive wireless applications. By applying the theory of effective capacity (EC), the maximum arrival rate with a guaranteed delay-outage probability constraint, is analyzed and investigated in terms of delay-constrained resource allocation and link-layer throughput analysis.

Firstly, a joint optimization problem of link-layer energy efficiency (EE) and EC in a single-user single-carrier communication system, is proposed and investigated, under a delay violation probability requirement and an average transmit power constraint. Formulated as a normalized multi-objective optimization problem (MOP), the problem is transformed into a weighted single-objective optimization problem (SOP), and then solved. The proposed optimal power value is proved to be sufficient for the Pareto optimal set of the original EE-EC MOP.

Secondly, a total EC maximization problem subject to the individual link-layer EE requirement as well as the per-user average transmit power limit, in a multi-user multi-carrier orthogonal frequency-division multiple access (OFDMA) system, is proposed and analyzed. Formulated as a combinatorial integer programming problem, the problem is decoupled into a frequency provisioning problem and an independent per-user multi-carrier EE-EC tradeoff problem. A low-complexity heuristic algorithm is proposed to obtain the subcarrier assignment solution coupled with a per-user optimal power allocation strategy, across frequency and time domains.

Finally, the achievable link-layer rate under the per-user delay QoS requirements is studied for a downlink M-user non-orthogonal multiple access (NOMA) network. The impact of the transmit signal-to-noise ratio (SNR) and the delay QoS requirement on the per-user achievable EC and the total link-layer rate is investigated and compared between NOMA and orthogonal multiple access (OMA) networks. All theoretical conclusions and closed-form expressions are confirmed with Monte Carlo results.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors, Prof. Qiang Ni and Dr. Leila Musavian, for their patient guidance and continuous support, during the past four years of my Ph.D. study. Their commitment to solid and high-quality research work impressed me most and will keep me motivated in the future.

My sincere appreciation also goes to all academic and administrative staff in the School of Computing and Communications, for their professional advice, valuable support, and interesting seminars held regularly. It has been my pleasure to work with everyone of you in such a warm group.

Last but not the least, I would like to thank my parents and my sister for their love and encouragement throughout my entire life. I also want to thank all my lovely friends and Mr. Qiao Cheng, for always being so supportive and encouraging.

List of Publications

Journal papers

1. **W. Yu**, L. Musavian and Q. Ni, "Tradeoff Analysis and Joint Optimization of Link-Layer Energy Efficiency and Effective Capacity Toward Green Communications", IEEE Trans. Wireless Commun., vol. 15, no. 5, pp. 3339-3353, Jan. 2016.
2. **W. Yu**, L. Musavian and Q. Ni, "Statistical Delay QoS Driven Energy Efficiency and Effective Capacity Tradeoff for Uplink Multi-User Multi-Carrier Systems", IEEE Trans. Commun., vol. 65, no. 8, pp. 3494-3508, Apr. 2017.
3. **W. Yu**, L. Musavian and Q. Ni, "Link-Layer Capacity of NOMA Under Statistical Delay QoS Guarantees", submitted to IEEE Trans. Commun.

Conference papers

1. **W. Yu**, L. Musavian and Q. Ni, "Weighted Tradeoff between Effective Capacity and Energy Efficiency", in IEEE Int. Conf. Commun. (ICC), London, UK, Jun. 2015, pp. 238-243.
2. **W. Yu**, L. Musavian and Q. Ni, "Multi-carrier Link-Layer Energy Efficiency and Effective Capacity Tradeoff", in IEEE Int. Conf. Commun. (ICC) Workshop, London, UK, Jun. 2015, pp. 2763-2768.
3. **W. Yu**, L. Musavian and Q. Ni, "Fundamental Capacity Limits of NOMA Under Statistical Delay QoS Guarantees", submitted to Globecom 2018.

Contents

List of Tables

List of Figures

List of Acronyms

3GPP	Third generation partnership project
5G	Fifth generation
AMC	Adaptive modulation and coding
ATM	Asynchronous transfer mode
AWGN	Additive white gaussian noise
BS	Base station
CB	Contention-based
CSI	Channel-sate information
EB	Effective bandwidth
EC	Effective capacity
EE	Energy efficiency
FDMA	Frequency-division multiple access
FIFO	First-In-First-Out
GDP	Gross domestic product
GHG	Green house gases
ICT	Information and communication technology
i.i.d.	independent and identically distributed
IoT	Internet of things
KKT	Karush-Kuhn-Tucker
LHS	Left-hand-side
LTE-A	Long term evolution advanced

MA	Multiple access
MB	Mutually beneficial
MEP	Minimum envelope process
MER	Minimum envelope rate
MGF	Moment generating function
MOP	Multi-objective optimization problem
MUST	Multiuser superposition transmission
NOMA	Non-orthogonal multiple access
OFDM	Orthogonal frequency division multiplexing
OFDMA	Orthogonal frequency-division multiple access
OMA	Orthogonal multiple access
PMF	Probability mass function
PDF	Probability density function
QoS	Quality-of-service
RHS	Right-hand-side
SE	Spectral efficiency
SIC	Successive interference cancellation
SISO	Single-input single-output
SNR	Signal-to-noise ratio
SOP	Single-objective optimization problem
SWIPT	Simultaneous wireless information and power transfer
TDMA	Time division multiple access
TWh	TeraWatt hour

List of Symbols and Mathematical Operators

Symbols

$X_1, X_2, \ldots$	a sequence of i.i.d. random variables
μ	Mean of random variable X_1
M_n	Empirical mean of random variables $X_1, X_2, \ldots$
θ	Delay QoS exponent
$M_X(\theta)$	Moment generating function of the random variable X
$I(x)$	Legendre transform
$a(t)$	The number of arrivals at time t
$A(t)$	Cumulative number of arrivals in the time interval $(0, t]$
$\hat{A}(t)$	Deterministic envelope process of $A(t)$
$A^*(t)$	Minimum deterministic envelope process of $A(t)$
a^*	Minimum envelope rate
$\hat{A}(\theta, t)$	θ-envelope process of $A(t)$
$A^*(\theta, t)$	θ-minimum envelope process of $A(t)$
$a^*(\theta)$	θ-minimum envelope rate
$q(t)$	The number of packets in the queue at time t
$c(t)$	The capacity of the link at time t
$E_{\mathrm{b}}(\theta)$	Effective bandwidth
$E_{\mathrm{c}}(\theta)$	Effective capacity
$D_{\max}$	Maximum delay bound

T_{f}	The length of each fading-block
B	Channel bandwidth
ϵ	Power amplifier efficiency
$P_{\mathcal{L}}$	Distance-based path-loss
$P_{\max}$	Average input power limit
P_{t}	Instantaneous transmission power
P_{r}	Scaled instantaneous transmission power
$\overline{P}_{\mathrm{t}}$	Average transmission power
$\overline{P}_{\mathrm{r}}$	Scaled average transmission power
Ψ_{EE}	Normalization value for energy efficiency function
Ψ_{SE}	Normalization value for spectral efficiency function
P_{norm}	Normalization factor
w_1	Importance weight of the energy efficiency function
P_{c}	Circuit power of the transmitter
$P_{\mathrm{c_r}}$	Scaled circuit power of the transmitter
K	The number of users in a multi-user multi-carrier OFDMA system
N	The number of subcarriers in a multi-user multi-carrier OFDMA system
$\boldsymbol{\theta}$	Delay QoS exponent vector
$\boldsymbol{\chi}_{\mathrm{EE}}$	EE requirement factor vector
$\boldsymbol{\phi}$	Subcarrie assignment indicator matrix
N_k	The number of allocated subcarriers for the k^{th} user
$\mathcal{N}_k$	Index set of subcarriers allocated to the k^{th} user
B_k	Bandwidth allocated to the k^{th} user
$\boldsymbol{P}_k$	Subcarrier power allocation vector for the k^{th} user
$\boldsymbol{P}_k^{\mathrm{r}}$	Scaled subcarrier power allocation vector for the k^{th} user
$\overline{P_k}$	Average transmission power for the k^{th} user

$P_{\max}^{k}$	Average input power limit for the k^{th} user
P_{c}^{k}	Circuit power for the k^{th} user
$\boldsymbol{\eta}_{\max}$	Maximum achievable link-layer EE matrix
$\boldsymbol{\eta}_{\text{req}}$	EE requirement vector
$\boldsymbol{S}_{\text{req}}$	Subcarrier requirement vector
M	The number of users in a downlink NOMA network
α_k	Power coefficient for the k^{th} user
ρ	Transmit SNR
$f_{(m)}\left(\gamma_m\right)$	PDF of the ordered channel power gains γ_m

Mathematical Operators

$P(A)$	Probability of an event A
$\exp, e$	Exponential function
$\mathbb{E}[\cdot]$	Expectation operation
$\ln(\cdot)$	Natural logarithm
$\log_i$	Logarithm base i
$\inf\limits_{\theta} f\left(\theta\right)$	Infimum of $f(\theta)$
$\sup\limits_{\theta} f(\theta)$	Supremum of $f(\theta)$
$\lim$	Limitation
$\limsup$	Limit superior
$\liminf$	Limit inferior
$\sum$	Summation operation
$\prod$	Product operation
$\int$	Integration of sets
max	Maximum function
min	Minimum function
$\in$	Set membership (Belongs to a set)

$\forall$	Universal quantifier (For all)
$\exists$	Existential quantifier (There exists)
$\cap$	Intersection
$\frac{\mathrm{d}f(x)}{\mathrm{d}x}, f'(x)$	First derivative of function f with respect to x
$\frac{\mathrm{d}^2f(x)}{\mathrm{d}x^2}, f''(x)$	Second derivative of function f with respect to x
$\frac{\partial f(x,y)}{\partial x}$	Partial derivative of function f with respect to x
$\lvert\mathcal{F}\rvert$	Cardinality of set $\mathcal{F}$
$[x]^+$	Maximum between zero and some real number x
$(\cdot)^T$	Transpose of a matrix
$\mathrm{sgn}(\cdot)$	Sign function
$(\cdot)_j$	Pochhammer symbol

Chapter 1

Introduction

1.1 Motivations

Delay quality-of-service (QoS) guarantees will play a critical role in 5G and beyond 5G wireless networks, due to the explosive growth of delay-sensitive wireless communication applications and networks, such as vehicular communications, E-health communication and Tactile Internet [1–9]. Extensive studies have been carried out, for systems with deterministic delay QoS requirements, where the delay is bounded within a certain threshold. However, satisfying a deterministic delay bound is practically infeasible for the time-varying fading channels, due to the random variations experienced in the channel conditions [10]. Specifically, in future mobile wireless networks, users are expected to tolerate various levels of delay for their service satisfactions [11]. Henceforth, to satisfy diverse users' delay requirements, a simple and flexible statistical delay QoS metric is imperative to be analyzed.

Note that conventional channel models directly characterize the fluctuations in the amplitude of a radio signal and then provide the physical-layer performance of wireless communication systems [10]. Hence, they can be called as the physical-layer channel models. However, physical-layer channel models cannot easily guarantee the delay QoS performance for a connection, such as queue distributions, buffer overflow probabilities, and delay violation probabilities [10, 12]. The reason is that, these complex delay QoS metrics need an analysis of the queueing behavior of the connection, which is hard to extract from the physical-layer models [10].

Recognizing the limitations of the physical-layer channel models in delay QoS support, the authors in [10] proposed a link-layer channel model termed effective capacity (EC), which is the dual of effective bandwidth (EB). The theory of EB was extensively studied in the early 90's with the emphasis on wired asynchronous transfer mode (ATM) networks [13–18]. By introducing a statistical envelope process which

deterministically bounds the moment generating function of the cumulative arrivals, the authors in [13] proposed the theory of EB, which gives the minimum service rate that is needed to support the probabilistic delay QoS requirements. As a dual of EB, EC, proposed in [10], denotes the maximum arrival rate that a given service process can support, on the condition that the required delay violation probability is guaranteed. Specifically, a comprehensive overview of the theory of EB and EC is provided in Chapter 2. Note that EC can be considered as the link-layer spectral efficiency (SE), while the link-layer energy efficiency (EE) can be formulated as the ratio of EC to the total power expenditure [19, 20]. Just like the inconsistent property of EE and SE in physical-layer channel model, the link-layer EE and EC are also incompatible [21]. In more detail, for a point-to-point communication system operating in a flat-fading channel, the EE versus EC curve is proved to be bell shape when non-zero circuit power is considered [22]. Hence, inspired by green communications, the focus first lies on designing an efficient resource allocation strategy to balance the three important QoS metrics, i.e., EE, SE and delay [22–26].

By applying the theory of EC, in this thesis, the delay-constrained resource allocation problem is first studied for delay-sensitive wireless communication networks. Focusing on a single-user single-carrier communication system, a multi-objective optimization problem (MOP) of link-layer EE and EC is first proposed and investigated, under a delay-outage probability constraint and an average transmit power limit. To solve the problem, an optimal power allocation strategy is proposed and proved to be sufficient for the Pareto optimal set of the original EE-EC MOP. To further balance these QoS metrics in a more practical scenario, a total EC maximization problem is proposed and investigated for the uplink transmission in a multi-user multi-carrier orthogonal frequency-division multiple access (OFDMA) system, subject to each user's link-layer EE requirement as well as the per-user average transmit power limit.

Apart from the delay-constrained resource allocation, the throughput analysis for different wireless communication networks is also of great importance. Note that non-orthogonal multiple access (NOMA) has been considered as a promising multiple access (MA) technique towards 5G networks. Current research work in NOMA-related areas mainly focuses on the topics such as cooperative design [27–29], subcarrier assignment and power control policy [30–32], physical layer security [33, 34], fairness analysis [35, 36], etc. Meanwhile, by adopting the theory of EC, the performance gain of NOMA over orthogonal multiple access (OMA), with a guaranteed statistical delay constraint, deserves elaborate study. Considering a downlink NOMA network with M users, the individual link-layer rate and the total achievable EC are studied

and analyzed in Chapter 5, while the per-user statistical delay QoS requirements are satisfied.

1.2 Thesis Outline and Contributions

1.2.1 Thesis Contributions

Motivated by the above discussions, this thesis focuses on different delay-sensitive wireless communication networks. Specifically, by applying the theory of EC and the link-layer channel model, the maximum achievable arrival rate with a guaranteed delay-outage probability constraint, is analyzed and investigated in terms of delay-constrained resource allocation and the link-layer throughput analysis. The main contributions of this thesis can be summarized as follows.

In Chapter 2, the theory of effective capacity, the convex optimization theory and literature review are provided. The background knowledge of large deviation theory and envelope processes is introduced first, which paves the way for deriving the theory of EB and EC. By deterministically bounding the moment generating function of the cumulative arrivals, the statistical envelope process proposed in [13] provides an upper bound on the traffic flows in a probabilistic manner. After applying the queueing theory and the large deviation theory, it is proved that the minimum envelope rate proposed in [13], which can be calculated from the Gärtner-Ellis limits, is the EB satisfying the required buffer overflow probability and the delay violation probability. Inspired by the theory of EB, the authors in [10] proposed the dual, i.e., the concept of EC. Specifically, EC denotes the maximum arrival rate that a given service process can support, on the condition that a target delay violation probability is guaranteed. After providing the concept of EC, the convex optimization theory is then briefly introduced in this chapter, followed by the literature review. Note that the included mathematical theorems and definitions regarding the EC theory and the convex optimization theory were from existing literature. However, this chapter only serves as a comprehensive overview, to help the readers to thoroughly understand the background knowledge.

In Chapter 3, the delay-QoS driven resource allocation problem is proposed and solved in a point-to-point single-user single-carrier communication system. To balance the three important QoS metrics, i.e., EE, EC and delay, a normalized link-layer EE-EC MOP is formulated on a Nakagami-m fading channel, under a delay-outage probability constraint and an average transmit power constraint. The MOP is then transformed into a power-constrained single-objective optimization problem (SOP),

by introducing two adjustable weights to the objectives. Focusing on the unconstrained EE-EC tradeoff problem first, a closed-form expression for the optimal power allocation strategy is derived to pave the way for the power-constrained problem. To solve the power-constrained EE-EC tradeoff problem, the Pseudocode of the optimal power allocation algorithm is then proposed. The impact of different system parameters on the optimal average power, such as the importance weight, normalization factor, circuit power, and power amplifier efficiency, is thoroughly analyzed. In more detail, this chapter has the following contributions:

- A generalized link-layer EE-EC MOP in a Nakagami-m fading channel under a delay-outage probability constraint and an average transmit power constraint is transformed into an SOP using weighted sum method. Specifically, two normalization values are introduced to balance the measurements and orders of magnitude of EE and EC.

- The unconstrained EE-EC tradeoff formulation is then proved to be continuously differentiable, strictly quasiconvex in the average power, which follows a cup shape curve. Henceforth, the global optimum is unique and can be achieved at a finite value.

- By using the Charnes-Cooper transformation and Karush-Kuhn-Tucker (KKT) conditions, an optimal power allocation scheme for the power-unconstrained link-layer EE-EC tradeoff problem is derived, and proved to be sufficient for the Pareto optimal set of the original EE-EC MOP. For the power-constrained tradeoff problem, the Pseudocode of the optimal power allocation algorithm is provided in Table 3.1.

- The average optimal power level is proved to be monotonically decreasing with the importance weight, but strictly increasing with the normalization factor, scaled circuit-to-noise power ratio and power amplifier efficiency.

- Finally, a proper guideline on how to choose the normalization factor and the importance weight to benefit either link-layer EE or EC is provided.

In Chapter 4, the focus lies on maximizing the total EC for the uplink transmission in a multi-user multi-carrier OFDMA network, subject to each user's required link-layer EE performance level and its individual resource limits. Firstly, the resource allocation problem is decoupled into two steps, i.e., the subcarrier assignment solution

and the optimal power allocation strategy for each user. In more detail, a low-complexity heuristic algorithm is proposed, which first allocates each served user the exact number of its required subcarriers, and then implements the optimal power allocation strategy for each user. Finally, the remaining subcarriers will be allocated by applying the strategy that the user with current minimum EC value has the allocation priority. To sum up, this chapter has the following contributions:

- A novel total EC maximization problem for the uplink transmission, in a multi-user multi-carrier OFDMA system, is formulated as a complex combinatorial integer programming problem, subject to each user's link-layer EE requirement and the individual's average input power limit. An adjustable EE requirement factor is introduced to further tune each user's EE constraint value, which transforms the formulated problem into a tradeoff problem between the total EC and the users' individual EE achievements.

- The formulated challenging problem is first decoupled into a frequency provisioning problem and an independent link-layer multi-carrier EE-EC tradeoff problem for each user. The traditional exhaustive algorithm and a fair-exhaustive algorithm are introduced first, followed by a low-complexity heuristic algorithm, which cares about user fairness, offers a close-to-optimal performance, and also has a complexity linearly relating to the size of the problem.

- The independent power-constrained link-layer EE-EC tradeoff problem is then solved and analyzed for each single-user multi-carrier system, given a subcarrier assignment matrix. The optimal power allocation strategy, which is across frequency and time domains, and the Pseudocode of the optimal power allocation algorithm are derived and proposed.

- The proposed per-user optimal average power level is proved to be monotonically decreasing with its EE requirement factor. Furthermore, the proposed per-user link-layer EE value is proved to be monotonically decreasing with its circuit power value, but increasing with its EE requirement factor.

- Simulation results reveal that when there is a link-layer EE constraint, each user's operational tradeoff EC value [1] will not show a monotonic trend with its delay QoS exponent. Further, the tradeoff EC value achieved with a smaller

[1]Here each user's operational tradeoff EC value is the calculated final EC value achieved at its EE requirement equality.

number of available subcarriers may be higher than the one obtained with larger number of subcarriers.

In Chapter 5, the achievable link-layer rate and the total achievable EC are studied for a downlink NOMA network with M users, under the per-user statistical delay QoS requirements. Specifically, the M users are assumed to be divided into multiple NOMA pairs, with coventional OMA applied for inter-NOMA-pairs multiple access. The performance gain of NOMA over OMA is investigated, by analyzing the impact of the transmit signal-to-noise ratio (SNR) and the delay QoS requirement on the performance of individual EC and the total link-layer rate. In more detail, this chapter has the following contributions:

- Focusing on a downlink M-user network, the individual EC and the total achievable link-layer rate are formulated and investigated. Assuming that M users are divided into multiple NOMA pairs, we prove that OMA achieves higher total EC than NOMA, at small SNRs. Further, simulation results show that NOMA outperforms OMA, at high SNRs.

- Focusing on a downlink M-user network, the total EC difference between NOMA and OMA becomes stable when the transmit SNR is extremely high.

- Focusing on a two-user network, the closed-form expressions for the link-layer rates for both users, in NOMA and OMA, are derived. The accuracy is then confirmed by comparing with the Monte Carlo simulation results.

- Focusing on a two-user network, the impact of the transmit SNR[2] and the delay QoS requirement is analyzed in two cases, for both NOMA and OMA scenarios. Case 1: consider delay-constrained users; Case 2: consider delay-unconstrained users.

- In Case 1 and Case 2, we characterize the region of the transmit SNR, in which NOMA outperforms OMA, in terms of the individual and the total EC for the two-user system.

[2]The transmit SNR is defined as the ratio of the transmission power to the noise power, in which the noise is assumed to be the additive white Gaussian noise. Further details will be provided in the next section.

1.2.2 Thesis Outline

The remainder of this thesis is organized as follows. In Chapter 2, the theory of EC and convex optimization is introduced first, followed by a brief literature review on the resource allocation towards green communications and the current research progress in NOMA networks. In Chapter 3, the delay-constrained resource allocation for a joint optimization problem of link-layer EE and EC is analyzed and solved in a single-user single-carrier communication system. As a natural extension of Chapter 3, Chapter 4 studies the delay-constrained resource allocation which solves the total EC maximization problem for the uplink transmission in a multi-user multi-carrier system, subject to each user's link-layer EE requirement and the individual power limit. In Chapter 5, the link-layer throughput analysis and the total achievable EC are studied and analyzed for a downlink M-user NOMA network, under the per-user statistical delay QoS requirements. Finally, in Chapter 6, this thesis is summarized and the recommendations for future research are provided.

Chapter 2

Background Theory and Literature Review

2.1 The Theory of Effective Capacity

2.1.1 Large Deviation Theory

Let $X_1, X_2, \ldots,$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean $\mu = \mathbb{E}[X_1] < \infty$, and let $M_n = \frac{1}{n}(X_1 + \cdots + X_n)$ denote the empirical mean. From the weak law of large numbers, it is noted that for any $\epsilon > 0$, $P(|M_n - \mu| > \epsilon) \to 0$, as $n \to \infty$ [37]. But how fast is this convergence? This falls into the scope of the theory of large deviations [38, 39]. Large deviation theory includes a set of techniques for turning difficult probability problems dealing with a class of rare events into analytic problems in the calculus of variations [38].

To find out the decay rate, the probability of the empirical mean exceeding a is considered, where a is a value larger than μ, i.e., $a > \mu$. Then, by fixing a positive parameter $\theta > 0$, we get [37]

$$P\left(\sum_{1\le i\le n} X_i > an\right) = P\left(e^{\theta \sum_{1\le i\le n} X_i} > e^{\theta an}\right) \tag{2.1}$$

$$\le \frac{\mathbb{E}\left[e^{\theta \sum_{1\le i\le n} X_i}\right]}{e^{\theta an}} \tag{2.2}$$

$$= \frac{\mathbb{E}\left[\prod_i e^{\theta X_i}\right]}{\left(e^{\theta a}\right)^n} \tag{2.3}$$

$$= \left(\frac{\mathbb{E}\left[e^{\theta X_1}\right]}{e^{\theta a}}\right)^n. \tag{2.4}$$

From (2.1) to (2.2), it is derived by utilizing the Markov inequality [38]. From (2.3)

to (2.4), it is due to the reason that the random variables X_i, $i \in [1, n]$, are i.i.d. [38]. Finally, (2.4) can be considered as an upper bound for the tail probability. For the bound to be meaningful and useful, $\mathbb{E}\left[e^{\theta X_1}\right]$ needs to exist and $\frac{\mathbb{E}\left[e^{\theta X_1}\right]}{e^{\theta a}}$ needs to be less than 1. Here, $\mathbb{E}\left[e^{\theta X_1}\right]$ is the moment generating function (MGF) of X_1 and can be denoted by $M_X(\theta)$[1].

Definition. Let X be a random variable. The MGF of X is defined by [40]

$$M_X(\theta) = \mathbb{E}[e^{\theta X}] = \begin{cases} \sum_x e^{\theta x} P_X(x), & \text{if X is discrete with PMF } P_X(x), \\ \int_{-\infty}^{\infty} e^{\theta x} f_X(x) dx, & \text{if X is continuous with PDF } f_X(x). \end{cases}$$

The domain D_X of $M_X(\theta)$ is defined as the set $D_X = \{\theta \in \mathbb{R} | M_X(\theta) < \infty\}$.

Henceforth, D_X is the set of θ for which the MGF is finite, i.e., when the sum or integral given above converges [41]. Furthermore, according to [37], it can be proved that the ratio $\frac{\mathbb{E}\left[e^{\theta X_1}\right]}{e^{\theta a}} < 1$, for sufficiently small positive θ values, for any $a > \mu$. Similarly, if $a < \mu$, $\frac{\mathbb{E}\left[e^{\theta X_1}\right]}{e^{\theta a}} < 1$ holds for sufficiently small negative θ values[2]. Hence, for $a > \mu$, one can conclude that for sufficiently small positive values of θ in D_X[3], (2.4) provides an exponential bound on the tail probability for the empirical mean.

Note that the reason for calling $M_X(\theta)$ moment generating function is due to the Taylor expansion of $e^{\theta X}$ [41, 42]. By assuming that it converges, we have

$$M_X(\theta) = \mathbb{E}[e^{\theta X}] = \mathbb{E}\left[1 + \theta X + \frac{1}{2}\theta^2 X^2 + \frac{1}{3!}\theta^3 X^3 + \dots\right] = \sum_{i=0}^{\infty} \frac{1}{i!}\theta^i \mathbb{E}[X^i].$$

The terms $\mathbb{E}[X^i]$ are called "moments" and include important information about the distribution. Through the MGF, all the moments of this distribution can be calculated. For example, if a MGF exists for a random variable X, then the mean of X can be found by evaluating the first derivative of the MGF at $\theta = 0$, i.e., $\mathbb{E}[X] = M'_X(0)$. The variance of X can be found by evaluating the first and second derivatives of the MGF at $\theta = 0$, i.e., $M''_X(0) - (M'_X(0))^2$. Another important property of MGF is that it has a one-to-one correspondence with the random variable's probability distribution. In other words, for any distribution there is a unique MGF that characterizes it (if it exists) and for each MGF there is a unique probability distribution it characterizes [43].

[1]Here we use $M_X(\theta)$ rather than $M_{X_1}(\theta)$, to denote the identical MGF for the i.i.d. sequence $\{X_i, i = 1, \dots, n\}$.

[2]The proof is omitted here for simplicity. Please refer to [37] for the complete information.

[3]The domain D_X is an interval containing zero [37].

Recall that for any $a > \mu$, by fixing a positive θ, an exponential bound on the tail probability $P\left(\sum_{1\leq i\leq n} X_i > an\right)$ has been found in (2.4). On the other hand, if $a < \mu$, by fixing a negative $\theta < 0$, one can get that

$$P\left(\sum_{1\leq i\leq n} X_i < an\right) = P\left(e^{\theta \sum_{1\leq i\leq n} X_i} > e^{\theta an}\right) \leq \left(\frac{M_X(\theta)}{e^{\theta a}}\right)^n. \tag{2.5}$$

Finally, the above findings can be summarized in the theorem below [37].

Theorem 1. *Given an i.i.d. sequence $X_1, \ldots, X_n$, the MGF $M_X(\theta)$ is assumed to be finite for all θ in some neighborhood $\mathcal{B}_0$ of $\theta = 0$. Let $a > \mu = \mathbb{E}[X_1]$. Then there exists $\theta > 0$, such that $\frac{M_X(\theta)}{e^{\theta a}} < 1$ and*

$$P\left(\frac{\sum_{1\leq i\leq n} X_i}{n} > a\right) \leq \left(\frac{M_X(\theta)}{e^{\theta a}}\right)^n. \tag{2.6}$$

Similarly, if $a < \mu$, then there exists $\theta < 0$, such that $\frac{M_X(\theta)}{e^{\theta a}} < 1$ and

$$P\left(\frac{\sum_{1\leq i\leq n} X_i}{n} < a\right) \leq \left(\frac{M_X(\theta)}{e^{\theta a}}\right)^n. \tag{2.7}$$

Proof. The proof follows the above analysis. □

Theorem 1 provides an exponential bound on the tail probability for the empirical average of $X_1, \ldots, X_n$. But how tight can the bound be? Since θ can be varied as long as $M_X(\theta)$ is finite, therefore the value of θ which minimizes the ratio $\frac{M_X(\theta)}{e^{\theta a}}$ needs to be found [37]. By rewriting $\left(\frac{M_X(\theta)}{e^{\theta a}}\right)^n$ as $e^{-n(\theta a - \ln M_X(\theta))}$, we get[4] [44]

$$\inf_\theta \left(\frac{M_X(\theta)}{e^{\theta a}}\right)^n = \inf_\theta e^{-n(\theta a - \ln M_X(\theta))} \tag{2.8}$$

$$= e^{\inf_\theta -n(\theta a - \ln M_X(\theta))} \tag{2.9}$$

$$= e^{-n \sup_\theta (\theta a - \ln M_X(\theta))}. \tag{2.10}$$

From (2.8) to (2.9), it is due to the reason that the exponential function is a monotonically increasing function. From (2.9) to (2.10), it is derived by applying the

[4]Here, $\sup_\theta f(\theta)$ is the supremum of $f(\theta)$, which represents the least upper bound. Meanwhile, $\inf_\theta f(\theta)$ is the infimum of $f(\theta)$, which denotes the greatest lower bound. Further, if the minimum (or maximum) exists, then it must be the infimum (or supremum).

Proposition 11.4 in [45], which says that if $c < 0, c \in \mathbb{R}$, then $\inf_A cf = c \sup_A f$, for a bounded function $f : A \to \mathbb{R}$. Then, let us apply the infimum of the upper bound to the first part of Theorem 1. Hence, (2.6) can be transformed into

$$P\left(\frac{\sum_{1\le i\le n} X_i}{n} > a\right) \le \inf_{\theta>0}\left(\frac{M_X(\theta)}{e^{\theta a}}\right)^n = e^{-n\sup_{\theta>0}(\theta a - \ln M_X(\theta))}. \tag{2.11}$$

According to [46], the tightness of the above bound can be confirmed. In other words, it can be proved that $\lim_{n\to\infty}\frac{1}{n}\log P\left(\frac{\sum_{1\le i\le n} X_i}{n} > a\right) = -\sup_{\theta>0}(\theta a - \ln M_X(\theta))$, based on the assumption that the supremum can be obtained at some interior point in the neighborhood $\mathcal{B}_0$. Further, $\sup_{\theta>0}(\theta a - \ln M_X(\theta))$ is called Legendre transform, defined below [37].

Definition. A Legendre transform of a convex function $\Lambda(\theta)$ is defined by $I(x) = \sup_\theta(\theta x - \Lambda(\theta))$. The domain D_x of $I(x)$ is given as $\{x \in \mathbb{R} | \sup_\theta(\theta x - \Lambda(\theta)) < \infty\}$.

Suppose that $\Lambda(\theta) = \ln M_X(\theta)$ and finite MGF $M_X(\theta)$ exists for all θ. According to [37,45], it is proved that $\ln M_X(\theta)$, which is called the log moment generating function or the cumulant generating function, is convex. Then, the Legendre transform[5] $I(x) = \sup_\theta(\theta x - \ln M_X(\theta)), x \in \mathbb{R}$, is proved to be a convex (being the supremum of linear, hence a convex function) and non-negative function, with its minimum $I(\mu) = 0$ obtained at the mean value $\mu = \mathbb{E}[X_1]$ [37, 38, 45]. Furthermore, it is shown to be an increasing function on $[\mu, \infty)$, and a decreasing function on $(-\infty, \mu]$ [37].

Note that the tail probability described in Theorem 1 considers relatively simple situations. To deal with more complicated rare events, like the likelihood of $P(\frac{\sum_{1\le i\le n} X_i}{n} \in A)$ for some set $A \subset \mathbb{R}$, a more generalized theorem is needed [47,48].

Theorem 2. *(Cramér Theorem) Let $X_1, \ldots, X_n$, be a sequence of i.i.d. real valued random variables with $S_n = \frac{\sum_{1\le i\le n} X_i}{n}$, which satisfies the large deviation principle with the convex rate function $I(x) = \sup_\theta(\theta x - \ln M_X(\theta))$.*

1. *For any closed set $F \subset \mathbb{R}$, $\limsup_{n\to\infty}\frac{1}{n}\log P(S_n \in F) \le -\inf_{x\in F} I(x)$,*
2. *For any open set $U \subset \mathbb{R}$, $\liminf_{n\to\infty}\frac{1}{n}\log P(S_n \in U) \ge -\inf_{x\in U} I(x)$.*

Proof. The proof is omitted here. Please refer to [37] for further information. □

[5]The Legendre transform of $\ln M_X(\theta)$ is also commonly called the rate function in the theory of Large Deviations.

Here, $\limsup_{n\to\infty} f(n)$ is defined as $\limsup_{n\to\infty} f(n) = \lim_{n\to\infty} g(n)$, where $g(n) = \sup_{k\geq n} f(k)$, representing the supremum of $f(k)$ with $k \geq n$. Similarly, $\liminf_{n\to\infty} f(n)$ is defined as $\liminf_{n\to\infty} f(n) = \lim_{n\to\infty} h(n)$, where $h(n) = \inf_{k\geq n} f(k)$, representing the infimum of $f(k)$ with $k \geq n$ [45]. Furthermore, according to [45, 49], it is noted that for any function $f(t)$, if its limit exists, i.e., $\lim_{t\to\infty} f(t) = f^*$ (where f^* is possibly infinite), then both the lim sup and lim inf of the function are equal to f^*, i,e, $\limsup_{t\to\infty} f(t) = \liminf_{t\to\infty} f(t) = f^*$. Conversely, if $\limsup_{t\to\infty} f(t) = \liminf_{t\to\infty} f(t)$, then the regular limit also exists and is equal to the same value [49].

To see that Theorem 2 is a generalization of Theorem 1, we provide the following analysis [44, 47]. Set $F = [a, \infty)$[6], and $U = (a, \infty)$, $a \geq \mu = \mathbb{E}[X_1]$. For $x \in [a, \infty)$, $I(x)$ monotonically increases with a minimum value achieved at $x = a$. Hence, by applying the first part of Cramér Theorem, we have $\limsup_{n\to\infty} \frac{1}{n} \log P\left(S_n \geq a\right) \leq -\inf_{x\geq a} I(x) = -\min_{x\geq a} I(x) = -I(a)$. Furthermore, due to the reason that $F \supset U$, one can get that $P(S_n \in F) \geq P(S_n \in U)$, i.e., $P(S_n \geq a) \geq P(S_n > a)$. By applying the second part of Cramér Theorem, we get $\liminf_{n\to\infty} \frac{1}{n} \log P\left(S_n \geq a\right) \geq \liminf_{n\to\infty} \frac{1}{n} \log P\left(S_n > a\right) \geq -\inf_{x>a} I(x) = -I(a)$[7]. Henceforth, one can conclude that $\limsup_{n\to\infty} \frac{1}{n} \log P\left(S_n \geq a\right) = \liminf_{n\to\infty} \frac{1}{n} \log P\left(S_n \geq a\right)$, which equals to the regular limit, i.e., $\lim_{n\to\infty} \frac{1}{n} \log P\left(S_n \geq a\right) = -I(a)$ [37, 44]. Since the limit is insensitive to whether the inequality is strict, hence we can get that $\lim_{n\to\infty} \frac{1}{n} \log P\left(S_n > a\right) = -I(a)$ [37], which confirms the analysis following Theorem 1.

Note that Cramér Theorem only applies to a sequence of i.i.d. random variables. For a sequence of not necessarily independent random variables, the Gärtner-Ellis Theorem provided below can be utilized to deal with large deviation events [50–53]. Consider a sequence of random variables $\{Y_n, n \geq 1\}$. Let $\Lambda_n(\theta) = \frac{1}{n} \log\left(\mathbb{E}\left[e^{\theta Y_n}\right]\right)$. Note that $\Lambda_n(\theta)$ can be proved to be a convex function via Hölder's inequality [52,53][8].

Theorem 3. *(Gärtner-Ellis Theorem [53])Assume*
*(**A1**)* $\lim_{n\to\infty} \Lambda_n(\theta) = \Lambda(\theta) < \infty$ *for all* $\theta \in R$,
*(**A2**)* $\Lambda(\theta)$ *is differentiable for all* $\theta \in R$.
Let $\Lambda^*(a) = \sup_{\theta} \theta a - \Lambda(\theta)$.

[6]Note that the set F is closed, since its complement $(-\infty, a)$ is an open set.
[7]This step requires the property of lim inf: $\liminf_{t\to\infty} f(t) \geq \liminf_{t\to\infty} g(t)$, if $f(t) \geq g(t)$, for all t.
[8]The proof is omitted here for simplicity. Please refer to [52] for further information.

1. *(Upper bound) For every close set* $F \subset R$, $\limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{Y_n}{n} \in F\right) \leq -\inf_{a\in F} \Lambda^*(a)$.

2. *(Lower bound) For every open set* $U \subset R$, $\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{Y_n}{n} \in U\right) \geq -\inf_{a\in U} \Lambda^*(a)$.

Proof. The proof is omitted here. Please refer to [37] for further information. □

Gärtner-Ellis Theorem states that when the scaled cumulant generating function of Y_n, i.e., $\Lambda_n(\theta)$, is differentiable and converges, the large deviation principle holds for $\frac{Y_n}{n}$ with a rate function $\Lambda^*(a)$ given by the Legendre-Fenchel transform of $\Lambda(\theta)$[9].

2.1.2 Envelope Process

To support quality-of-service (QoS) guarantees in communication networks, it is important to characterize the source traffic and the network service, matched using a First-In-First-Out (FIFO) buffer [10]. The most widely used approach for traffic characterization, is to require that the cumulative arrival traffic $A(t)$ of a flow over any interval of length t conforms to an upper bound, called the traffic envelope $\hat{A}(t)$. Correspondingly, the service characterization is a guarantee of a minimum service level, specified by a service envelope $\hat{C}(t)$ [10]. Such traffic and service envelopes could be deterministic (i.e., strict bounds) or statistical (i.e., violation is allowed, but with a small probability), which can be used for provisioning of deterministic or statistical service guarantees [54], such as a bounded delay or a delay violation probability. In this section, the deterministic envelope process is briefly introduced first, followed by the statistical envelope process, which leads to the concept of effective bandwidth.

2.1.2.1 Deterministic Envelope Process

We first describe a discrete-time arrival process of a traffic source by a sequence of variables $\{a(t), t = 0, 1, 2, \dots\}$. Let $A(t_1, t_2)$ be the cumulative number of arrivals in the time interval $(t_1, t_2]$, i.e., $A(t_1, t_2) = \sum_{t_1+1}^{t_2} a(t)$. Assume that there is no arrival at time 0, and that $A(t)$ is nondecreasing, i.e., $A(t_1) \leq A(t_2)$, for all $t_1 \leq t_2$.

Generally, for such traffic flows, a deterministic envelope process could be any nondecreasing, nonnegative function, as long as the cumulative traffic is bounded [13, 54], i.e., $A(t_1, t_2) \leq \hat{A}(t_2 - t_1), \forall t_1 \leq t_2$. In this section, we only focus on the

[9]Here, the Legendre-Fenchel transform is a generalization of the Legendre transform, which applies to non-convex functions.

simple linear envelope process proposed in [55][10]:

$$A(t_1, t_2) \leq \hat{a}(t_2 - t_1) + \sigma, \qquad \forall t_1 \leq t_2, \tag{2.12}$$

where σ is called the burstiness parameter and $\hat{a}$ can be considered as an upper bound on the long-term average rate of the traffic flow [55][11]. Since $A(t_1, t_2)$ is the number of arrivals in the interval $(t_1, t_2]$, hence, the linear envelope process in (2.12) basically imposes an upper bound on the number of arrivals within a time interval. Furthermore, the proposed linear envelope process is formulated as a function of the time interval $\tau = t_2 - t_1$, regardless where the interval begins [13].

Since envelope processes are not unique, therefore it is important to find the tightest one, which satisfies $A^*(t) = \sup_{s \geq 0} A(s, s+t)$. That is, $A^*(t)$ is called the minimum envelope process (MEP) of $A(t)$. According to [13], it is known that the MEP $A^*(t)$ is increasing and subadditive[12], and the minimum envelope rate (MER) is defined as $a^* = \lim_{t \to \infty} \frac{A^*(t)}{t}$.

While the deterministic traffic bound looks intuitive, a drawback is that it generally considers the "hard" performance guarantees, such as the worst-case delay bounds and no packet dropped in the network [54]. As a consequence, it cannot take advantage of the statistical nature of traffic [57]. In addition, hard performance guarantees might be an overkill for some applications, where a certain amount of loss or delay violation is tolerable. For example, in fading communication networks, it is especially challenging and unnecessary to satisfy a strict deterministic delay bound, due to the random variations experienced in channel conditions, user mobility and changing environment.

2.1.2.2 Statistical Envelope Process

As opposed to the deterministic approach, a statistical envelope process bounds traffic flows in a probabilistic manner, and provides "soft" QoS guarantees, statistically [54]. There are various statistical envelope processes, but in this section the focus lies on the stochastic traffic characterization proposed in [13]. Specifically, this envelope process deterministically bounds the moment generating function of the cumulative arrival $A(t)$ and supports probabilistic delay QoS guarantees. Note that if the MGF of a random variable X is bounded by a finite constant D as $\left(\mathbb{E}[e^{\theta X}]\right)^{1/\theta} \leq D$, then

[10] For more generalized deterministic envelope processes, please refer to [54] for further information.

[11] An arbitrary traffic flow can be policed to be confined to the linear envelope process, by using a token bucket with a token rate $\hat{a}$ and a token bucket size σ [54, 56].

[12] A process $A^*(t)$ is subadditive if $A^*(t_1 + t_2) \leq A^*(t_1) + A^*(t_2)$.

from Chernoff's bound [58], its distribution is bounded exponentially with respect to θ as

$$P(X \geq x) \leq D^{\theta} e^{-\theta x}, \qquad \text{for all } x. \tag{2.13}$$

Henceforth, by deterministically bounding the MGF of the cumulative arrival $A(t)$, the arrival traffic itself can be bounded in a probabilistic way.

The mathematical expression of the statistical envelope process proposed in [13] can be given as

$$\frac{1}{\theta} \log \left(\mathbb{E} \left[e^{\theta A(t_1, t_2)} \right] \right) \leq \hat{A}(\theta, t_2 - t_1), \quad \forall t_1 \leq t_2, \tag{2.14}$$

where $\hat{A}(\theta, t)$ is called an θ-envelope process of $A(t)$ [13]. Similar to the deterministic case, the θ-MEP of $A(t)$ is defined as $A^*(\theta, t) = \sup_{s \geq 0} \frac{1}{\theta} \log \left(\mathbb{E} \left[e^{\theta A(s, s+t)} \right] \right)$, and the θ-MER is given as $a^*(\theta) = \limsup_{t \to \infty} \frac{A^*(\theta, t)}{t}$ [13].

By applying the Chernoff bound in (2.13) and the Lindley's equation from queueing theory, the above statistical envelope process can be used to derive the probabilistic delay QoS measures, which leads to the concept of effective bandwidth.

2.1.3 Effective Bandwidth

Consider a discrete-time FIFO queue with a single link. Let $a(t)$ and $q(t)$ be the number of arrivals at time t and the number of packets in the queue at time t, respectively. Assume that the buffer size is infinite and that the link can serve $c(t)$ packets per unit of time, which means that the capacity of the link at time t is $c(t)$. If the link has a constant capacity, then $c(t) = c$, for all t. The link works under a work-conserving policy, i.e., a policy that does not allow idling when there are packets in the queue. Further, $q(t)$ converges to a steady state $q(\infty)$, if both $a(t)$ and $c(t)$ are stationary and ergodic, and $\mathbb{E}[a(t)] < \mathbb{E}[c(t)]$ [43, 53].

Let $A(t_1, t_2) = \sum_{t=t_1+1}^{t_2} a(t)$ be the total number of arrivals in the time interval $(t_1, t_2]$, and $C(t_1, t_2) = \sum_{t=t_1+1}^{t_2} c(t)$. Before deriving the theory of effective bandwidth (EB), the authors in [16, 53] proposed a theorem as follows.

Theorem 4. *Let us make the following assumptions first.*
*(**A1**) $a(t)$ and $c(t)$ are independent.*
*(**A2**) For all $\theta \in R$, $\lim_{t \to \infty} \frac{1}{t} \log \left(\mathbb{E} \left[e^{\theta A(0,t)} \right] \right) = \Lambda_A(\theta)$ and $\Lambda_A(\theta)$ is differentiable.*

*(**A3**) For all $\theta \in R$, $\lim_{t\to\infty} \frac{1}{t} \log\left(\mathbb{E}\left[e^{\theta C(0,t)}\right]\right) = \Lambda_C(\theta)$ and $\Lambda_C(\theta)$ is differentiable.*
*(**A4**) Both $a(t)$ and $c(t)$ are stationary and ergodic, and $\mathbb{E}[a(t)] < \mathbb{E}[c(t)]$.*
If there exists a unique $\theta^* > 0$ such that

$$\Lambda_A(\theta^*) + \Lambda_C(-\theta^*) = 0, \tag{2.15}$$

then we can get

$$\lim_{x\to\infty} \frac{\log\left(\Pr\left(q(\infty) \geq x\right)\right)}{x} = -\theta^*. \tag{2.16}$$

Proof. The proof is omitted here for simplicity. Please refer to [16, 53]. □

Specifically, when the capacity is a fixed constant, i.e., $c(t) = c$ for all t, $\Lambda_C(-\theta^*)$ reduces to

$$\Lambda_C(-\theta^*) = \lim_{t\to\infty} \frac{1}{t} \log\left(e^{-\theta^* ct}\right) = -\theta^* c. \tag{2.17}$$

By inserting (2.17) into (2.15), we get that $\frac{\Lambda_A(\theta^*)}{\theta^*} = c$. Henceforth, one can conclude that, when the capacity is fixed as a constant c, the condition needed to satisfy the queue overflow probability is that $\frac{\Lambda_A(\theta^*)}{\theta^*} = c$. In other words, $\frac{\Lambda_A(\theta^*)}{\theta^*}$ can be considered as the bandwidth (approximated) needed to guarantee the queue overflow probability [53]. Hence, $\frac{\Lambda_A(\theta)}{\theta}$, denoted by $E_{\mathrm{b}}(\theta)$, is called the effective bandwidth of the arrival process, on the condition that the tail distribution of the queue length has the decay rate θ [16].

Furthermore, the assumptions (A2-3), known as the Gärtner-Ellis limits [50, 51], connects the large deviation theory with the θ-MER introduced in Section 2.1.2.2 [53]. To establish the connection, we recall that $\{a(t), t \geq 0\}$ is stationary and ergodic. Hence, the θ-MEP of the arrival process, i.e., $A^*(\theta, t)$, equals to $\frac{1}{\theta} \log\left(\mathbb{E}\left[e^{\theta A(0,t)}\right]\right)$. Further, the θ-MER of $A(t)$, i.e., $a^*(\theta)$, defined as $\limsup_{t\to\infty} \frac{1}{\theta t} \log\left(\mathbb{E}\left[e^{\theta A(0,t)}\right]\right)$, equals to $\lim_{t\to\infty} \frac{1}{\theta t} \log\left(\mathbb{E}\left[e^{\theta A(0,t)}\right]\right)$[13], due to the reason that the Gärtner-Ellis limit exists. Hence, one can conclude that the θ-MER of $A(t)$, equal to $\frac{\Lambda_A(\theta)}{\theta}$, is the effective bandwidth of the arrival process [53], when the Gärtner-Ellis limit of $A(t)$ exists.

[13]Note that for any function $f(t)$, if its limit exists, i.e, $\lim_{t\to\infty} f(t) = f^*$, then $\limsup_{t\to\infty} f(t) = f^*$.

2.1.4 Effective Capacity

Inspired by the theory of EB, the authors in [10] proposed the concept of effective capacity (EC), as a dual of EB. Let $\{c(t), t = 0, 1, 2, \dots\}$ be a discrete-time service process, which is stationary and ergodic. $C(t_1, t_2) = \sum_{t=t_1+1}^{t_2} c(t)$ denotes the partial sum. Assume that the Gärtner-Ellis limit of $C(t)$, expressed as $\lim_{t\to\infty} \frac{1}{t} \log\left(\mathbb{E}\left[e^{\theta C(0,t)}\right]\right) = \Lambda_C(\theta)$, exists and is a differentiable convex function for all $\theta \in R$ [16]. Consider that the arrival rate is a constant, i.e., a. Therefore, by applying Theorem 4 in Section 2.1.3, one can get that

$$\Lambda_C(-\theta^*) = -\Lambda_A(\theta^*) = -\theta^* a, \tag{2.18}$$

where θ^* is the unique delay QoS exponent satisfying (2.16). From (2.18), it is noted that $-\frac{\Lambda_C(-\theta^*)}{\theta^*} = a$, which can be considered as the effective capacity of the service process, on the condition that the queue overflow probability can be guaranteed with a decay rate θ^*. Finally, $-\frac{\Lambda_C(-\theta^*)}{\theta^*}$, denoted by $E_{\mathrm{c}}(\theta)$, can be calculated from the Gärtner-Ellis limit:

$$E_{\mathrm{c}}(\theta) = -\frac{\Lambda_C(-\theta^*)}{\theta^*} = -\lim_{t\to\infty} \frac{1}{\theta t} \log\left(\mathbb{E}\left[e^{-\theta C(0,t)}\right]\right). \tag{2.19}$$

Furthermore, let us define ϵ as the required queue overflow probability limit. In other words, the maximum queue overflow probability that can be afforded is given as ϵ. In this case, by applying (2.16), the minimum decay rate θ^* can be calculated as $\theta^* = -(\log\epsilon)/x$. Inserting the minimum decay rate θ^* into (2.19), a maximum value of $E_{\mathrm{c}}(\theta^*)$ satisfying the queue overflow probability limit can be found, since EC is a monotonically decreasing function with the delay QoS exponent. Hence, one can say that, in order to guarantee a required queue overflow probability limit, the calculated effective capacity $E_{\mathrm{c}}(\theta^*)$ represents the maximum constant arrival rate that the service process can support.

Further, when the sequence $\{c(t), t = 0, 1, 2, \dots\}$ is uncorrelated, the effective

capacity reduces to

$$E_{\mathrm{c}}(\theta) = -\lim_{t\to\infty}\frac{1}{\theta t}\log\left(\mathbb{E}\left[e^{-\theta\sum_{i=1}^{t}c(i)}\right]\right) \tag{2.20}$$

$$= -\lim_{t\to\infty}\frac{1}{\theta t}\log\left(\mathbb{E}\left[\prod_{i=1}^{t}e^{-\theta c(i)}\right]\right) \tag{2.21}$$

$$= -\lim_{t\to\infty}\frac{1}{\theta t}\log\left(\prod_{i=1}^{t}\mathbb{E}\left[e^{-\theta c(i)}\right]\right) \tag{2.22}$$

$$= -\lim_{t\to\infty}\frac{1}{\theta t}\log\left(\mathbb{E}\left[e^{-\theta c(i)}\right]\right)^{t} \tag{2.23}$$

$$= -\frac{1}{\theta}\log\left(\mathbb{E}\left[e^{-\theta c(i)}\right]\right). \tag{2.24}$$

From (2.21) to (2.22), it is due to the reason that the sequence $\{c(t), t = 0, 1, 2, \dots\}$ is uncorrelated. From (2.22) to (2.23), it is because that the service process is stationary and ergodic. Apparently, when the service process is uncorrelated, the EC expression in (2.24) only depends on marginal statistics, which is much simpler than the general expression given in (2.19), where the higher-order statistics are required [11]. Since the block fading channel generates an i.i.d., hence uncorrelated, service process, it can greatly simplify the EC expressions [11].

Note that the above introduction of EC assumes the constant arrival rate. Actually, it can be generalized to investigate the delay QoS performance of any stationary arrival process [11]. By rewriting (2.15) in Theorem 4 in Section 2.1.3, we can get that if there exists a unique $\theta^* > 0$ such that

$$\frac{\Lambda_A(\theta^*)}{\theta^*} = -\frac{\Lambda_C(-\theta^*)}{\theta^*}, \tag{2.25}$$

then we have

$$\lim_{x\to\infty}\frac{\log\left(\Pr\left(q(\infty)\geq x\right)\right)}{x} = -\theta^*. \tag{2.26}$$

Since $\dfrac{\Lambda_A(\theta^*)}{\theta^*} = E_{\mathrm{b}}(\theta^*)$ denotes the EB and $-\dfrac{\Lambda_C(-\theta^*)}{\theta^*} = E_{\mathrm{c}}(\theta^*)$ is the EC, hence, (2.25) and (2.26) indicate that the EB function intersects with the EC function at the point where the delay QoS exponent is θ^*. Here, θ^* is the one which guarantees the queue overflow probability limit.

To thoroughly understand the relationship between EB and EC when the time-varying arrival process and service process are considered, Fig. 2.1 is included which shows the curves of EB and EC versus the delay QoS exponent θ [11]. Set $\mu_a =$

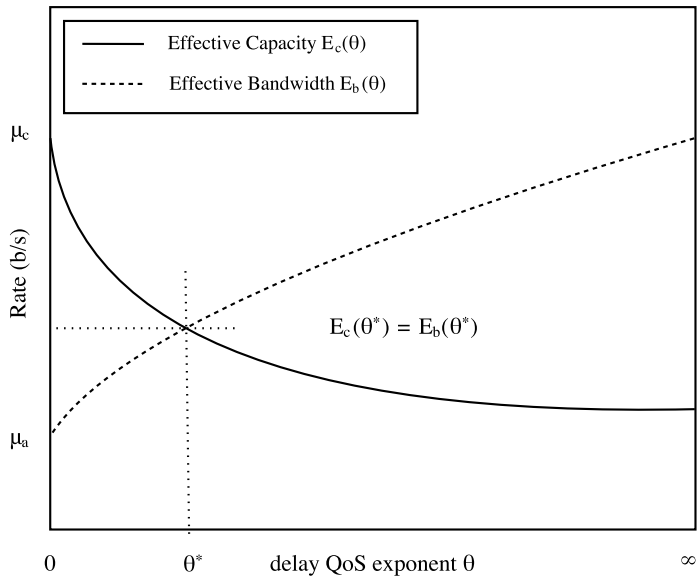


Figure 2.1: EC and EB, as functions of the delay QoS exponent θ.

$\lim_{\theta\to 0} E_{\rm b}(\theta)$, and $\mu_c = \lim_{\theta\to 0} E_{\rm c}(\theta)$. From Fig. 2.1, it shows that when the minimum value of EB is larger than the maximum value of EC, i.e., $\mu_a > \mu_c$, there is no solution for $\theta^* > 0$ existing. In this case, the service process cannot support the required delay QoS for the given arrival process, which is consistent with the conclusion from queueing theory that, if $\mathbb{E}[a(t)] > \mathbb{E}[c(t)]$, both queue length and the queueing delay will approach to infinity. This is because that, when $\theta \to 0$, the EB is equal to the average arrival rate of the traffic process, i.e., $\mu_a = \lim_{\theta\to 0} E_{\rm b}(\theta) = \mathbb{E}[a(t)]$. Meanwhile, when $\theta \to 0$, the EC is equal to the average service rate of the service process, i.e., $\mu_c = \lim_{\theta\to 0} E_{\rm c}(\theta) = \mathbb{E}[c(t)]$ [11].

In the above analysis, the buffer overflow probability was considered as the delay QoS measurements. When the focus is on the delay experienced by a source packet arriving at time t, defined by $D(t)$, an expression analogous to (2.26) can be estimated as [10, 53]

$$\Pr\left(D(t) > D_{\max}\right) \approx \Pr\left(q(t) > 0\right) e^{-\theta\mu D_{\max}}, \tag{2.27}$$

where $D_{\max}$ denote the delay bound, and $\Pr\left(q(t) > 0\right)$ is the probability of a non-empty buffer, which can be approximated by the ratio of the average arrival rate and the average service rate [11], i.e., $\dfrac{\mathbb{E}[a(t)]}{\mathbb{E}[c(t)]}$. Furthermore, from [11], we note that $\mu = E_{\rm c}(\theta) = E_{\rm b}(\theta)$, when a time-varying arrival process is considered.

Considering the delay violation probability in (2.27) as a function of θ, one can notice that the parameter θ plays an important role for statistical QoS guarantees, by indicating the exponential decay rate of the delay QoS violation probability [11]. A smaller θ corresponds to a slower decay rate, which implies that the system can tolerate a looser QoS guarantee, while a larger θ indicates a faster decay rate, which means that a more stringent QoS requirement can be supported. In particular, when $\theta \to 0$, the system can tolerate an arbitrarily long delay. When $\theta \to \infty$, it indicates that the system cannot tolerate any delay [12].

2.2 Convex Optimization Theory

2.2.1 Convex Optimization Problems

Normally, an optimization problem has the following form [59, 60]

$$\begin{aligned} \min \quad & f_0(\boldsymbol{x}) && \text{(2.28a)} \\ \text{subject to:} \quad & f_i(\boldsymbol{x}) \leq b_i,\ i = 1, \ldots, m && \text{(2.28b)} \\ & h_i(\boldsymbol{x}) = 0,\ i = 1, \ldots, p. && \text{(2.28c)} \end{aligned}$$

Here, the vector $\boldsymbol{x} = [x_1, \ldots, x_n]$ is the optimization variable of the problem, the function $f_0 : \boldsymbol{R}^n \to \boldsymbol{R}$ is the objective function, the functions $f_i : \boldsymbol{R}^n \to \boldsymbol{R}$, $i = 1, \ldots, m$, are the inequality constraint funtions, the constants $b_1, \ldots, b_m$ are the limits for the constraints, and the functions $h_i : \boldsymbol{R}^n \to \boldsymbol{R}$, $i = 1, \ldots, p$ are called the equality constraint functions [59]. If a vector $\boldsymbol{x}^*$ provides the minimum objective value among all feasible vectors which satisfy the constraints, then it is an optimal solution.

Then, a convex optimization problem has the following form [59]

$$\begin{aligned} \min \quad & f_0(\boldsymbol{x}) && \text{(2.29a)} \\ \text{subject to:} \quad & f_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, m && \text{(2.29b)} \\ & a_i^T \boldsymbol{x} = b_i,\ i = 1, \ldots, p, && \text{(2.29c)} \end{aligned}$$

where $f_0, \ldots, f_m$ are convex functions. Comparing (2.29) with the general form (2.28), one can notice that the convex problem has the following requirements: 1) the objective function must be convex; 2) the inequality constraint functions must be convex; 3) the equality constraint functions $h_i(\boldsymbol{x}) = a_i^T \boldsymbol{x} - b_i$ must be affine [59]. Here, we note that a function $f_i : \boldsymbol{R}^n \to \boldsymbol{R}$ is convex if its domain $\mathbf{dom}\ f_i$ is a convex set and if for all $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbf{dom}\ f_i$ with $0 \leq \alpha \leq 1$, we have [59]

$$f_i(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}). \tag{2.30}$$

A function f_i is strictly convex if strict inequality holds in (2.30) whenever $\boldsymbol{x} \neq \boldsymbol{y}$ and $0 < \alpha < 1$. Further, a function f_i is concave if $-f_i$ is convex, and strictly concave if $-f_i$ is strictly convex. Since an affine function always holds the equality in (2.30), therefore one can note that an affine function is both convex and concave [60].

Normally, if we can formulate a practical problem as a convex optimization problem, then we can solve it efficiently. However, sometimes the formulations can be nonconvex [60]. For example, if f_0 is quasiconvex instead of convex, then the problem (2.29) becomes a quasiconvex optimization problem [61]. Here, we note that a function $f : \boldsymbol{R}^n \to \boldsymbol{R}$ is called quasiconvex if its domain and all its sublevel sets $S_\alpha = \{\boldsymbol{x} \in \mathbf{dom}\ f \mid f(\boldsymbol{x}) \leq \alpha\}$, for $\alpha \in \boldsymbol{R}$, are convex. The sublevel sets of convex functions are convex, therefore convex functions are quasiconvex. But the converse is not true. For some quasiconvex problems following specific structures, e.g., convex fractional programming [62,63], they can be transformed into equivalent convex problems and then get solved efficiently. The calculated optimal solutions can be proved to be optimal for the original quasiconvex problems [62, 63].

2.2.2 Lagrangian Dual and KKT Conditions

In this section, we briefly introduce the Lagrangian dual and the Karush-Kuhn-Tucker (KKT) condtions, which will be applied in Chapter 3 and Chapter 4 to solve the optimization problems and to derive the optimal power allocation strategies.

The basic idea in Lagrangian duality is to take the constraints in (2.28) into account by augmenting the objective function with a weighted sum of the constraint functions [59]. The Lagrangian $\mathcal{L} : \boldsymbol{R}^n \times \boldsymbol{R}^m \times \boldsymbol{R}^p \to \boldsymbol{R}$ associated with the problem (2.28) is defined as follows [59]

$$\mathcal{L}(\boldsymbol{x}, \lambda, v) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p} v_i h_i(\boldsymbol{x}). \tag{2.31}$$

Here, λ_i is the Lagrangian multiplier associated with the i^{th} inequality constraint $f_i(\boldsymbol{x}) \leq 0$, and v_i is the Lagrangian multiplier associated with the i^{th} equality constraint $h_i(\boldsymbol{x}) = 0$. The vectors λ and v are called the Lagrangian multiplier vectors associated with the optimization problem [59].

Then, the Lagrangian dual function $g : \boldsymbol{R}^m \times \boldsymbol{R}^p \to \boldsymbol{R}$ is defined as the minimum value of the Lagrangian over $\boldsymbol{x}$: for $\lambda \in \boldsymbol{R}^m$, $v \in \boldsymbol{R}^p$,

$$g(\lambda, v) = \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \lambda, v) = \inf_{\boldsymbol{x}} f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p} v_i h_i(\boldsymbol{x}). \tag{2.32}$$

The Lagrangian dual function $g(\lambda, v)$ is concave even when the orginal problem (2.28) is not convex, since the dual function is the pointwise infimum of a family of affine functions of (λ, v) [59]. For each pair (λ, v) with $\lambda \succeq 0$[14], the Lagrangian dual function provides a lower bound on the optimal value p^* of the optimization problem (2.28). Then, in order to find the best lower bound that can be obtained from the Lagrangian function, the following optimization problem needs to be solved [59]:

$$\begin{aligned} \max \quad & g(\lambda, v) && \text{(2.33a)} \\ \text{subject to:} \quad & \lambda \succeq 0. && \text{(2.33b)} \end{aligned}$$

This problem is called the Lagrangian dual problem associated with the problem (2.28). Correspondingly, the orginal problem (2.28) can be called as the primal problem. Apparently, the Lagrangian dual problem (2.33) is a convex optimization problem, since the objective function is concave and the constraint function is convex [59]. This does not depend on the convexity of the primal problem (2.28) [59].

Then, we assume that the functions $f_0, \ldots, f_m$, $h_1, \ldots, h_p$ are differentiable. From [59], we note that if a convex optimization problem with differentiable objective and constraint functions satisfies Slater's condition, then the KKT conditions provide necesseary and sufficient conditions for optimality. Hence, by assuming that f_i functions are convex and h_i functions are affine, and $\boldsymbol{x}^*$, λ^*, v^* are any points that satisfy the KKT conditions

$$f_i(\boldsymbol{x}^*) \leq 0, \quad i = 1, \ldots, m \tag{2.34a}$$

$$h_i(\boldsymbol{x}^*) = 0, \quad i = 1, \ldots, p \tag{2.34b}$$

$$\lambda_i^* \geq 0, \quad i = 1, \ldots, m \tag{2.34c}$$

$$\lambda_i^* f_i(\boldsymbol{x}^*) = 0, \quad i = 1, \ldots, m \tag{2.34d}$$

$$\nabla f_0(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(\boldsymbol{x}^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(\boldsymbol{x}^*) = 0, \tag{2.34e}$$

then $\boldsymbol{x}^*$, λ^*, v^* are primal and dual optimal, with zero duality gap [59].

2.3 Literature Review

2.3.1 Resource Allocation Towards Green Communications

According to International Telecommunication Union, the number of mobile subscriptions worldwide has dramatically increased in recent years [64]. In addition, many

[14]Here, the curled inequality symbol $\succeq$ denotes generalized inequality. For vectors, it represents componentwise inequality. For symmetric matrices, it represents matrix inequality.

new wireless applications, such as autonomous driving, smart cities, smart homes and appliances have emerged from research ideas to concrete systems [1]. The explosive growth of wireless communication applications coupled with the proliferation of mobile devices dramatically speeds up the progress of wireless networks, which results in a higher-quality human life and rapid economic growth. Meanwhile, many technical challenges still remain unsolved in wireless network designs, e.g., the need for reducing energy consumption and end-to-end latency [1].

According to [65], for every 1 TeraWatt hour (TWh) energy consumption, the information and communication technology (ICT) sector is responsible for approximately 0.75 million tons of CO_2 gas emissions. If no action is taken, the overall costs and risks of climate change, as a result of the increasing green house gases (GHG) emissions, will be equivalent to losing at least 5% of global gross domestic product (GDP) every year [66]. Nevertheless, it is also well known that ICT industry has the potential to reduce more than 23% of its current GHG emissions [66]. Interestingly, if one-third of the GHG emissions is reduced, the generated economical benefit will be higher than the required investment [67]. As an important part of ICT, wireless communication sector needs to take the responsibility to save more energy. Green communication technology, which emphasizes energy efficiency (EE) in addition to spectral efficiency (SE), has thereby been proposed as an effective solution which not only benefits communication technology sector, but also promotes economic and ecological sustainability. However, considering the compromise between network performance and energy savings, designing an efficient resource allocation strategy to limit the network energy consumption is a real challenge [68–70].

In this trend, an energy-efficient optimization problem to maximize the EE of the worst-case link was formulated and studied in [71], subject to the rate requirements, transmit power, and subcarrier assignment constraints. Price-driven algorithms for joint power and admission control are proposed to characterize the tradeoff between the total energy consumption and the system capacity in [72]. Considering the cognitive radio networks, a multi-objective optimization was formulated in [73], in which the ergodic capacity was maximized and the total transmission power of femtocell base stations was minimized. A general power consumption model in multi-user orthogonal frequency division multiple access (OFDMA) systems, including the transmission power, signal processing power, and circuit power from both the transmitter and receiver sides, was first established in [74]. Then the authors in [74] proposed a joint optimization method to iteratively find the optimal solution for the EE-maximization problem, subject to a peak transmit power constraint and a

minimum system data rate requirement. EE and SE tradeoff, based on Shannon limit, has also been extensively studied for different kinds of wireless communication networks, such as energy-constrained wireless multi-hop networks with a single source-destination pair [75], multi-user downlink OFDMA networks [76], general narrowband interference-limited systems [77] and OFDMA-based cooperative cognitive radio networks [78]. Further, the relationship between EE and SE for downlink multiuser distributed antenna systems with proportional fairness was investigated in [79]. Specifically, the EE-maximization problem was first converted into a multi-objective optimization problem (MOP), by maximizing the numerator of EE while minimizing its denominator. Then, the MOP was transformed into a single-objective optimization problem (SOP) using weighted sum method, and the optimal power value was provided by applying Lagrangian method and sub-gradient iteration approach. Considering imperfect channel estimation in an orthogonal frequency division multiplexing (OFDM) network, the inverse of EE and inverse of SE were combined into a weighted optimization problem in [80]. The problem was then transformed into a convex problem, namely, to jointly minimize the total power consumption and maximize the channel capacity, which was solved using Lagrangian method.

In the aforementioned studies [75–80], Shannon limit was utilized as the system throughput, which is mostly considered as the suitable capacity metric for communication systems with no link-layer delay QoS requirements. Nevertheless, for delay-sensitive mobile multimedia applications, such as video conferencing, autonomous driving and online gaming, provisioning QoS requirements is critical. Actually, 5G, the next generation of mobile communication technology, has been anticipated to not only offer >1 Gbps downlink data rate, but also sub-1ms end-to-end latency and 90% reduction in network energy usage [1]. This infers that the future wireless communication networks are targeted at satisfying the end-user applications' QoS requirements, while at the same time increasing EE and SE for green communications.

In order to fulfill these requirements, extensive studies in the context of power control, scheduling, and admission control have been widely provided in [10, 19, 20, 23, 24, 26, 81–87]. A cross-layer optimization framework for delay-sensitive applications over a single wireless link was formulated in [81], in which some characteristics, e.g., delay deadlines, dependencies, distortion impacts, were considered and discussed. The authors in [82] provided energy-efficient transmission techniques for a group of M packets subject to the individual packet transmission delay constraint. The above works all characterize the delay QoS requirement for a dynamic queuing system in

a deterministic way, where the delay is bounded within a certain threshold [83]. Although this sounds reasonable for real-time services, satisfying fixed QoS guarantees is especially challenging in fading communication scenarios, due to the random variations experienced in channel conditions, user mobility and changing environment [84], which could lead to settling for non-necessarily low data rates. In this direction, the delay-limited capacity, i.e., the zero-outage capacity, which is defined as the maximum rate achievable with a prescribed strict delay bound, was derived and analyzed in [88] and [89]. Since delay-limited capacity is a performance level that can be attained regardless of the values of the fading states, it can be seen as a stringent and deterministic service guarantee [84]. However, the attempt to provide a strict lower bound on delay may result in extremely conservative guarantees [10]. For example, the only lower bound that can be deterministically guaranteed in a Rayleigh fading channel is a capacity of zero [10]. In contrast to the above deterministic delay QoS bounds, in this thesis, the statistical delay QoS requirement is considered, which confines the delay bound violation probability to a required value range. In this direction, the authors in [10] introduced a link-layer capacity notion supporting statistical delay QoS requirements, which is the concept of EC.

EC, as a generalized link-level capacity notion which specifies the maximum arrival rate with a target delay-outage probability requirement, has recently received a lot of attention [10]. Specifically, EC can be regarded as the link-layer SE while the link-layer EE can be formulated as the ratio of EC to the total power expenditure. However, just like the inconsistent property of EE and SE in physical-layer channel model, the link-layer EE and EC also can be incompatible [21]. In more detail, for a point-to-point communication system operating in a flat-fading channel, the EE versus EC curve is bell shape when non-zero circuit power is considered [22]. Comparing with the physical-layer EE and SE, the link-layer EE and EC experience a much more pronounced tradeoff [22–24]. Therefore, how to allocate the system resource to efficiently balance the two conflicting metrics deserves elaborate study. Towards this direction, considering frequency flat-fading channels, an optimal power allocation strategy to maximize EC subject to an EE constraint, for delay-limited mobile multimedia applications was introduced in [23]. [24] analyzed the tradeoff between EE and EC by providing the mutually beneficial (MB) region and the contention-based (CB) region. In more detail, the MB region refers to the case when EE and EC can mutually optimize, whereas in the CB region, the trends of EE and EC conflict. However, the adjustable tradeoff between EE and EC, as well as a closed-form power allocation strategy, was not involved in [24]. The EE-EC relationship was exploited

and plotted, by expressing signal-to-noise ratio (SNR) in terms of SE using a curve fitting method in [25]. However, according to the users' diverse preferences, various application types and dynamic surrounding circumstances, a more flexible and tractable tradeoff function is preferable, which is not provided in [22–25].

Furthermore, the optimal power allocation strategy proposed in the above mentioned papers focus on the point-to-point single-channel communication systems. Note that based on the theory of Shannon limit, the total average rate of a multi-carrier system is a linear summation of each subcarrier's achievable average rate. This, however, does not apply to systems with limited statistical delay requirements. Specifically, in delay-constrained systems, the concavity and monotonicity of the EC do not remain homogeneous for single-carrier and multi-carrier systems [20]. In addition, for systems with statistical delay QoS constraints, it has been proven that the optimal power allocation strategy for single-carrier communications cannot be simply extended to the multi-carrier communications [20]. Hence, considering a single-user multi-carrier link over a frequency-selective fading channel, the delay-constrained EC maximization and EE maximization problem were separately addressed in [20] and [19], respectively. However, the link-layer EE-EC tradeoff problem for the multi-carrier communications is not investigated and analyzed in the existing literature. Especially, when a multi-user multi-carrier network is considered, the link-layer EE-EC tradeoff problem becomes more challenging. The formulated problem will be a complex combinatorial integer programming problem, rather than a convex optimization problem in [20] which can be solved using Lagrangian method. In [26] and [87], an EE optimization problem with a statistical delay provisioning and the per-user's EC requirement constraint was analyzed for a downlink multi-user OFDMA network. In these two papers, the power allocation for each subcarrier is assumed to be only related to this subcarrier's channel power gains, and not related to the same user's other subcarriers' channel power gains. Therefore, based on this assumption and the i.i.d. property of all subcarriers, the EC for a single-user multi-carrier system can be formulated as a linear summation of all subcarriers' EC values. Although this independent optimization approach is optimal in maximizing the Shannon capacity (e.g., water-filling power control for multi-carrier transmissions), it is not an optimal policy to maximize the EC-based problems for an arbitrary statistical delay provisioning [20]. In this thesis, we will not make this assumption, and aim to derive the optimal power allocation strategy for each user, which is not only across the time domain, but also across the frequency domain.

2.3.2 Current Research Progress in NOMA Networks

Due to the explosive growth of mobile data and the Internet of Things (IoT) applications which exponentially accelerate the demand for high data rates, 5G has been anticipated to offer much higher data rate, less end-to-end latency and a significant reduction in network energy usage [1]. When it comes to the proposed multiple access (MA) techniques for 5G, non-orthogonal multiple access (NOMA) has been attracting a lot of attention as a promising scheme, due to the fact that it can offer improved spectral efficiency [90], higher cell-edge throughput [91] and low transmission latency [92], over conventional orthogonal multiple access (OMA) techniques. Current available NOMA techniques can be broadly divided into two categories, i.e., power-domain and code-domain NOMA [93]. The power-domain NOMA[15] allows multiple users to simultaneously transmit using the same radio resources, either in time, frequency, or in code [93]. At the transmitter side, power-domain user-multiplexing can be enabled using superposition coding [90]. At the receiver side, multiuser separation techniques, such as successive interference cancellation (SIC), can be utilized to decode the signal [94, 95].

Current research work in NOMA-related areas mainly focuses on the topics such as cooperative design [27–29], subcarrier assignment and power control policy [30–32], physical layer security [33, 34], fairness analysis [35, 36], etc. For example, a cooperative NOMA scheme was analyzed in [27], in which the users with the stronger channel conditions were used as relays to improve the reception reliability for users with poorer connections. It was concluded that the cooperative NOMA scheme can achieve the maximum diversity gain for all the users [27]. The impact of user pairing on the performance of two different NOMA systems, i.e., NOMA with fixed power allocation (F-NOMA) and cognitive radio inspired NOMA (CR-NOMA), was studied in [96]. It was found that F-NOMA can offer a larger sum rate than OMA, and the performance gain between the two techniques can be enlarged by selecting users whose channel conditions are more distinctive [96]. In [28], the application of simultaneous wireless information and power transfer (SWIPT) to NOMA networks with randomly located users was investigated. Closed-form expressions for the outage probability and system throughput were derived to characterize the performance of the proposed user selection schemes. Further, considering a downlink NOMA transmission, an EE maximization problem was studied in [30], in which both the subcarrier assignment and the power allocation algorithms were provided for multiplexed users. The physical

[15]The power-domain NOMA will be simplified as NOMA, in the following sections.

layer secrecy issue of NOMA was discussed in [33], in which the secrecy sum rate of a single-input single-output (SISO) NOMA system consisting of a transmitter, multiple legitimate users and an eavesdropper, was maximized subject to per-user minimum data rate requirement. In [35], two different objectives, i.e., the sum rate and the minimum rate, were maximized respectively, to propose a suitable proportional fairness scheduling for a two-user NOMA system. It was shown that the proportional fairness scheduling that maximizes the minimum normalized rate can not only provide proportional fairness, but also small variation of transmission rates [35]. Furthermore, the optimal power allocation technique to maximize the user fairness in a downlink NOMA network was investigated in [36], under two different assumptions: 1) when all users' data rates are adapted to the instantaneous channel state information (CSI), and 2) when all users have fixed data rates under the average CSI.

However, all the aforementioned studies were based on Shannon limit theory, without taking into consideration the users' delay requirements. For systems with delay-sensitive applications, the physical-layer based performance analysis and power adaptive techniques may not be efficient. By applying the concept of EC and the link-layer channel model, a suboptimal power control policy was proposed in [97] to maximize the sum EC, for a two-user downlink NOMA network. However, the theoretical conclusions regarding to the advantage of NOMA over OMA on the link-layer rate performance, was not provided in [97], as well as the closed-form expressions for the individual EC in a two-user NOMA network.

Chapter 3

Single-User Single-Carrier Link-Layer EE-EC Tradeoff

3.1 Introduction

In this chapter, the delay-constrained resource allocation problem is proposed and analyzed in a single-user single-carrier communication system. A joint optimization problem of link-layer energy efficiency (EE) and effective capacity (EC) in a Nakagami-m fading channel is formulated as a normalized multi-objective optimization problem (MOP) first, under a delay-outage probability constraint and an average transmit power constraint. Then, it is transformed into a power-constrained single-objective optimization problem (SOP), by applying the weighted sum method. To solve the power-constrained SOP, the power-unconstrained SOP is considered and solved first. Firstly, the objective function is proved to be continuously differentiable and strictly quasiconvex in the optimum average input power, which confirms a cup shape curve. Then, the power-unconstrained SOP is solved by applying Charnes-Cooper transformation and Karush-Kuhn-Tucker (KKT) conditions. The proposed optimal power allocation, which includes the optimal strategy for the link-layer EE-maximization problem and the EC-maximization problem as extreme cases, is proved to be sufficient for the Pareto optimal set of the original EE-EC MOP. Finally, the power-constrained link-layer EE-EC tradeoff problem is analyzed and solved, by following the Pseudocode of the optimal power allocation algorithm in Table 3.1. To obtain more insight, the impact of different system parameters on the optimal power level is analyzed, such as the importance weight, normalization factor, scaled circuit-to-noise power ratio, and power amplifier efficiency. Simulation results confirm the analytical derivations and further show the effects of fading severeness and transmission power limit on the tradeoff performance.

The remainder of this chapter is organized as follows. In Section 3.2, the system model and a general tradeoff problem formulation are provided. The theory of link-layer EC and EE is introduced in Section 3.3. Further, the optimal power allocation strategy is derived and analyzed in this section, followed by the analysis of the impact of importance weight, normalization factor, scaled circuit-to-noise power ratio, and power amplifier efficiency on the average power level. Finally, numerical results are given in Section 3.4, followed by conclusions in Section 3.5.

3.2 System Model and Problem Formulation

3.2.1 System Model

A point-to-point wireless communication link over a Nakagami-m flat-fading channel is considered in this chapter. Different from the physical-layer channel model which has limitations in quality-of-service (QoS) support, the link-layer model depicted in Fig. 3.1(a) captures a generalized link-level capacity notion of the fading channel, under a delay QoS requirement [10], [98]. Firstly, the upper-layer packets are divided into frames at the data-link layer. Then, the source traffic and the network service are matched using a first-in-first-out (FIFO) buffer, which prevents loss of packets that could occur when the source rate is higher than the service rate, at the expense of increasing the delay [10]. At the physical layer, the frames stored at the buffer are split into bit streams. Adaptive coding and power allocation strategy are applied at the transmitter [12], using the channel-state information (CSI) fed back from the receiver, and the predetermined delay QoS requirement. The bit streams are read out of the FIFO buffer and transmitted through the wireless fading channel. Finally, the reverse operations are performed at the receiver and the frames are recovered for further processing.

The wireless channel is assumed to be block fading, i.e., the channel gain is invariant during each fading-block, but independently varies from one fading-block to another. The length of each fading-block, denoted by T_{f}, is assumed to be an integer multiple of the symbol duration T_{s}. Ideal Nyquist transmission symbol rate is also assumed to be satisfied, which means that the symbol duration $T_{\mathrm{s}} = \dfrac{1}{B}$, where B is the channel bandwidth. In addition, the service rate process using adaptive transmission, $\{R[t], t = 1, 2, \dots\}$, is considered to be stationary and ergodic [98]. The instantaneous service rate, in b/s/Hz, at the t^{th} fading-block is given by

$$R\left[t\right] = \log_2\left(1 + P_{\mathrm{t}}[t]\frac{\gamma\left[t\right]}{P_{\mathcal{L}}\sigma_{\mathrm{n}}^2}\right), \tag{3.1}$$

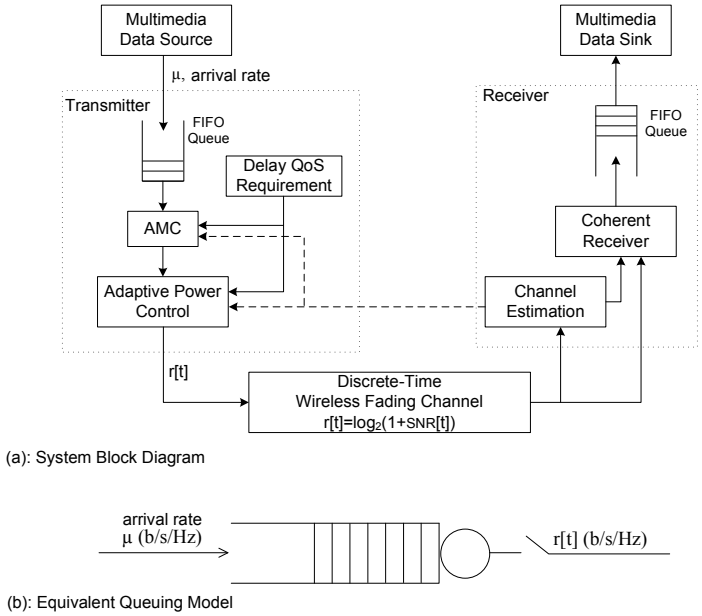


Figure 3.1: Block diagram for a point-to-point wireless communication link.

where $P_{\mathrm{t}}[t]$ denotes the transmission power, $P_{\mathcal{L}}$ shows the distance-based path-loss, and $\sigma_{\mathrm{n}}^2 = N_0 B$, with N_0 indicating the single-sided noise spectral density. In addition, $\gamma[t]$ represents the channel power gain of the unit-variance Nakagami-m block fading channel with the probability density function (PDF)[1] [99]

$$f_\gamma(\gamma) = \frac{m^m \gamma^{m-1}}{\Gamma(m)} \, \mathrm{e}^{-m\gamma},$$

where $\Gamma(z) = \int_0^\infty w^{z-1} \mathrm{e}^{-w} dw$ is the Gamma function [100]. To be specific, the Nakagami-m fading distribution is parameterized by the fading parameter m [101]. For $m = 1$, the distribution matches Rayleigh fading, whereas, for $m = \dfrac{(K+1)^2}{(2K+1)}$, the distribution is approximately Rician fading with parameter K [101]. The case of $m \to \infty$ describes the Additive White Gaussian Noise (AWGN) channels [101].

3.2.2 Problem Formulation

Note that spectral efficiency (SE) denotes the maximum achievable data rate in b/s/Hz, and EE, defined as the ratio of SE to the total power expenditure, represents the number of delivered information per energy consumed at the transmitter side, in b/J/Hz [102]. EE and SE, either in physical-layer channel model, or in link-layer channel model, conflict with each other [80] [21]. Therefore, the intention

[1]The block index t of $P_{\mathrm{t}}[t]$ and $\gamma[t]$ will be omitted for simplicity.

of simultaneously optimizing both of them, over a feasible set determined by constraint functions [103], falls into the scope of an MOP. To get rid of the different measurements and orders of magnitude of EE and SE, we normalize them with two normalization values, Ψ_{EE} and Ψ_{SE}, respectively. The normalized MOP is, hence, formulated as:

$$Q1: \quad \max_{P_{\mathrm{t}}} \ \frac{\mathrm{EE}}{\Psi_{\mathrm{EE}}} \quad \text{and} \quad \max_{P_{\mathrm{t}}} \ \frac{\mathrm{SE}}{\Psi_{\mathrm{SE}}} \tag{3.2a}$$

$$\text{subject to:} \quad \overline{P}_{\mathrm{t}} \leq P_{\max}, \tag{3.2b}$$

where $\overline{P}_{\mathrm{t}} = \mathbb{E}_{\gamma}[P_{\mathrm{t}}]$ indicates the expectation of the transmission power and $P_{\max}$ denotes the average input power limit. Here, $\mathbb{E}_{\gamma}[\cdot]$ indicates the expectation over the PDF of γ. Ψ_{EE} and Ψ_{SE} are assumed to be the EE and SE values achieved at the same normalization factor, denoted as P_{norm} satisfying $P_{\mathrm{norm}} > 0$. In more detail, $\Psi_{\mathrm{EE}} = \mathrm{EE}\,|_{\overline{P}_{\mathrm{t}}=P_{\mathrm{norm}}}$ and $\Psi_{\mathrm{SE}} = \mathrm{SE}\,|_{\overline{P}_{\mathrm{t}}=P_{\mathrm{norm}}}$.

Since EE is generally defined as the ratio of SE to the total power expenditure, the inverse of the two functions in problem $Q1$ can be minimized to make SE as the common denominator, yielding

$$Q2: \quad \min_{P_{\mathrm{t}}} \ \frac{\Psi_{\mathrm{EE}}}{\mathrm{EE}} \quad \text{and} \quad \min_{P_{\mathrm{t}}} \ \frac{\Psi_{\mathrm{SE}}}{\mathrm{SE}} \tag{3.3a}$$

$$\text{subject to:} \quad \overline{P}_{\mathrm{t}} \leq P_{\max}. \tag{3.3b}$$

Lemma 1. *The MOP, $Q2$, is equivalent to the MOP, $Q1$.*

Proof. The proof is provided in Appendix A. □

For an MOP, instead of having a single global solution, a set of points which all fit Pareto optimality is provided. To be specific, a Pareto optimal set includes solutions that cannot be improved in one objective function without deteriorating the performance in at least one of the rest of objective functions. Henceforth, Lemma 1 implies that if a point is Pareto optimal for problem $Q2$, it also belongs to the Pareto optimal set for problem $Q1$, and vice-versa.

In order to solve the MOP Q2 and to achieve the Pareto optimal solutions, one general way is to convert the MOP into an SOP, using weighted sum method [104], [105]. As such, the optimization problem Q2 can be transformed into:

$$Q3: \quad \min_{P_{\mathrm{t}}} \ w_1 \frac{\Psi_{\mathrm{EE}}}{\mathrm{EE}} + (1 - w_1) \frac{\Psi_{\mathrm{SE}}}{\mathrm{SE}} \tag{3.4a}$$

$$\text{subject to:} \quad \overline{P}_{\mathrm{t}} \leq P_{\max}, \tag{3.4b}$$

where $w_1 \in [0,1]$ is the importance weight. Specifically, w_1 and $1-w_1$ represent the relative importance of the two objective functions, EE and SE, respectively. When $w_1 = 0$, the tradeoff problem reduces to an SE-maximization problem, while when $w_1 = 1$, the problem $Q3$ is simplified into an EE-maximization problem. In other words, the importance of EE gradually grows as w_1 increases from 0 to 1.

In order to guarantee the Pareto optimal solutions for problem $Q2$, the following theorem is provided to demonstrate the relationship between the weighted optimal point and Pareto optimal solutions of the MOP $Q2$.

Theorem 5. *The unique optimal solution $\hat{P}$ of the weighted optimization problem, $\min \sum_{i=1}^{q} w_i f_i(P)$, $P \in [0, P_{\max}]$, for a given $\mathbf{w} = \{[w_i]_{1\times q} | w_i \in [0,1], \sum_{i=1}^{q} w_i = 1\}$, is Pareto optimal for the MOP, $\min f_i(P)$, $i = 1, \ldots, q$, $P \in [0, P_{\max}]$.*

Proof. The proof is provided in Appendix B. □

Implicitly, Lemma 1 and Theorem 5 illustrate that if $\hat{P}$ is a unique optimal solution for the weighted optimization problem $Q3$, it is Pareto optimal for the original MOP $Q1$.

3.3 Link-layer EE-EC tradeoff

In this section, the theory of EC and link-layer EE is introduced to incorporate the link-level delay-QoS metrics. The tradeoff performance is optimized by adaptively distributing the transmit power over time, based on the channel condition and the system delay requirement. An optimal power allocation strategy for the power-unconstrained EE-EC tradeoff problem is first developed and investigated, to pave the way for the power-constrained tradeoff problem. Further, the influence of various system parameters on the tradeoff performance is analyzed in this section.

3.3.1 Effective Capacity and Link-layer Energy Efficiency

From Chapter 2.1, we note that by assuming that the Gärtner-Ellis theorem [52, Pages 34-36] is satisfied, EC of an independent and identically distributed (i.i.d.) block fading channel can be expressed as [10]

$$E_{\mathrm{c}} = -\frac{1}{\theta T_{\mathrm{f}} B} \ln\left(\mathbb{E}\left[e^{-\theta B T_{\mathrm{f}} R[t]}\right]\right) \quad (\mathrm{b/s/Hz}), \tag{3.5}$$

where the parameter θ ($\theta > 0$) denotes the exponential decay rate of the QoS violation probability. A slower decay rate can be represented by a smaller θ, which indicates

that the system can tolerate a looser QoS guarantee, while a more stringent QoS requirement is expressed by a larger θ. In order to guarantee a required queue overflow probability limit given in (2.16), the corresponding EC, which can also be considered as the link-layer SE, represents the maximum constant arrival rate that the service process can support.

Finally, the link-layer EE for a delay-sensitive system can be defined as the ratio of EC to the sum of the transmitter's circuit power $P_{\rm c}$ and the average transmission power scaled by the power amplifier efficiency ϵ, yielding

$$\mathrm{EE} = \frac{E_{\rm c}}{P_{\rm c} + \frac{1}{\epsilon}\overline{P}_{\rm t}}, \quad 0 \leq \epsilon \leq 1. \tag{3.6}$$

3.3.2 Optimal Power Allocation

Using (3.4a)-(3.4b) and (3.6), the link-layer EE-EC tradeoff problem can be expressed as

$$Q5: \min_{P_{\rm t}} \quad w_1 \frac{\Psi_{\rm EE}\left(P_{\rm c} + \frac{1}{\epsilon}\overline{P}_{\rm t}\right)}{E_{\rm c}} + (1-w_1)\frac{\Psi_{\rm EC}}{E_{\rm c}} \tag{3.7a}$$

$$\text{subject to:} \quad \overline{P}_{\rm t} \leq P_{\max}, \tag{3.7b}$$

where $\Psi_{\rm EC}$ is the normalization value for EC, which is defined as the EC value achieved at the normalization factor, $P_{\rm norm}$, e.g., $\Psi_{\rm EC} = E_{\rm c}\,|_{\overline{P_{\rm t}}=P_{\rm norm}}$.

Replacing EC in (3.7a) with (3.5) and then inserting (3.1), the EE-EC tradeoff problem can be transformed into

$$Q6: \min_{P_{\rm r}\geq 0} \quad \frac{w_1\Psi_{\rm EE}K_\ell\left(P_{\rm c_r} + \frac{1}{\epsilon}\overline{P_{\rm r}}\right) + (1-w_1)\Psi_{\rm EC}}{-\frac{1}{\theta T_{\rm f} B}\ln\left(\mathbb{E}_\gamma\left[(1+P_{\rm r}\gamma)^{-\alpha(\theta)}\right]\right)} \tag{3.8a}$$

$$\text{subject to:} \quad \overline{P_{\rm r}} \leq \frac{P_{\max}}{K_\ell}, \tag{3.8b}$$

The scaled transmission power, $P_{\rm r} = \dfrac{P_{\rm t}}{K_\ell}$ is the optimization variable in (3.8a), which can be any nonnegative real value, i.e., $P_{\rm r} \geq 0$[2]. Further, $\overline{P_{\rm r}} = \dfrac{\overline{P}_{\rm t}}{K_\ell}$ denotes the scaled average input power, $P_{\rm c_r} = \dfrac{P_{\rm c}}{K_\ell}$ represents the circuit-to-noise power ratio, $K_\ell = P_{\mathcal{L}}\sigma_{\rm n}^2$,

[2]Since the fading coefficient is uncountable, the optimization variable $P_{\rm r}$, which is adapted to the fading coefficient, also forms an uncountable set.

and $\alpha(\theta) = \dfrac{\theta T_{\mathrm{f}} B}{\ln 2}$. After deleting the negative constant, $-\dfrac{1}{\theta T_{\mathrm{f}} B}$, the minimization problem (3.8a) reduces to a maximization problem. Then, by inverting the objective function, it can be converted back into a minimization problem, yielding[3]

$$Q7: \min_{P_{\mathrm{r}} \geq 0} \quad \frac{\ln\left(\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}\gamma)^{-\alpha(\theta)}\right]\right)}{w_1 \Psi_{\mathrm{EE_r}}\left(P_{\mathrm{c_r}} + \dfrac{1}{\epsilon}\overline{P_{\mathrm{r}}}\right) + (1-w_1)\,\Psi_{\mathrm{EC}}} \tag{3.9a}$$

$$\text{subject to:} \ \ \overline{P_{\mathrm{r}}} \leq \frac{P_{\max}}{K_{\ell}}, \tag{3.9b}$$

where $\Psi_{\mathrm{EE_r}} = \Psi_{\mathrm{EE}} K_{\ell}$.

3.3.2.1 Optimum Power Allocation with No Input Power Constraint

In this section, the unconstrained SOP is tackled to pave the way for the optimal power allocation algorithm of the power-constrained SOP. To fully understand the unconstrained SOP, we start by investigating the properties of this case with a predetermined importance weight w_1, which are summarized in the following theorem.

Theorem 6. *Define U7 as the objective function of the tradeoff problem Q7, i.e., the minimization function in* (3.9a)*. For a predetermined importance weight, U7 has the following properties:*

1. *U7 is continuously differentiable and strictly quasiconvex in* $\overline{P_{\mathrm{r}}}$*,*
2. *U7 first decreases and then increases with* $\overline{P_{\mathrm{r}}}$*, which turns out to be a cup shape curve,*
3. $$U7' \begin{cases} > 0 & \text{if} \quad U7 < \dfrac{\epsilon}{w_1 \Psi_{\mathrm{EE_r}}} f(\overline{P_{\mathrm{r}}})' \\ = 0 & \text{if} \quad U7 = \dfrac{\epsilon}{w_1 \Psi_{\mathrm{EE_r}}} f(\overline{P_{\mathrm{r}}})', \\ < 0 & \text{if} \quad U7 > \dfrac{\epsilon}{w_1 \Psi_{\mathrm{EE_r}}} f(\overline{P_{\mathrm{r}}})' \end{cases}$$

where $f(\overline{P_{\mathrm{r}}}) = \ln\left(\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}\gamma)^{-\alpha(\theta)}\right]\right)$, $U7' = \dfrac{dU7}{d\overline{P_{\mathrm{r}}}}$, *and* $f(\overline{P_{\mathrm{r}}})' = \dfrac{df(\overline{P_{\mathrm{r}}})}{d\overline{P_{\mathrm{r}}}}$.

Proof. The proof is provided in Appendix C. □

[3]The objective function in problem $Q7$ is similar to equation (4) developed in [98]. The difference is the second addend and the introduced adjustable parameters in the denominator of (3.9a).

In Theorem 6, Property 1) reveals the differentiability of (3.9a) and guarantees the existence and uniqueness of the global minimum, for a predetermined weight value. Property 2) indicates that the global optimum is always achieved at a finite power value. From Property 2) and Property 3), one can notice that when $\overline{P_{\mathrm{r}}} \to 0$, $U7' < 0$, which means now $U7 > \dfrac{\epsilon}{w_1 \Psi_{\mathrm{EE_r}}} f(\overline{P_{\mathrm{r}}})'$. With $\overline{P_{\mathrm{r}}}$ increasing, $U7$ gradually decreases until it equals to $\dfrac{\epsilon}{w_1 \Psi_{\mathrm{EE_r}}} f(\overline{P_{\mathrm{r}}})'$. After that point, $U7$ starts to increase with $\overline{P_{\mathrm{r}}}$. Further, Property 3) connects the sign of the first derivative with the relative difference of $U7$ and the scaled first derivative of $f(\overline{P_{\mathrm{r}}})$.

To solve the unconstrained SOP, we apply the Charnes-Cooper transformation.

Lemma 2. *A ratio problem* $(P): \min\limits_{x \in S} \quad \dfrac{f(x)}{g(x)}$, *where* f *is convex and* g *is affine and positive,* $f, g : S \to R$, $S \subseteq R^n$, *can be transformed into a convex program*

$$\begin{aligned} (P'): \min_{y/\phi \in S} \quad & \phi f(y/\phi) \\ \textit{subject to:} \quad & \phi g(y/\phi) = 1, \end{aligned}$$

by using the Charnes-Cooper transformation $y = \dfrac{1}{g(x)} x$, $\phi = \dfrac{1}{g(x)}$, *where* $\phi > 0$.

Proof. The proof is provided in Appendix D. □

According to Lemma 2, the power-unconstrained minimization problem (3.9a) reduces to the following equivalent problem $Q8$, by applying the Charnes-Cooper transformation and one further step of substitution[4].

$$Q8: \quad \min_{P_{\mathrm{r}} \geq 0} \quad \phi \ln \left(\mathbb{E}_\gamma \left[(1 + P_{\mathrm{r}} \gamma)^{-\alpha(\theta)} \right] \right) \tag{3.12a}$$

$$\text{subject to:} \quad \phi \left(w_1 \Psi_{\mathrm{EE_r}} \left(P_{\mathrm{c_r}} + \frac{1}{\epsilon} \overline{P_{\mathrm{r}}} \right) + (1 - w_1) \Psi_{\mathrm{EC}} \right) = 1. \tag{3.12b}$$

Note that problem $Q8$ is not jointly convex in P_{r} and ϕ. But, by regarding ϕ as a parameter, problem $Q8$ becomes a convex program in P_{r}, since the objective function is convex [98] and the constraint is an affine function in P_{r}. The KKT conditions are, hence, sufficient and necessary for the optimal solution. Set $\lambda \in \mathbb{R}_+, \mathbb{R}_+ \equiv [0, \infty]$ as the Lagrange multiplier, the Lagrangian function can be expressed as

$$\begin{aligned} &\mathcal{L}(P_{\mathrm{r}}, \lambda) \\ &= \phi \ln \left(\mathbb{E}_\gamma \left[(1 + P_{\mathrm{r}} \gamma)^{-\alpha(\theta)} \right] \right) + \lambda \left(\phi \left(w_1 \Psi_{\mathrm{EE_r}} \left(P_{\mathrm{c_r}} + \frac{1}{\epsilon} \overline{P_{\mathrm{r}}} \right) + (1 - w_1) \Psi_{\mathrm{EC}} \right) - 1 \right). \end{aligned}$$

[4]The Charnes-Cooper transformation is first utilized to achieve the convex program (P'), then problem $Q8$ is derived by substituting $x = \dfrac{y}{\phi}$ in problem (P').

The KKT condition $\dfrac{\partial \mathcal{L}(P_{\mathrm{r}},\lambda)}{\partial P_{\mathrm{r}}}=0$ can be expanded as

$$\alpha(\theta)\int_0^{\infty}(1+P_{\mathrm{r}}\gamma)^{-\alpha(\theta)-1}\,\gamma f(\gamma)\mathrm{d}\gamma=\frac{\lambda w_1\Psi_{\mathrm{EE_r}}}{\epsilon}\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}\gamma)^{-\alpha(\theta)}\right]\int_0^{\infty}f(\gamma)\mathrm{d}\gamma.$$

Finally, it can be expressed as

$$\alpha(\theta)\gamma\,(1+P_{\mathrm{r}}^{*}\gamma)^{-\alpha(\theta)-1}=\frac{\lambda w_1\Psi_{\mathrm{EE_r}}}{\epsilon}\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}^{*}\gamma)^{-\alpha(\theta)}\right], \tag{3.13a}$$

and the optimum power distribution scheme can be found as

$$P_{\mathrm{r}}^{*}=\left[\frac{\alpha(\theta)^{\frac{1}{1+\alpha(\theta)}}}{(w_1\nu)^{\frac{1}{1+\alpha(\theta)}}\,\gamma^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}-\frac{1}{\gamma}\right]^{+}, \tag{3.14}$$

where $\nu=\dfrac{\lambda\Psi_{\mathrm{EE_r}}}{\epsilon}\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}^{*}\gamma)^{-\alpha(\theta)}\right]$ is referred to as the scaled-Lagrangian-multiplier and $[x]^{+}=\max\{0,x\}$.

Now the optimal value of ϕ can be found. Since all unknowns have been expressed as explicit functions of ν, this reduces to finding ν^{*} from the following equation

$$\nabla_{\phi}\mathcal{L}=\ln\left(\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}^{*}\gamma)^{-\alpha(\theta)}\right]\right)+\lambda\left(w_1\Psi_{\mathrm{EE_r}}\left(P_{\mathrm{c_r}}+\frac{1}{\epsilon}\overline{P_{\mathrm{r}}^{*}}\right)+(1-w_1)\Psi_{\mathrm{EC}}\right)=0. \tag{3.15}$$

By substituting the power allocation (3.14) into (3.15), the optimal value for ν (referred to as ν^{*}) can be easily found using the following equation

$$\begin{aligned}&\Psi_{\mathrm{EE_r}}\mathbb{E}_{\gamma}\left[\left(1+\left[\frac{(\gamma\alpha(\theta))^{\frac{1}{1+\alpha(\theta)}}}{(w_1\nu^{*})^{\frac{1}{1+\alpha(\theta)}}}-1\right]^{+}\right)^{-\alpha(\theta)}\right]\ln\left(\mathbb{E}_{\gamma}\left[\left(1+\left[\frac{(\gamma\alpha(\theta))^{\frac{1}{1+\alpha(\theta)}}}{(w_1\nu^{*})^{\frac{1}{1+\alpha(\theta)}}}-1\right]^{+}\right)^{-\alpha(\theta)}\right]\right)\\&+\epsilon\nu^{*}\left(w_1\Psi_{\mathrm{EE_r}}\left(P_{\mathrm{c_r}}+\frac{1}{\epsilon}\mathbb{E}_{\gamma}\left[\frac{\alpha(\theta)^{\frac{1}{1+\alpha(\theta)}}}{(w_1\nu^{*})^{\frac{1}{1+\alpha(\theta)}}\,\gamma^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}-\frac{1}{\gamma}\right]^{+}\right)+(1-w_1)\Psi_{\mathrm{EC}}\right)=0.\end{aligned} \tag{3.16}$$

Lemma 3. *For the Nakagami-m fading channel, the closed-form expressions of* $\overline{P_{\mathrm{r}}^{*}}$ *and* $\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}^{*}\gamma)^{-\alpha(\theta)}\right]$ *are given in* (3.18a) *and* (3.18b)*, wherein* $\Gamma(a,x)$ *is the upper incomplete gamma function, i.e.,* $\Gamma(a,x)=\displaystyle\int_x^{\infty}z^{a-1}\mathrm{e}^{-z}\,dz$*, and* $E_1(x)=\displaystyle\int_x^{\infty}\frac{\mathrm{e}^{-z}}{z}\,dz$ *indicates the exponential integral [100]*[5]*.*

[5]It is assumed that the path of integration excludes the origin and does not cross the negative real axis [100].

$$\overline{P_{\mathrm{r}}^{*}} = \begin{cases} \frac{\left(\frac{\alpha(\theta)}{w_1\nu^*}\right)^{\frac{1}{1+\alpha(\theta)}} m^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma(m)\left(m-\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)}\left[-\left(\frac{w_1\nu^* m}{\alpha(\theta)}\right)^{\left(m-\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)} \mathrm{e}^{-\frac{w_1\nu^* m}{\alpha(\theta)}} + \Gamma\left(m+\frac{1}{1+\alpha(\theta)}, \frac{w_1\nu^* m}{\alpha(\theta)}\right)\right] \\ -\frac{m}{\Gamma(m)(m-1)}\left[-\left(\frac{w_1\nu^* m}{\alpha(\theta)}\right)^{m-1} \mathrm{e}^{-\frac{w_1\nu^* m}{\alpha(\theta)}} + \Gamma\left(m, \frac{w_1\nu^* m}{\alpha(\theta)}\right)\right], \text{when } m \neq 1, m \neq \dfrac{\alpha(\theta)}{\alpha(\theta)+1}, \\ \left(\frac{\alpha(\theta)}{w_1\nu^*}\right)^{\frac{1}{1+\alpha(\theta)}} \Gamma\left(\frac{1}{1+\alpha(\theta)}, \frac{w_1\nu^*}{\alpha(\theta)}\right) - \mathrm{E}_1\left(\frac{w_1\nu^*}{\alpha(\theta)}\right), \qquad \text{when } m = 1, \\ \frac{\left(\frac{\alpha(\theta)}{w_1\nu^*}\right)^{\frac{1}{\alpha(\theta)+1}}\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)^{\frac{\alpha(\theta)}{\alpha(\theta)+1}}}{\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)} \mathrm{E}_1\left(\frac{w_1\nu^*}{1+\alpha(\theta)}\right) + \frac{\alpha(\theta)}{\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)}\Bigg[-\mathrm{e}^{-\frac{w_1\nu^*}{\alpha(\theta)+1}}\left(\frac{w_1\nu^*}{\alpha(\theta)+1}\right)^{-\frac{1}{\alpha(\theta)+1}} \\ +\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}, \frac{w_1\nu^*}{\alpha(\theta)+1}\right)\Bigg], \qquad \text{when } m = \dfrac{\alpha(\theta)}{\alpha(\theta)+1}, \end{cases} \tag{3.18a}$$

$$\mathbb{E}_{\gamma}\left[\left(1+P_{\mathrm{r}}^{*}\gamma\right)^{-\alpha(\theta)}\right] = \begin{cases} \left(\frac{w_1\nu^*}{\alpha(\theta)}\right)^{\frac{\alpha(\theta)}{1+\alpha(\theta)}} \frac{m^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma(m)\left(m-\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)}\Bigg[-\left(\frac{w_1\nu^* m}{\alpha(\theta)}\right)^{m-\frac{\alpha(\theta)}{1+\alpha(\theta)}} \mathrm{e}^{-\frac{w_1\nu^* m}{\alpha(\theta)}} \\ +\Gamma\left(m+\frac{1}{1+\alpha(\theta)}, \frac{w_1\nu^* m}{\alpha(\theta)}\right)\Bigg] + 1 - \frac{\Gamma\left(m, \frac{w_1\nu^* m}{\alpha(\theta)}\right)}{\Gamma(m)}, \quad \text{when } m \neq \dfrac{\alpha(\theta)}{\alpha(\theta)+1}, \\ \frac{\left(\frac{w_1\nu^*}{\alpha(\theta)+1}\right)^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)} \mathrm{E}_1\left(\frac{w_1\nu^*}{\alpha(\theta)+1}\right) + 1 - \frac{\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}, \frac{w_1\nu^*}{\alpha(\theta)+1}\right)}{\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)}, \text{ when } m = \dfrac{\alpha(\theta)}{\alpha(\theta)+1}. \end{cases} \tag{3.18b}$$

Proof. The proof is provided in Appendix E. □

After inserting the closed-form expressions (3.18a)-(3.18b) into (3.16), the optimal value for ν, i.e., ν^*, can be solved from (3.16) using root-finding functions, e.g., fzero in Matlab. The optimal input power level $\overline{P_{\mathrm{t}}^{*}}$ can then be found by inserting ν^* into (3.18a), namely

$$\overline{P_{\mathrm{t}}^{*}} = K_{\ell} \times \overline{P_{\mathrm{r}}^{*}}\,|_{\nu=\nu^*}\,. \tag{3.17}$$

Since the channel is assumed to be stationary and ergodic, henceforth, its average will not be affected by the shift in the time origin. Further, the pointwise mapping between P_{r} and γ is fixed for each fading realization and is determined by the power allocation policy that depends on $\overline{P_{\mathrm{r}}}$.

The above equations conclude the power-unconstrained EE-EC tradeoff solution. Now we provide the following analysis to pave the way for the power-constrained EE-EC tradeoff problem, that is presented in next section. Let us assume the optimal average power $\overline{P_{\mathrm{t}}^{*}}$ which solves the power-unconstrained tradeoff problem is

found. Then, the power-unconstrained EE-EC tradeoff problem simplifies into an EC-maximization problem with an input power constraint, yielding

$$\max_{P_{\mathrm{r}} \geq 0} \quad -\frac{1}{\theta T_{\mathrm{f}} B} \ln\left(\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}\gamma)^{-\alpha(\theta)}\right]\right) \tag{3.19a}$$

$$\text{subject to:} \quad \overline{P_{\mathrm{r}}} \leq \frac{\overline{P_{\mathrm{t}}^{*}}}{K_{\ell}}. \tag{3.19b}$$

3.3.2.2 Optimal Power Allocation under Average Input Power Constraint

In this section, the optimization problem (3.9a)-(3.9b) can be solved using the results from Subsection 3.3.2.1. After the unique optimum average power value $\overline{P_{\mathrm{t}}^{*}}$ for the power-unconstrained problem is calculated, it needs to be compared with the input average power limit $P_{\max}$. If $\overline{P_{\mathrm{t}}^{*}} \leq P_{\max}$, it means that the system has enough power to support the optimal tradeoff performance found in Subsection 3.3.2.1. Otherwise, $\overline{P_{\mathrm{t}}^{*}} \geq P_{\max}$ means that $P_{\max}$ is too small to support the power allocation strategy (3.14)-(3.18b) and the system has to operate at the maximum available power $P_{\max}$. Therefore, the operational input average power value becomes $\min(\overline{P_{\mathrm{t}}^{*}}, P_{\max})$.

Hence, the power-constrained EE-EC tradeoff problem in (3.9a)-(3.9b) simplifies to an EC-maximization problem with two input power constraints, yielding

$$\max_{P_{\mathrm{r}} \geq 0} \quad -\frac{1}{\theta T_{\mathrm{f}} B} \ln\left(\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}\gamma)^{-\alpha(\theta)}\right]\right) \tag{3.20a}$$

$$\text{subject to:} \quad \overline{P_{\mathrm{r}}} \leq \frac{\overline{P_{\mathrm{t}}^{*}}}{K_{\ell}}, \tag{3.20b}$$

$$\overline{P_{\mathrm{r}}} \leq \frac{P_{\max}}{K_{\ell}}. \tag{3.20c}$$

The optimal power allocation to solve (3.9a)-(3.9b) is according to (3.14), wherein, the optimal ν^{*} is found such that $K_{\ell}\overline{P_{\mathrm{r}}} \mid_{\nu=\nu^{*}} = \min(\overline{P_{\mathrm{t}}^{*}}, P_{\max})$. To summarize, the Pseudocode of the optimal power allocation algorithm to solve (3.9a)-(3.9b) is illustrated in Table 3.1.

Furthermore, the optimal power allocation strategy (3.14)-(3.18b) has the following properties:

Properties 1. *1. For every given weight value, the proposed optimal solution* (3.14)-(3.18b) *is sufficient for the Pareto optimal set of the original EE-EC MOP* $Q1$.

2. The proposed optimal solution includes the optimal power allocation strategy for the link-layer EE-maximization problem (when $w_1 = 1$*) and also the one for the EC-maximization problem (when* $w_1 = 0$*), as extreme cases.*

Table 3.1: Optimal Power Allocation Algorithm for a Single-User Single-Carrier System

Input: Initialization parameters

w_1	importance weight of EE
P_{norm}	normalization factor
Ψ_{EE}	normalization value of EE, e.g., $\Psi_{\text{EE}} = \text{EE}\,\vert_{\overline{P_{\text{t}}}=P_{\text{norm}}}$
Ψ_{EC}	normalization value of EC, e.g., $\Psi_{\text{EC}} = E_{\text{c}}\,\vert_{\overline{P_{\text{t}}}=P_{\text{norm}}}$
θ	exponential decay rate of the QoS violation probability
$P_{\max}$	input average power limit
P_{c}	circuit power
m	Nakagami fading parameter
ϵ	power amplifier efficiency
K_ℓ	pathloss and noise factor, e.g., $K_\ell = P_{\mathcal{L}}\sigma_{\text{n}}^2$
T_{f}	fading block duration
B	channel bandwidth

Step 1:

Create (3.16), using closed-form expressions given in (3.18a) and (3.18b).
Find ν^* which solves (3.16) using root-finding functions, e.g., fzero in Matlab.
Insert ν^* in (3.14) to calculate P_{r}^* and then get P_{t}^*, using $P_{\text{t}}^* = K_\ell \times P_{\text{r}}^*$.
Insert ν^* in (3.18a) to calculate $\overline{P_{\text{r}}^*}$ and then get $\overline{P_{\text{t}}^*}$, using $\overline{P_{\text{t}}^*} = K_\ell \times \overline{P_{\text{r}}^*}$.

Step 2:

If $P_{\max} > \overline{P_{\text{t}}^*}$

Calculate E_{c} using (3.5) and EE using (3.6), by applying P_{t}^* and $\overline{P_{\text{t}}^*}$.

Else

Create $\overline{P_{\text{t}}^*} = P_{\max}$ and use $\overline{P_{\text{r}}^*} = \dfrac{P_{\max}}{K_\ell}$ to update ν^* by solving (3.18a).
Insert ν^* in (3.14) to calculate P_{r}^* and then get P_{t}^*, using $P_{\text{t}}^* = K_\ell \times P_{\text{r}}^*$.
Calculate E_{c} using (3.5) and EE using (3.6), by applying P_{t}^* and $\overline{P_{\text{t}}^*}$.

End

Output: $\left[P_{\text{t}}^*, \overline{P_{\text{t}}^*}, E_{\text{c}}, \text{EE}\right]$

3. *When $\theta \to 0$, EC is equivalent to the ergodic capacity. For the weighted physical-layer EE-SE tradeoff problem, the optimum power allocation strategy is the traditional water-filling approach, with the water level to be chosen so that the maximum tradeoff performance can be achieved [98].*

4. *When $\theta \to \infty$, EC is equivalent to the zero-outage capacity, and the optimum power allocation strategy is to maintain a constant received signal-to-noise ratio (SNR), at a level that maximizes the tradeoff performance [106].*

In more detail, we first note that with a predetermined importance weight, the unique optimal solution of $Q8$ is sufficient for the optimal solution of the weighted

tradeoff problem $Q7$ [62, 63]. Then, by applying Lemma 1, Theorem 5 and Theorem 6, one can show that the optimal power allocation strategy (3.14)-(3.18b) for every determined weight value, is sufficient for the Pareto optimal set of the original EE-EC MOP $Q1$.

Furthermore, the optimal solution (3.14)-(3.18b) is similar to the optimal power allocation strategy of the link-layer EE-maximization problem in [98], with a different value of the optimal scaled-Lagrangian-multiplier ν^*. Specifically, when $w_1 = 1$, the proposed optimal solution (3.14) reduces to the one developed in [98]. It means that the optimal solution in [98] is a special case of the optimal power allocation strategy for the weighted EE-EC tradeoff problem in this chapter. Specifically, in [98], the optimal operational average power equals to $\min(P^*_{\mathrm{EE}}, P_{\max})$, while the optimal average power varies between $[P^*_{\mathrm{EE}}, P_{\max}]$, for a typical EE-EC tradeoff problem.

When $\theta \to 0$, by following similar steps, the optimal power allocation strategy for the weighted tradeoff problem can be derived as

$$P_{\mathrm{r}} = \left(\frac{1}{\rho} - \frac{1}{\gamma}\right)^{+}, \tag{3.21}$$

which is the well-known water-filling approach and ρ can be found from the KKT condition

$$\mathbb{E}_{\gamma}\left[\left(\ln\left(\frac{\gamma}{\rho}\right)\right)^{+}\right] - \rho\left(\left(\epsilon P_{\mathrm{c_r}} + \mathbb{E}_{\gamma}\left[\left(\frac{1}{\rho} - \frac{1}{\gamma}\right)^{+}\right]\right) + \frac{\epsilon\left(1 - w_1\right)\Psi_{\mathrm{EC}}}{w_1 \Psi_{\mathrm{EE_r}}}\right) = 0. \tag{3.22}$$

When $\theta \to \infty$, a system with extremely stringent delay requirement is considered, which means in this case, EC is equivalent to the zero-outage capacity [98].

3.3.3 The Impact of w_1, $P_{\mathbf{norm}}$, $P_{\mathrm{c_r}}$ and ϵ on the EE-EC Tradeoff

From (3.14)-(3.18b), it is noted that the calculated tradeoff optimal power value can be influenced by four factors, which are the importance weight w_1, normalization factor P_{norm}, scaled circuit-to-noise power ratio $P_{\mathrm{c_r}}$, and power amplifier efficiency ϵ. In order to thoroughly understand the effects of these factors on the tradeoff performance, we provide the following lemmas.

Lemma 4. *The optimal average power value $\overline{P^*_{\mathrm{t}}}$ monotonically decreases with w_1, but strictly increases with P_{norm}.*

Proof. The proof is provided in Appendix F. □

To understand Lemma 4, we note that Ψ_{EE} and Ψ_{EC} can not only function as the normalization values, but also can be regarded as two weights. Hence, the complete weights of EE and EC can be viewed as $w_1\Psi_{\mathrm{EE}}$ and $(1-w_1)\Psi_{\mathrm{EC}}$, respectively. In order to compare the relative importance of the two objective functions, the complete weights need to be compared, yielding

$$\frac{\mathrm{W}_{\mathrm{EE}}}{\mathrm{W}_{\mathrm{EC}}} = \frac{w_1\Psi_{\mathrm{EE}}}{(1-w_1)\Psi_{\mathrm{EC}}} = \frac{1}{K_\ell\left(\frac{1}{w_1}-1\right)\left(P_{\mathrm{c_r}}+\frac{1}{\epsilon}P_{\mathrm{norm}}\right)}, \tag{3.23}$$

where $\mathrm{W}_{\mathrm{EE}} = w_1\Psi_{\mathrm{EE}}$ and $\mathrm{W}_{\mathrm{EC}} = (1-w_1)\Psi_{\mathrm{EC}}$ denote the complete weights of EE and EC, respectively. From (3.23), one can note that $\mathrm{W}_{\mathrm{EE}}/\mathrm{W}_{\mathrm{EC}}$ increases with w_1, which means that with the increase of w_1, the importance of EC drops. Hence, when w_1 increases, the system prefers to sacrifice more EC to achieve a better EE. Therefore, the optimum average transmit power $\overline{P_{\mathrm{t}}^*}$ will be shifted from $P_{\max}$-side to P_{EE}^*-side. On the other hand, (3.23) indicates that when P_{norm} grows, the ratio of W_{EE} to W_{EC} decreases. This means that with the increase of P_{norm}, the system prefers to improve EC, with certain deteriorations of EE. Therefore, following the same trend with EC, $\overline{P_{\mathrm{t}}^*}$ will increase with P_{norm}.

Lemma 4 provides a proper guideline for users to design a more flexible and favorable system, based on diverse preferences and different system requirements. For example, if the system prefers a better EC, a larger P_{norm} as well as a smaller w_1 should be chosen to offer a larger optimal transmit power, and in turn, a relatively larger EC. In contrast, if a user prefers more EE, a smaller P_{norm} as well as a larger w_1 will be more beneficial.

To investigate the effects of the scaled circuit-to-noise power ratio $P_{\mathrm{c_r}}$ and the power amplifier efficiency ϵ, we introduce the following lemma.

Lemma 5. *The average optimal power $\overline{P_{\mathrm{t}}^*}$ monotonically increases with the scaled circuit-to-noise power ratio $P_{\mathrm{c_r}}$, as well as the power amplifier efficiency ϵ.*

Proof. The proof is omitted here due to the page limit, but can be found by following similar steps as in Appendix F. □

3.4 Numerical Results

In this section, the impact of the normalization factor P_{norm}, fading severness parameter m, scaled circuit-to-noise power ratio $P_{\mathrm{c_r}}$, importance weight w_1, and transmission power constraint on the link-layer EE-EC tradeoff problem is numerically investigated

for a flat block-fading channel with delay-outage probability constraints. In the following figures, it is assumed that the fading-block duration $T_{\mathrm{f}} = 2$ms, bandwidth $B = 250$kHz, input average power limit $P_{\max} = 10$dB, power amplifier efficiency $\epsilon = 0.5$, fading parameter $m = 1$, and the QoS exponent $\theta = 10^{-2}$, unless otherwise indicated.

Fig. 3.2 includes the plots for EC (on the left-hand-side (LHS) y-Axis, in solid lines with markers) and EE[6] (on the right-hand-side (RHS) y-Axis, markers only) versus the importance weight w_1, for various scaled circuit-to-noise power ratio values, with normalization factor $P_{\mathrm{norm}} = 0.5P_{\mathrm{EE}}^{*}$[7]. This figure reveals that, when $w_1 \in [0.18, 1]$, the link-layer EE increases whereas EC gradually decreases with w_1. This happens because the increase of w_1 raises the importance of EE and diminishes the priority of EC, which confirms our design intention. Moreover, when $P_{\mathrm{c_r}} = 5$dB, there is a flat region, i.e., $w_1 \in [0, 0.18]$, wherein EE and EC remain constant. It happens because, in this region, the calculated optimum average power $\overline{P_{\mathrm{t}}^{*}}$ is larger than $P_{\max}$. Since the power-constrained tradeoff system performs at $\min(\overline{P_{\mathrm{t}}^{*}}, P_{\max})$, which, in this case, equals to $P_{\max}$, hence, the final operational average power is a fixed constant. Therefore, the constant EE and EC will be observed in this region. Furthermore, although the flat region exists for both settings of $P_{\mathrm{c_r}}$, the one obtained when $P_{\mathrm{c_r}} = 5$dB is larger than the case when $P_{\mathrm{c_r}} = -5$dB. In Section 3.3.3, it has been proved that when $P_{\mathrm{c_r}}$ increases, the optimum average input power $\overline{P_{\mathrm{t}}^{*}}$ increases, which means that $\overline{P_{\mathrm{t}}^{*}}$ will remain larger than $P_{\max}$ and EC will stabilize at its maximum value for a longer period of w_1. In addition, Fig. 3.2 also demonstrates that, with fixed w_1, when $P_{\mathrm{c_r}}$ increases from -5dB to 5dB, the value of EE decreases. This is due to the fact that EE varies inversely with $\overline{P_{\mathrm{t}}^{*}}$, while the optimum average power $\overline{P_{\mathrm{t}}^{*}}$ increases monotonically with $P_{\mathrm{c_r}}$. Therefore, EE decreases with the circuit-to-noise power ratio $P_{\mathrm{c_r}}$.

The results of EC and EE versus $P_{\max}$, for various values of w_1, with $P_{\mathrm{c_r}} = 0$dB and $P_{\mathrm{norm}} = P_{\mathrm{EE}}^{*}$, are plotted in Fig. 3.3 and Fig. 3.4, respectively. From Fig. 3.3, it shows that when $w_1 = 0.5$ and $w_1 = 1$, EC first continuously increases, and then it remains stable, after a break-point. This is because, for the weighted tradeoff problem with $w_1 = 0.5$ or $w_1 = 1$, the operational average power limit is settled at $\min(\overline{P_{\mathrm{t}}^{*}}, P_{\max})$. Specifically, when $P_{\max} \leq \overline{P_{\mathrm{t}}^{*}}$, the system operates at $P_{\max}$, whereas when $P_{\max} \geq \overline{P_{\mathrm{t}}^{*}}$, the tradeoff system will not consume all the available power, but

[6]In this section, all EE figures show the scaled EE values calculated with the scaled circuit-to-noise power ratio and the scaled average transmission power. In other words, the actual EE values equal to $1/K_{\ell}$ multiplied with the values on the y-Axis.

[7]Here P_{EE}^{*} is the optimum average power value for the EE-maximization problem.

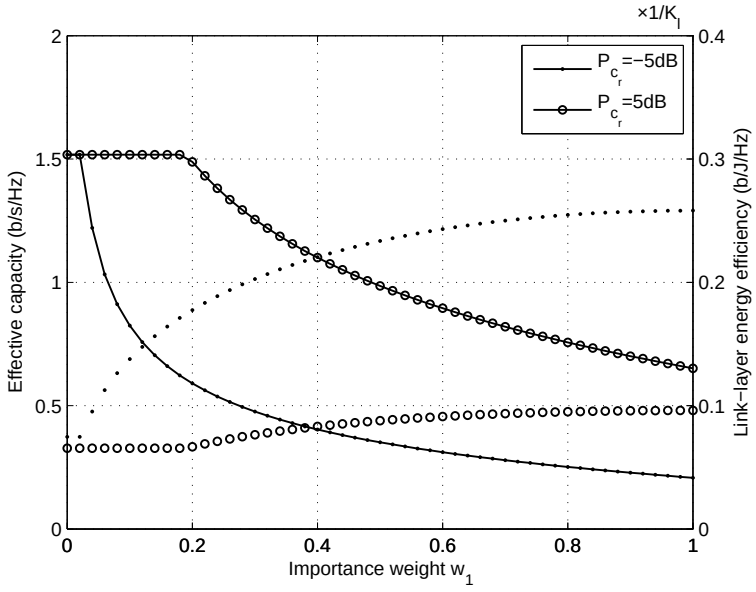


Figure 3.2: EC and link-layer EE versus importance weight w_1 for various values of $P_{\mathrm{c_r}}$ in Rayleigh fading channels.

rather operates at $\overline{P_{\mathrm{t}}^*}$, which leads to a constant EC. These observations, however, do not apply to the case of $w_1 = 0$, which represents the EC-maximization problem. In this case, EC continuously increases with $\dfrac{P_{\max}}{K_\ell}$ while EE, shown in Fig. 3.4, decreases after reaching its peak value. This is due to the fact that the allocation strategy for the EC-maximization problem consumes the whole available input power, resulting in continuously growing EC, and simultaneously losing EE, after its maximum value.

Similarly, from Fig. 3.4, one can find that when $w_1 = 0.5$, EE first increases until it reaches its peak value, at which $P_{\max} = P_{\mathrm{EE}}^*$, then EE gradually decreases until $\overline{P_{\mathrm{t}}} = \overline{P_{\mathrm{t}}^*}$, after which it stabilizes. This demonstrates that the operational optimal average power, $\min(\overline{P_{\mathrm{t}}^*}, P_{\max})$, is always achieved between $[P_{\mathrm{EE}}^*, P_{\max}]$. And, for any $P_{\max} \geq \overline{P_{\mathrm{t}}^*}$, the tradeoff system performs at $\overline{P_{\mathrm{t}}^*}$, which leads to a constant EE in Fig. 3.4. In addition, when $w_1 = 1$, which indicates the EE-maximization problem, Fig. 3.4 shows that the link-layer EE gradually increases until its peak value, achieved at P_{EE}^*, after which it remains constant. This is due to the fact that the average optimal power limit for the EE-maximization problem is always achieved at $\min(P_{\mathrm{EE}}^*, P_{\max})$ [98], which means that when $P_{\max} \leq P_{\mathrm{EE}}^*$, the system operates at the most achievable power value $P_{\max}$, and when $P_{\max} \geq P_{\mathrm{EE}}^*$, it performs at the global optimal power level P_{EE}^*. Although Fig. 3.3 and Fig. 3.4 are plotted using the link-layer rate, the same trend can be observed in physical-layer tradeoff problem.

The plots for $\overline{P_{\mathrm{r}}^*}$ versus w_1 for various fading parameters, with $P_{\mathrm{norm}} = 0.5P_{\mathrm{EE}}^*$

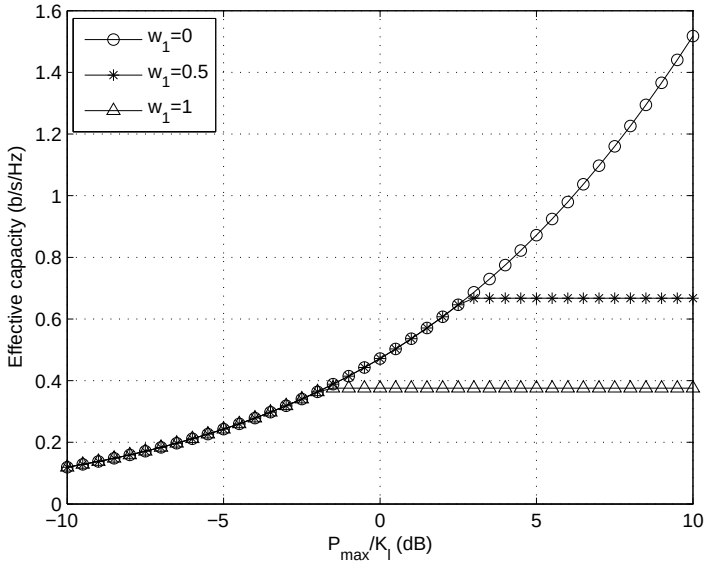


Figure 3.3: EC versus scaled average input power limit $\frac{P_{\max}}{K_\ell}$ for various values of importance weight w_1 in Rayleigh fading channels.

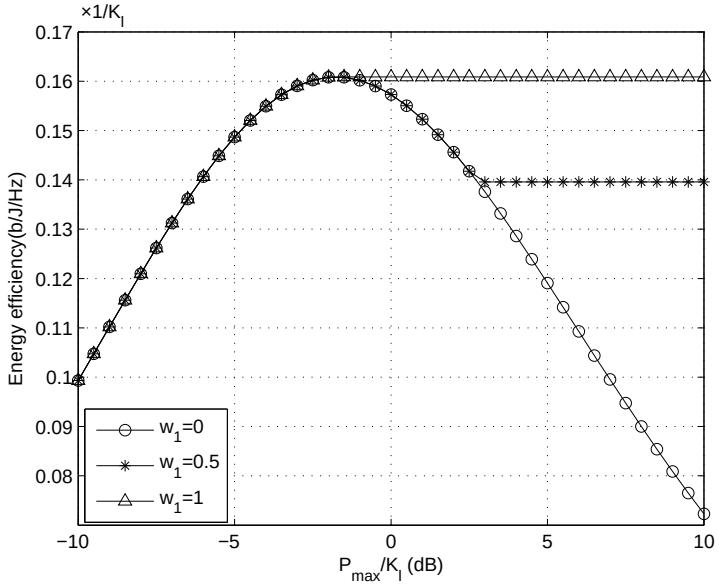


Figure 3.4: Maximum achievable EE versus scaled average power limit $\frac{P_{\max}}{K_\ell}$ for various values of w_1 in Rayleigh fading channels.

and $P_{\mathrm{c_r}} = -5$dB, is given in Fig. 3.5. Noting that the increase of w_1 increases the importance of EE in the tradeoff problem, therefore, when w_1 increases, $\overline{P_{\mathrm{r}}^*}$ monotonically decreases, which can be confirmed from Fig. 3.5. Further, this figure also

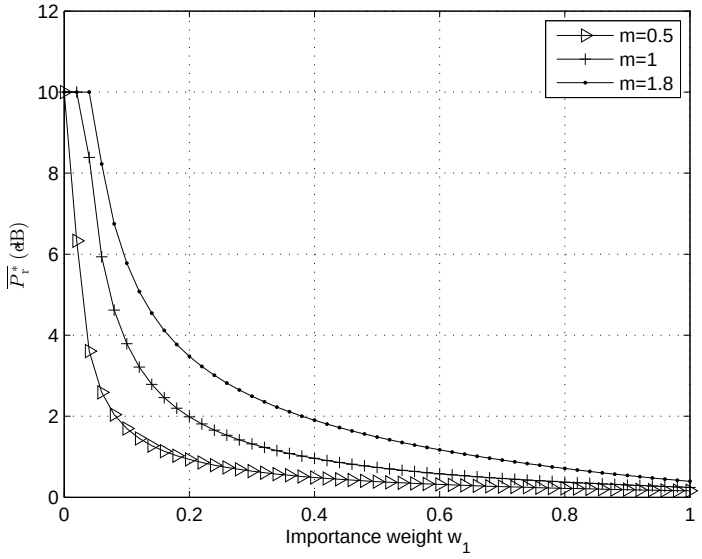


Figure 3.5: Normalized optimum average power value $\overline{P_{\rm r}^*}$ versus importance weight w_1 for various values of fading parameter m.

shows that, for a fixed w_1, when m increases, $\overline{P_{\rm r}^*}$ increases. This happens because with less channel fluctuations, the probability of the received data remaining in the FIFO buffer will be dropped, and therefore, EC and $\overline{P_{\rm r}^*}$ will increase. Further, from Fig. 3.5, it is noted that, for the cases of $m = 1$ and $m = 1.8$, $\overline{P_{\rm r}^*}$ first stabilizes at its maximum value $P_{\max}$, when w_1 is very small. This is because, in this region, the required optimal average power $\overline{P_{\rm t}^*}$ is larger than the available transmit power $P_{\max}$, therefore the tradeoff system can only operate at $P_{\max}$.

To show the tradeoff relationship between the link-layer EE and EC, the plots for EE versus EC, for various values of m, with $P_{\rm norm} = 0.5P_{\rm EE}^*$ and $P_{\rm c_r} = -5$dB, is plotted in Fig. 3.6. It shows that when $m = 1.8$, the MOP achieves the largest EE and EC, while the curve with the smallest m, i.e., $m = 0.5$, provides the least values of EE and EC.

Fig. 3.7 and Fig. 3.8 include the plots for EE and EC versus the importance weight w_1, for various values of $P_{\rm norm}$, with $P_{\rm c_r} = -5$dB, respectively. From Fig. 3.7, one can note that, when w_1 is relatively large, e.g., $w_1 \in [0.46, 1]$, EE shows a consistently upward trend with the increase of w_1, for all considered values of $P_{\rm norm}$. When w_1 is small, e.g., $w_1 \in [0, 0.46]$ and $P_{\rm norm} = P_{\max}$, EE initially remains constant until reaching a break-point, then gradually increases toward its maximum value. On the other hand, Fig. 3.8 shows that, when $P_{\rm norm} = P_{\max}$, EC levels off at its maximum value for a longer period of w_1, in comparison with the other EC curves

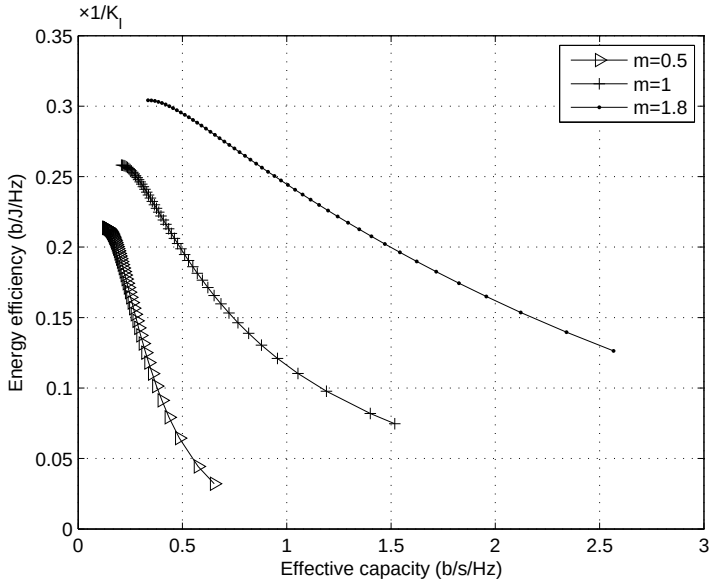


Figure 3.6: Maximum achievable EE versus EC for various values of Nakagami fading parameter m.

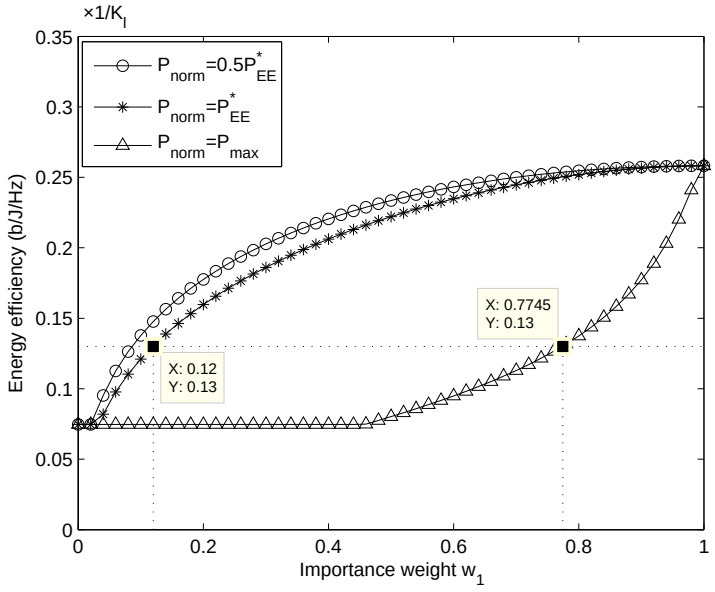


Figure 3.7: Maximum achievable EE versus importance weight w_1 for various values of P_{norm} in Rayleigh fading channels.

with $P_{\text{norm}} = 0.5P_{\text{EE}}^*$ and $P_{\text{norm}} = P_{\text{EE}}^*$. This provides a guideline for an EC-desired system and indicates that with a larger normalization factor P_{norm}, there is a better chance to make EC remain around its maximum value for a longer scope of varying w_1. Moreover, Fig. 3.7 and Fig. 3.8 demonstrate that the ranges of EC and EE,

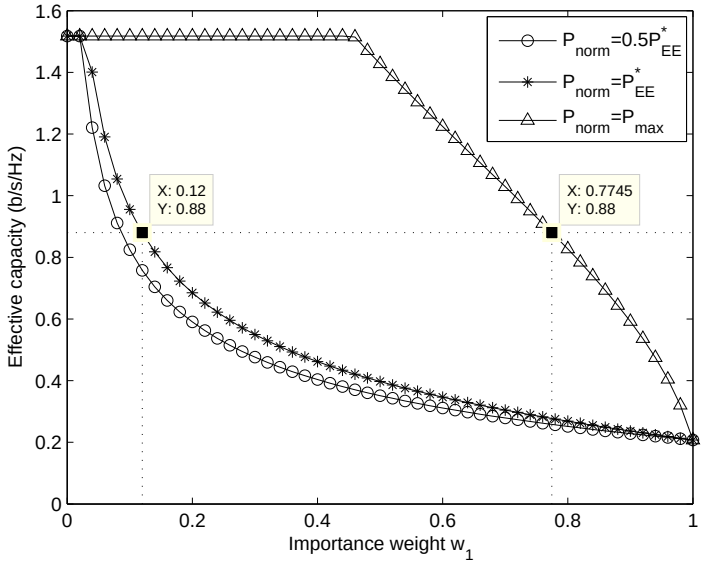


Figure 3.8: EC versus importance weight w_1 for various values of normalization factor P_{norm} in Rayleigh fading channels.

covered by $w_1 \in [0, 1]$, are always fixed, regardless of the different definitions of P_{norm}. For example, from Fig. 3.7, the EE curve with $P_{\text{norm}} = P^*_{\text{EE}}$, and the one with $P_{\text{norm}} = P_{\max}$, achieve the same value of the scaled EE, 0.13 b/J/Hz, at $w_1 = 0.12$ and $w_1 = 0.7745$, respectively. Meanwhile, in Fig. 3.8, EC obtained at $w_1 = 0.12$, and $P_{\text{norm}} = P^*_{\text{EE}}$, equals to the EC value achieved at $w_1 = 0.7745$, and $P_{\text{norm}} = P_{\max}$.

The results of EC (on the LHS y-Axis, in solid lines with markers) and EE (on the RHS y-Axis, markers only) versus w_1, for various values of θ, with $P_{\text{c}_\text{r}} = -5$dB, and $P_{\text{norm}} = P^*_{\text{EE}}$, are included in Fig. 3.9. As we discussed in Section 3.3.2.2, the case of $\theta \to 0$ refers to a system with no delay requirement, hence, EC is equivalent to the ergodic capacity. Although Fig. 3.9 indicates that, when $\theta \to 0$, EC and EE are larger than those obtained when $\theta = 10^{-2}$, one can also notice that when $\theta \to 0$, e.g., $\theta < 10^{-5}$, the delay-outage probability equals to 1, from Fig. 3.14. This is due to the fact that, for the physical-layer EE-SE tradeoff problem, no delay requirement means that the delay-outage probability can be very high. Further, Fig. 3.9 also shows that the physical-layer EC and EE, i.e., when $\theta \to 0$, follow the same trend with the link-layer EC and EE, with $\theta = 10^{-2}$.

The plots for EC versus the delay QoS exponent θ, under different power allocation policies, with $w_1 = 0.5$, $P_{\text{norm}} = P^*_{\text{EE}}$ and $P_{\text{c}_\text{r}} = -5$dB, is included in Fig. 3.10. Specifically, this figure compares the EC values under the optimal link-layer power allocation solution, which is derived in this chapter, and the traditional physical-layer

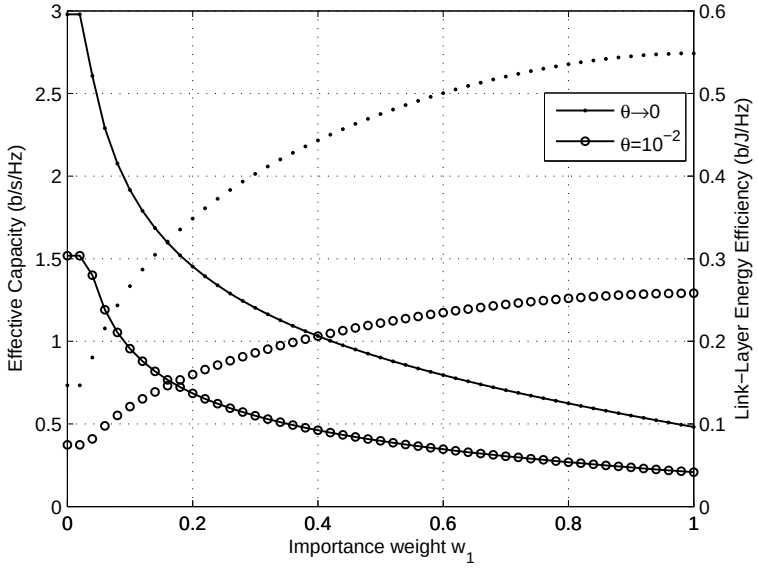


Figure 3.9: EC and link-layer EE versus importance weight w_1 for various values of θ in Rayleigh fading channels.

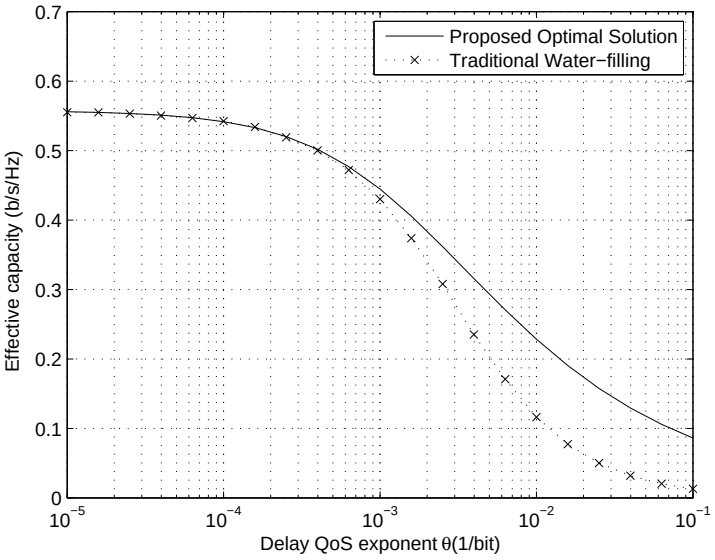


Figure 3.10: EC versus delay QoS exponent θ under different power allocation policies in Rayleigh fading channels.

water-filling approach. From Section 3.3.2.2, it is noted that when θ becomes very small, e.g., $\theta < 10^{-4}$, EC approaches to ergodic capacity. In this case, the proposed optimal power allocation strategy (3.14)-(3.18b) converges to the traditional water-filling strategy. Therefore, in Fig. 3.10, one can note that, when $\theta < 10^{-4}$, the values

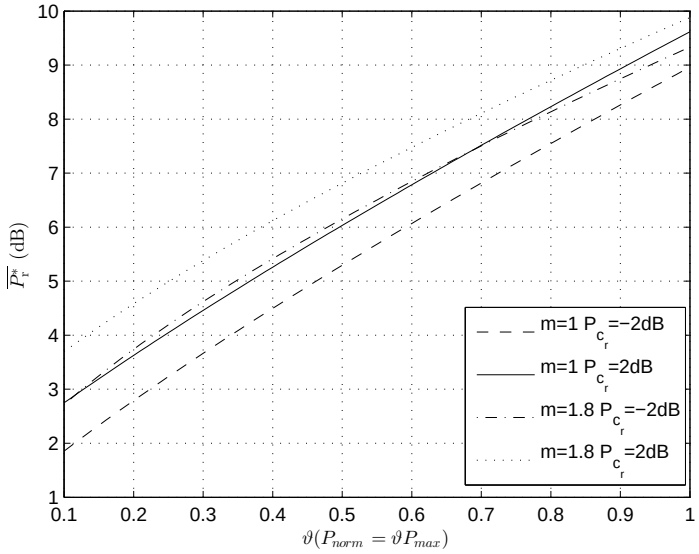


Figure 3.11: Normalized optimum average power value $\overline{P_{\rm r}^*}$ versus ϑ for various values of fading parameter m and scaled circuit-to-noise power ratio $P_{\rm c_r}$.

of EC for the two different power policies are equal. When θ becomes larger, e.g., $\theta \geq 10^{-3}$, which refers to a system with a stringent delay requirement, Fig. 3.10 indicates that the proposed link-layer optimal power allocation strategy guarantees a better performance than the traditional water-filling approach, with the water-filling performance approaching to zero when $\theta \to 0.1$.

Fig. 3.11 and Fig. 3.12 include the plots for the optimal average power and EE versus ϑ, where ϑ describes the ratio of $P_{\rm norm}$ to $P_{\rm max}$, i.e., $P_{\rm norm} = \vartheta P_{\rm max}$, for various values of the fading parameter m and the scaled circuit-to-noise power ratio $P_{\rm c_r}$. Specifically, a typical tradeoff system is considered and $w_1 = 0.5$. From Fig. 3.11 and Fig. 3.12, it is shown that, the optimal average power $\overline{P_{\rm r}^*}$ increases, while EE decreases with ϑ, for different values of m. This happens because when ϑ increases, $P_{\rm norm}$ becomes larger, which indicates that the priority of EE reduces and the importance of EC increases. Hence, the increase of ϑ leads to a lower value of EE, and a higher value of $\overline{P_{\rm r}^*}$. Furthermore, Fig. 3.11 indicates that for a fixed value of m, a system with a bigger $P_{\rm c_r}$ results in a larger $\overline{P_{\rm r}^*}$, which confirms the conclusion in Lemma 5.

Fig. 3.13 includes the plots for $\overline{P_{\rm r}^*}$ versus the delay QoS exponent θ for various values of w_1 and $P_{\rm norm}$, with $P_{\rm c_r} = -10$dB. When $w_1 = 0$, $\overline{P_{\rm r}^*}$ levels out at the maximum transmit power limit $P_{\rm max}$, which confirms that the EC-maximization system always consumes all the available power [12]. When $w_1 = 0.5$ and $P_{\rm norm} = P_{\rm max}$, $\overline{P_{\rm r}^*}$

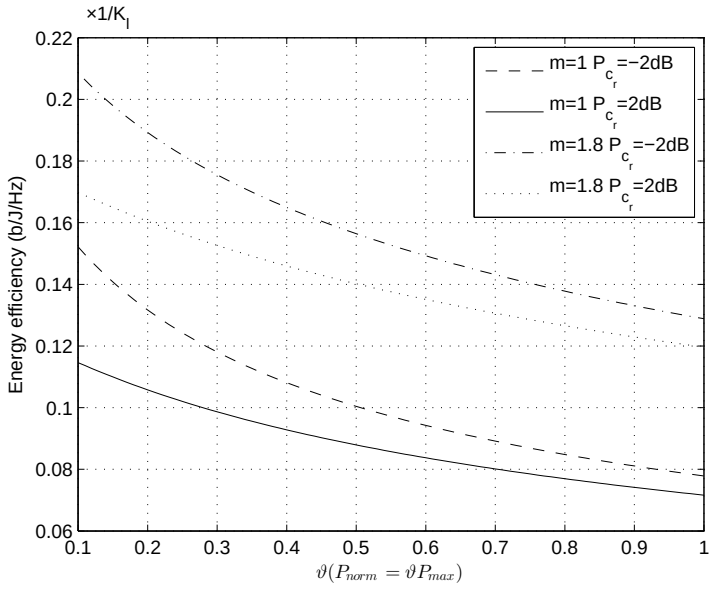


Figure 3.12: Maximum achievable EE versus ϑ for various values of Nakagami fading parameter m and scaled circuit-to-noise power ratio $P_{\mathrm{c_r}}$.

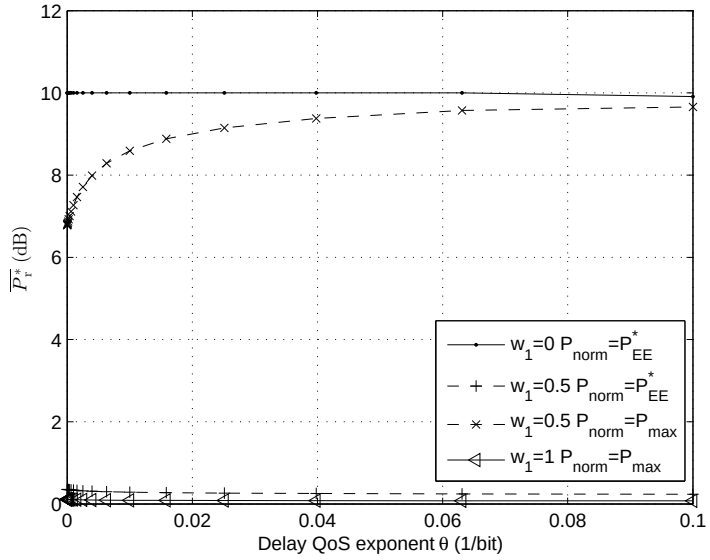


Figure 3.13: Normalized optimum average power value $\overline{P_{\mathrm{r}}^*}$ versus θ for various values of w_1 and P_{norm} in Rayleigh fading channels.

increases with θ, until it remains stable at a certain value, which is just under $P_{\max}$. In contrast, For cases of $w_1 = 0.5$ and $P_{\mathrm{norm}} = P_{\mathrm{EE}}^*$, and $w_1 = 1$ and $P_{\mathrm{norm}} = P_{\max}$, the optimum average power levels are achieved at minimal values. Furthermore, when $w_1 = 0.5$, $\overline{P_{\mathrm{r}}^*}$ obtained with $P_{\mathrm{norm}} = P_{\max}$ is higher, comparing to the case with

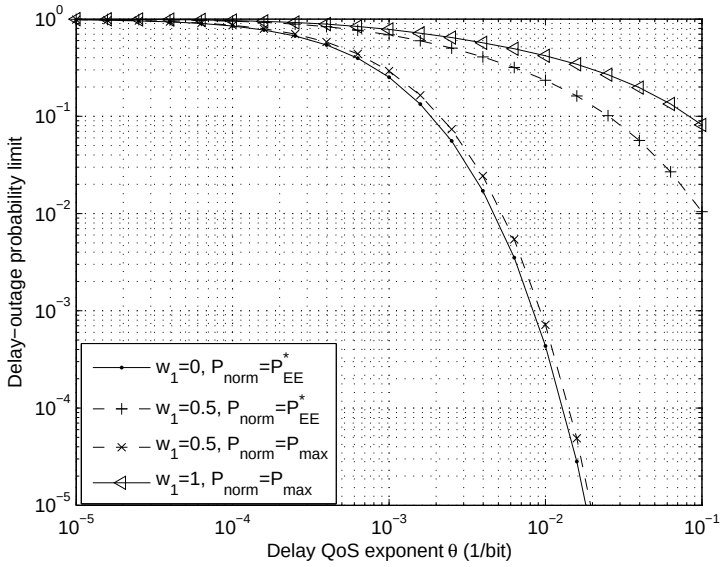


Figure 3.14: Delay-outage probability versus θ for various values of w_1 and normalization factor P_{norm} in Rayleigh fading channels.

$P_{\text{norm}} = P^*_{\text{EE}}$. This is due to the fact that, a larger P_{norm} reduces the priority of EE and raises the importance of EC, which results in a larger $\overline{P^*_{\text{r}}}$ and a smaller EE.

The delay-outage probability limit $P^{\text{out}}_{\text{delay}}$ versus the delay QoS exponent for various values of w_1 and P_{norm}, with a maximum tolerable delay threshold $D_{\max} = 500$ and the circuit-to-noise power ratio P_{c_r} = -10dB, is illustrated in Fig. 3.14. This figure indicates that for loose delay-constrained systems, e.g., $\theta \to 10^{-5}$, different values of w_1 will not affect the achievable $P^{\text{out}}_{\text{delay}}$ significantly. Also, in this case, the delay-outage probability approaches to 1, which means that the probability of the delay exceeding the maximum delay bound $D_{\max}$ approaches to 1. Further, for larger values of θ, e.g., $\theta \geq 10^{-3}$, the delay-outage probability increases with w_1. This happens because a smaller w_1 represents a system which prefers the EC-maximization approach. Hence, a higher EC will be achieved in this case and the probability that the delay exceeds a maximum delay-bound $D_{\max}$ will reduce. Furthermore, for a fixed θ, when $w_1 = 0.5$ and $P_{\text{norm}} = P^*_{\text{EE}}$, the delay-outage probability limit is larger than that with the same w_1 and $P_{\text{norm}} = P_{\max}$. This is due to the fact that a system with a larger P_{norm} offers a larger EC, which means that the probability of data remaining in the FIFO buffer will be dropped, and hence, the delay-outage probability $P^{\text{out}}_{\text{delay}}$ will be smaller.

3.5 Summary

A joint optimization problem of link-layer EE and EC in a Nakagami-m fading channel under a delay-outage probability constraint and an average transmit power constraint was considered and investigated in this chapter. First, a normalized MOP was formulated and transformed into an SOP, by applying the weighted sum method. Then, the formulated SOP was proved to be continuously differentiable and strictly quasiconvex in the optimum average input power. The weighted quasiconvex EE-EC tradeoff problem was then solved by applying the Charnes-Cooper transformation and KKT conditions. After obtaining the optimal power allocation scheme, which includes the optimal strategy for the link-layer EE-maximization problem and the EC-maximization problem as extreme cases, the proposed scheme was further proved to be sufficient for the Pareto optimal set of the original EE-EC MOP. In order to thoroughly analyze the tradeoff performance, the impact of the normalization factor, importance weight, scaled circuit-to-noise power ratio and power amplifier efficiency was analyzed and investigated. Furthermore, a proper guideline on how to choose the normalization factor and importance weight to build a more favorable system towards EE or EC was also provided.

In this chapter, the resource allocation problem in a point-to-point single-user single-carrier wireless communication system was proposed and investigated. As a natural extension, in the next chapter, the resource allocation problem, including the subcarrier allocation and the optimal power allocation strategy, is proposed and solved for the uplink transmission, in a multi-user multi-carrier orthogonal frequency division multiple access (OFDMA) system.

Chapter 4

Multi-User Multi-Carrier Link-Layer EE-EC Tradeoff

4.1 Introduction

As a natural extension of Chapter 3, Chapter 4 studies the delay quality-of-service (QoS) driven resource allocation for the uplink transmission in a multi-user multi-carrier orthogonal frequency division multiple access (OFDMA) system. A total effective capacity (EC) maximization problem is formulated as a combinatorial integer programming problem, subject to each user's link-layer energy efficiency (EE) requirement as well as the per-user average transmission power limit. To solve this challenging problem, it is first decoupled into a frequency provisioning problem and an independent multi-carrier link-layer EE-EC tradeoff problem for each user. In order to obtain the subcarrier assignment solution, a low-complexity heuristic algorithm is proposed, which not only offers close-to-optimal solutions, while serving as many users as possible, but also has a complexity linearly relating to the size of the problem. After obtaining the subcarrier assignment matrix, the original formulated problem reduces to a link-layer EE-EC tradeoff problem for each single-user multi-carrier system. Although in Chapter 3, the optimal power allocation strategy for the link-layer EE-EC tradeoff problem has been proposed for a single-user single-carrier system, it cannot be simply extended to the multi-carrier communications [20]. Considering the multi-carrier EE-EC tradeoff problem for each user, the per-user optimal power allocation strategy, across both frequency and time domains, is derived and analyzed. The impact of the circuit power and the EE requirement factor on each user's link-layer EE level and optimal average power is then theoretically investigated. Simulation results compare the proposed low-complexity heuristic algorithm with the traditional exhaustive algorithm and a fair-exhaustive algorithm, which confirm our

design intentions, and further show the effects of delay QoS exponent, the total number of users and the number of subcarriers on the system tradoeff performance.

The remainder of this chapter is organized as follows. In Section 4.2, the system model and problem formulation are introduced. Then, the optimal and suboptimal resource allocation solutions are proposed in Section 4.3, including frequency provisioning algorithms, optimal power allocation strategy for each single-user multi-carrier system, and also theoretical analysis of some system parameters' influence. Finally, simulation results are given in Section 4.4, followed by conclusions in Section 4.5.

4.2 System Model and Problem Formulation

4.2.1 System Model

We consider the uplink transmission, where the K active users send their own information to the base station, in a multi-user multi-carrier OFDMA system depicted in Fig. 4.1. A total bandwidth of B is divided into N subcarriers, each with a bandwidth of $\frac{B}{N}$. Assume that each subcarrier is exclusively assigned to at most one user at each time to avoid interference among different users. The total number of allocated subcarriers for all users does not exceed the available frequency resources. Therefore, a feasible subcarrier assignment indicator matrix can be denoted as $\boldsymbol{\phi}$, which satisfies

$$\boldsymbol{\phi} \in \Phi \triangleq \left\{ [\phi_{k,n}]_{K\times N} \mid \phi_{k,n} \in \{0,1\}, \sum_{k=1}^{K} \phi_{k,n} \leq 1, \sum_{k=1}^{K}\sum_{n=1}^{N} \phi_{k,n} \leq N, k \in \mathcal{K}_0, n \in \mathcal{N}_0 \right\}. \tag{4.1}$$

Here, Φ denotes the set of all possible subcarrier allocation indicator matrices, and $\mathcal{K}_0 = \{1, 2, \ldots, K\}$, $\mathcal{N}_0 = \{1, 2, \ldots, N\}$ denote the set of all users and all subcarriers, respectively. The number of allocated subcarriers for the k^{th} user is denoted by N_k, namely, $N_k = \sum_{n=1}^{N} \phi_{k,n}$, and the bandwidth allocated to the k^{th} user is denoted by B_k, i.e., $B_k = N_k \frac{B}{N}$.

Each transmitter implements a first-in-first-out (FIFO) buffer, which prevents loss of packets that could occur when the source rate is higher than the service rate, at the expense of increasing the delay [10]. The upper-layer packets are divided into frames at the data-link layer and are stored at the transmit buffer. The frames are then split into bit streams at the physical layer. By utilizing perfect channel state information (CSI) knowledge fed back from the receiver and the predetermined statistical QoS

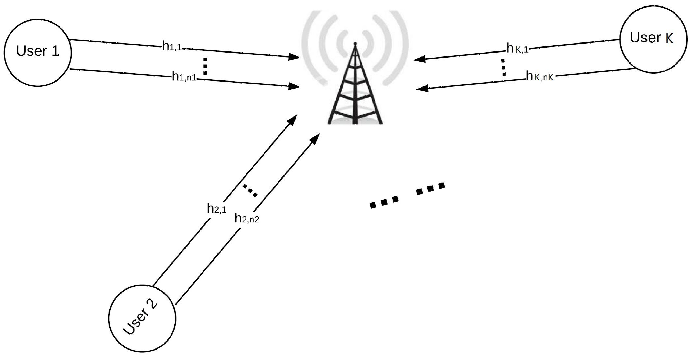


Figure 4.1: Uplink transmission in a multi-user multi-carrier network.

constraint, adaptive modulation and coding (AMC) and adaptive power control policy are applied at the transmitter side [20]. Then, the bit streams are read out of the buffer and are transmitted through the wireless fading subcarriers. At the receiver side, the reverse operations are performed and the frames are recovered for further processing. Each subcarrier is assumed to be block fading, i.e., the channel gains are invariant within a fading-block's time duration T_{f}, but independently varies from one fading block to another. In addition, the length of each fading-block, T_{f}, is considered to be an integer multiple of the symbol duration T_{s}, and is assumed to be less than the fading coherence time [20].

For the k^{th} user on the n^{th} subcarrier at the fading-block index t, the subcarrier power gain is denoted by $\gamma_{k,n}[t], k \in \mathcal{K}_0, n \in \mathcal{N}_0$. Also, each subcarrier is assumed to experience independent and identically distributed (i.i.d.) additive white Gaussian noise (AWGN) with the single-sided power spectral density η_0. Therefore, the instantaneous maximum achievable rate of the k^{th} user on the n^{th} subcarrier at the t^{th} fading-block is given by

$$R_{k,n}[t] = \frac{B}{N} T_{\mathrm{f}} \log_2 \left(1 + P_{k,n}[t] \frac{\gamma_{k,n}[t]}{P_{\mathcal{L}}^{k} \eta_0 \left(\frac{B}{N}\right)} \right) \text{ (bits)}, \tag{4.2}$$

where $P_{\mathcal{L}}^{k}$ denotes the distance-based path-loss power and $P_{k,n}[t]$ is the nonnegative transmission power for the k^{th} user on the n^{th} subcarrier, at the t^{th} fading-block, i.e., $P_{k,n}[t] \geq 0$. Specifically, for the k^{th} user, the vector of subcarrier power allocation values is denoted as $\boldsymbol{P}_k[t] = \left[P_{k,1}[t], \ P_{k,2}[t], \ \ldots \ P_{k,N}[t]\right]$[1]. The total achievable rate over all allocated subcarriers for the k^{th} user, which depends on the subcarrier allocation indicator matrix $\boldsymbol{\phi}$ and the subcarrier power allocation vector $\boldsymbol{P}_k$, can

[1]Since the service rate process of the k^{th} user on the n^{th} subcarrier is considered to be stationary and ergodic [20], hereafter, the block index t could be omitted for simplicity.

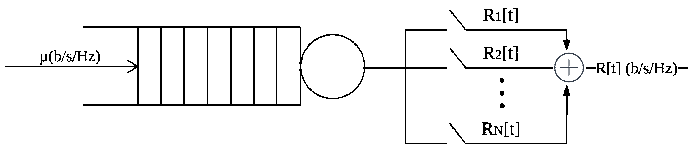


Figure 4.2: Queuing system model for each transmitter.

be denoted as $\mathbf{R}_k(\boldsymbol{\phi}, \boldsymbol{P}_k) = \sum_{n \in \mathcal{N}_k} \phi_{k,n} R_{k,n}$, where $\mathcal{N}_k$ is the index set of subcarriers allocated to the k^{th} user.

4.2.2 Effective Capacity and Link-Layer Energy Efficiency

For each transmitter, the FIFO buffer can be considered as a dynamic queueing system which connects the stationary ergodic arrival and service processes, depicted in Fig. 4.2 [13]. From Chapter 2.1, we note that by using the large deviation theory, the queue length process $Q(t)$ converges in distribution to a steady-state queue length $Q(\infty)$ such that [13]

$$-\lim_{x \to \infty} \frac{\ln\left(\Pr\{Q(\infty) > x\}\right)}{x} = \theta, \tag{4.3}$$

where $\Pr\{a > b\}$ shows the probability that $a > b$ holds. This implies that the probability of the queue length exceeding a certain threshold x decays exponentially fast as x increases [39]. Note that in (4.3), the parameter $\theta\,(\theta > 0)$ indicates the exponential decay rate of the QoS violation probability. A smaller θ denotes a looser QoS requirement, while a larger θ implies a lower probability of violating the queue length and a more stringent delay constraint. Particularly, when $\theta \to 0$, refering to a system with no delay constraint, the optimum power allocation strategy is the traditional water-filling approach and the maximum achievable rate is ergodic capacity [86]. For a transmitter with $\theta \to \infty$, the optimum power allocation is the channel inversion with fixed rate transmission technique, under which the delay-limited capacity can be achieved [98]. In other words, the ergodic capacity and the delay-limited capacity can be considered as two extreme cases of the concept of EC.

Furthermore, when the focus is on the delay experienced by a source packet arriving at time t, defined by $D(t)$, the delay-outage probability $P_{\text{delay}}^{\text{out}}$ can be given in (2.27). From Chapter 2.1, we note that, in order to meet a target delay-bound violation probability limit, $P_{\text{delay}}^{\text{out}}$, EC represents the maximum constant arrival rate that the current service process can support.

Assume that the Gärtner-Ellis theorem [52, Pages 34-36] is satisfied. For the k^{th} user, the EC over a multi-carrier transmission with a total bandwidth B_k can be expressed as [10]

$$E_{\text{c}}^{k}\left(\theta_k, \boldsymbol{\phi}, \boldsymbol{P}_k\right) = -\frac{1}{\theta_k T_{\text{f}} B_k} \ln\left(\mathbb{E}\left[e^{-\theta_k \mathbf{R}_k(\boldsymbol{\phi}, \boldsymbol{P}_k)}\right]\right) \quad \text{(b/s/Hz)}, \tag{4.4}$$

where θ_k stands for the delay QoS exponent of the k^{th} user which is associated with the statistical delay QoS requirement and $\mathbb{E}[\cdot]$ indicates the expectation operator. Henceforth, the EC of the k^{th} user becomes a function of θ_k, $\boldsymbol{\phi}$, and $\boldsymbol{P}_k$.

By expanding $\mathbf{R}_k\left(\boldsymbol{\phi}, \boldsymbol{P}_k\right)$ and inserting it into (4.4), the EC of the k^{th} user can be further expressed as

$$E_{\text{c}}^{k}\left(\theta_k, \boldsymbol{\phi}, \boldsymbol{P}_k\right) = -\frac{1}{\theta_k T_{\text{f}} B_k} \ln\left(\mathbb{E}\left[e^{-\theta_k \sum\limits_{n\in\mathcal{N}_k} \phi_{k,n} R_{k,n}}\right]\right) \quad \text{(b/s/Hz)}. \tag{4.5}$$

For the multi-user OFDMA network, the total EC can be expressed as

$$E_{\text{c}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbb{P}\right) = \frac{\sum_{k=1}^{K} N_k E_{\text{c}}^{k}\left(\theta_k, \boldsymbol{\phi}, \boldsymbol{P}_k\right)}{\sum_{k=1}^{K} N_k} \quad \text{(b/s/Hz)}, \tag{4.6}$$

where $\boldsymbol{\theta} = \begin{bmatrix}\theta_1, & \theta_2, & \ldots & \theta_K\end{bmatrix}$ is the $1 \times K$ vector of delay exponents for all K users. $\mathbb{P}$ denotes the transmission power allocation matrix, for all users over all subcarriers, i.e., $\mathbb{P} \in \mathcal{P} \triangleq \left\{[P_{k,n}]_{K\times N} \in \mathbb{R}_{+} \mid \mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N} \phi_{k,n} P_{k,n}\right] \leq P_{\max}^{k}, k \in \mathcal{K}_0\right\}$[2]. Here, $\mathcal{P}$ is all the possible power allocation matrices, and $\mathbb{E}_{\boldsymbol{\gamma}_k}[\cdot]$ indicates the expectation over the probability density function (PDF) of $\boldsymbol{\gamma}_k$, where $\boldsymbol{\gamma}_k$ is the k^{th} user's subcarrier power gains, i.e., $\boldsymbol{\gamma}_k = \begin{bmatrix}\gamma_{k,1}, & \gamma_{k,2}, & \ldots & \gamma_{k,N_k}\end{bmatrix}$. Further, $P_{\max}^{k}$ represents the maximum average power limit of the k^{th} user.

Moreover, for the k^{th} user, the link-layer EE can be defined as the ratio of EC to the sum of its circuit power P_{c}^{k}, and the average transmission power scaled by the power amplifier efficiency ϵ, yielding

$$\text{EE}^{k}\left(\theta_k, \boldsymbol{\phi}, \boldsymbol{P}_k\right) = \frac{E_{\text{c}}^{k}\left(\theta_k, \boldsymbol{\phi}, \boldsymbol{P}_k\right)}{P_{\text{c}}^{k} + \frac{1}{\epsilon}\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum\limits_{n\in\mathcal{N}_k} \phi_{k,n} P_{k,n}\right]}. \tag{4.7}$$

[2]We note that these are other kinds of power constraints adopted in the literature, such as the peak power constraints, power outage probability constraints, etc. In this paper, we consider the expectation power constraints, i.e., the average power is limited to a maximum value.

4.2.3 Problem Formulation

From a system point of view, the total EC needs to be maximized to achieve the best system performance. On the other hand, from the individual user point of view, each user has its own link-layer EE requirement, average transmission power limit and delay QoS constraint. Therefore, considering a multi-user multi-carrier network, the overall system throughput maximization problem, subject to each user's resource constraints, can be formulated as

$$
\begin{aligned}
Q1: \max_{\boldsymbol{\phi}\in\Phi,\mathbb{P}\in\mathcal{P}} \quad & E_{\mathrm{c}}\left(\boldsymbol{\theta},\boldsymbol{\phi},\mathbb{P}\right) && \text{(4.8a)}\\
\text{subject to:} \quad & \mathrm{EE}^{k}\left(\theta_k,\boldsymbol{\phi},\boldsymbol{P}_k\right)\geq\eta_{\mathrm{req}}^{k},\ \forall k, && \text{(4.8b)}\\
& \mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N}\phi_{k,n}P_{k,n}\right]\leq P_{\max}^{k},\ \forall k, && \text{(4.8c)}\\
& \sum_{k=1}^{K}\phi_{k,n}\leq 1,\ \forall n, && \text{(4.8d)}\\
& \sum_{k=1}^{K}\sum_{n=1}^{N}\phi_{k,n}\leq N, && \text{(4.8e)}\\
& \phi_{k,n}\in\{0,1\},\ \forall k,\forall n, && \text{(4.8f)}\\
& P_{k,n}\geq 0,\ \forall k,\forall n, && \text{(4.8g)}
\end{aligned}
$$

where η_{req}^{k} is the k^{th} user's required link-layer EE level, defined by a certain ratio of its maximum achievable link-layer EE value, i.e., $\eta_{\mathrm{req}}^{k}=\chi_{\mathrm{EE}}^{k}\times\eta_{\max}^{k,N}$. Here, $\eta_{\max}^{k,N}=\mathrm{EE}^{k}\Big|_{\overline{P_k}=\overline{P_{\mathrm{EE}}^{k*}}}^{N_k=N}$ denotes the k^{th} user's maximum achievable EE value, when all N subcarriers in the system are allocated to it. $\overline{P_{\mathrm{EE}}^{k*}}$ is the operational average input power which achieves $\eta_{\max}^{k,N}$. Further, $\chi_{\mathrm{EE}}^{k}\in[0,1]$ is an adjustable EE requirement factor, which reveals the strictness of the k^{th} user's required EE level and directly influences the system performance. In particular, $\chi_{\mathrm{EE}}^{k}=0$ indicates that the k^{th} user has no EE requirement, while $\chi_{\mathrm{EE}}^{k}=1$ means that the user k requires a maximum EE value $\eta_{\max}^{k,N}$. Since $\eta_{\max}^{k,N}$ depends on the individual user's delay QoS exponent and its maximum averge power limit, therefore this value is different for each user. Hence, defined as a ratio of $\eta_{\max}^{k,N}$, the k^{th} user's required EE level η_{req}^{k} is also different from the other users, even when they have the same EE requirement factors.

Due to the conflicting property of the total EC and each user's personal EE achievement, after introducing χ_{EE}^{k}, the formulated problem $Q1$ becomes an adjustable tradeoff problem. To be more specific, if the total system throughput has a

high priority, each user's EE requirement factor can be required to be very low, which results in a low link-layer EE level for each user. Correspondingly, if the total system throughput has a low priority, each user's EE requirement factor can be relatively high, so that each user will have a satisfied high level of link-layer EE.

4.3 Optimal and Sub-optimal Solutions

Since it is assumed that one subcarrier can be assigned to only one user at a time, therefore there could be K^N possible subcarrier assignments [107]. Hence, the complexity of the above combinatorial integer programming problem in finding the jointly optimal subcarrier and power allocation grows exponentially with the number of subcarriers. Furthermore, it is very difficult to jointly obtain the optimal subcarrier allocation sets and all power allocation values in every frame, due to the reasons below. Firstly, from (4.5), one can notice that the EC formulation of the k^{th} user not only requires the multiplication of two unknown parameters, i.e., $\phi_{k,n}$, and $R_{k,n}$, but also involves the expectation over the joint PDF of all subcarriers' channel power gains, i.e., $\boldsymbol{\gamma}_k$. Secondly, the expectation and the multiplication operations cannot be interchanged, even if all subcarriers are assumed to be i.i.d., and that is because the power allocation value on each subcarrier is related to the other subcarriers.

Henceforth, in order to make the formulated problem $Q1$ tractable, we divide the solving process into two steps: frequency provisioning which decides the number of subcarriers to be allocated to each user; and optimal power allocation for each user over all its allocated subcarriers. Specifically, the proposed frequency provisioning algorithms, which are independent of the instantaneous CSI knowledge in each frame, will be implemented only once within a period of time. On the other hand, for each user, the proposed optimal power allocation strategy on each subcarrier, not only relies on the instantaneous CSI of this subcarrier, but also depends on the other subcarriers' CSI knowledge in each frame.

We start from introducing three frequency provisioning algorithms: traditional exhaustive algorithm[3], fair-exhaustive algorithm and the proposed low-complexity heuristic frequency allocation algorithm. After obtaining the subcarrier assignments, the optimal power allocation strategy for each single-user multi-carrier system will then be derived and obtained in Section 4.3.2.

[3]In order to clearly compare the three frequency provisioning algorithms, the traditional exhaustive algorithm is also briefly introduced in Section 4.3.1, although it is generally well-known.

4.3.1 Frequency Provisioning Algorithms

By applying frequency provisioning, it is assumed that all subcarriers follows the same distribution. It is the number of designated subcarriers which matters, regardless where those subcarriers are located in the frequency band [107]. To reduce the problem complexity and the solving time, a pre-calculated offline database $\mathcal{D}$ is first built which stores all users' maximum achievable link-layer EE values, i.e., $\boldsymbol{\eta}_{\max} = \left[\boldsymbol{\eta}^1_{\max}, \ \boldsymbol{\eta}^2_{\max}, \ \ldots \ \boldsymbol{\eta}^K_{\max}\right]^{\mathrm{T}}$, in terms of certain settings of P_{c}, θ and N. Here, $\boldsymbol{\eta}^k_{\max}$ is a $1 \times N$ vector of the k^{th} user's maximum achievable EE values with different number of allocated subcarriers, i.e., $\boldsymbol{\eta}^k_{\max} = \left[\eta^{k,1}_{\max}, \ \eta^{k,2}_{\max}, \ \ldots \ \eta^{k,N}_{\max}\right]$. Specifically, in order to calculate $\boldsymbol{\eta}^k_{\max}$, the optimal power allocation strategy introduced in [19] for an EE-maximization problem in a single-user multi-carrier system is utilized. In this case, we assume that the average input power limit is large enough to support the calculated optimum power value. Define $\boldsymbol{\eta}_{\mathrm{req}} = \left[\eta^1_{\mathrm{req}}, \ \eta^2_{\mathrm{req}}, \ \ldots \ \eta^K_{\mathrm{req}}\right]^{\mathrm{T}}$ as the $K \times 1$ vector of the EE requirement values for all K users. Then, we can transform $\boldsymbol{\eta}_{\mathrm{req}}$ to a $K \times 1$ vector which specifies all users' required number of subcarriers, i.e., $\boldsymbol{S}_{\mathrm{req}} = \left[S^1_{\mathrm{req}}, \ S^2_{\mathrm{req}}, \ \ldots \ S^K_{\mathrm{req}}\right]^{\mathrm{T}}$.

Let us consider the k^{th} user as an example. Its required link-layer EE value is denoted by η^k_{req}, and correspondingly, its subcarrier requirement value will be stored as S^k_{req}. To obtain S^k_{req}, a flowchart is provided in Fig. 4.3 to compare η^k_{req} with $\boldsymbol{\eta}^{\boldsymbol{k}}_{\max}$. If the maximum achievable EE value obtained with i subcarriers is larger than the required EE value, i.e., $\eta^{k,i}_{\max} \geq \eta^k_{\mathrm{req}}$, then one can conclude that the minimum number of subcarriers required to satisfy the k^{th} user's EE requirement η^k_{req}, is i, i.e., $S^k_{\mathrm{req}} = i$. Henceforth, all K users' EE requirements in $\boldsymbol{\eta}_{\mathrm{req}}$ can be transformed to the subcarrier requirement vector $\boldsymbol{S}_{\mathrm{req}}$, by utilizing the flowchart in Fig. 4.3. In this way, the feasibility of each user's EE constraint can be easily checked by comparing the number of allocated subcarriers with the number of required subcarriers.

4.3.1.1 Traditional Exhaustive Algorithm

The traditional way to solve the formulated NP-hard problem, i.e., (4.8a)-(4.8g), is to carry out an exhaustive search, which systematically enumerates all possible combinations and finally locates the solution which optimizes the objective function and satisfies all the problem constraints [107]. Specifically, for problem $Q1$, the set of feasible combinations is found first. Then, the optimal power allocation strategy proposed in the next section, will be applied to all the feasible combinations. Finally, the feasible combination which offers the maximum system throughput will be chosen

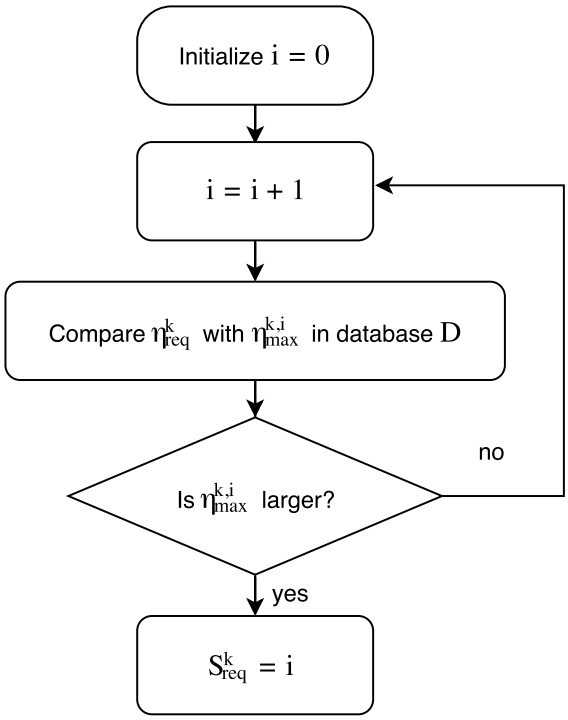


Figure 4.3: Transform η_{req}^{k} to S_{req}^{k}.

as the optimal solution. Although exhaustive search is able to find the optimal frequency provisioning solution, it also lacks user fairness and has a high computational complexity which exponentially grows with the size of the problem.

4.3.1.2 Fair-Exhaustive Algorithm

To further find the optimal frequency provisioning solution which not only maximizes the total system EC value, while satisfying each user's link-layer EE requirement, but also serves the maximum number of users that can be allowed, we propose a fair exhaustive algorithm. Firstly, the sum of all users' required subcarriers is compared with the total number of subcarriers N to find the maximum number of users that can be served. For example, let us assume $N = 8$, $K = 4$, and the subcarrier requirement vector for all users is $[1, 2, 2, 4]$. Hence, the total available subcarriers can serve 3 users at most. Secondly, the set of feasible subcarrier allocation vectors is found, in which each allocation vector not only satisfies all served users' subcarrier requirements, but also serves the maximum allowed number of users. Then, the optimal power allocation strategy proposed in Section 4.3.2 will be applied to all feasible allocation vectors to locate the fair and optimal solution which outperforms the others.

Clearly, by enumerating all possible subcarrier allocation vectors which can serve

the allowed maximum number of users, the above proposed algorithm exhaustively find the optimal solution in a fair way. Although the fair-exhaustive algorithm is less complex compared to the traditional exhaustive algorithm, but its computational complexity is still very high, especially when the number of available subcarriers N is large. To further reduce the solving complexity, we provide the following heuristic algorithm, which is simple, fair and close-to-optimal.

4.3.1.3 Heuristic Algorithm

There are three steps included in the proposed heuristic frequency provisioning algorithm, which are allocation process, calculation process and check process. Firstly, in order to serve as many users as possible, in the allocation process, we start from the user which requires the minimum number of subcarriers. Each served user will be allocated the exact number of its required subcarriers, so that all the allocated users can satisfy their EE requirements. The allocation will be repeated until the remaining subcarriers run out, or there are not enough subcarriers to satisfy the next user's EE requirement, or all users' subcarrier requirements have already been satisfied. Then, the calculation process starts, in which each served user operates the optimal power allocation strategy described in Table 4.2 to obtain its current EC value. In the check process, the aim is to maximize the system throughput, based on the strategy that the user with current minimum EC value has the allocation priority. Therefore, the remaining subcarriers will be assigned one-by-one to the user which has the current minimum EC value, until all subcarriers run out.

Assume the final subcarrier allocation vector is denoted by $\boldsymbol{N} = \begin{bmatrix} N_1, & N_2, & \dots & N_K \end{bmatrix}$. The Pseudocode of the proposed heuristic algorithm is illustrated in Table 4.1. Note that the proposed algorithm only needs at most $K-1$ comparisons per iteration, given that each user's EC values with various number of subcarriers are pre-calculated off-line and are stored in a database. Therefore, the heuristic algorithm offers a relatively low computational complexity comparing to the two exhaustive algorithms whose complexity exponentially increase with the number of subcarriers. On the other hand, later, in simulation results, it is shown that the proposed low-complexity algorithm offers a close performance with the fair-exhaustive algorithm.

Now, let us analyze and explain the strategies utilized in the proposed heuristic algorithm. Firstly, in the allocation process, the heuristic algorithm starts the allocation from the user which has the minimum subcarrier requirement. Assume that user i has the relatively small number of required subcarriers, S^i_{req}. By regarding

Table 4.1: Heuristic Algorithm for a Multi-User Multi-Carrier System

Initialization:

Calculate $\boldsymbol{S}_{\text{req}}$, using $\boldsymbol{\eta}_{\text{req}}$ and the pre-calculated database $\mathcal{D}$.

Define $S_{\text{tol}} = N$, $\boldsymbol{H} = \boldsymbol{S}_{\text{req}}$.

Allocation Process:

While $S_{\text{tol}} > 0$

If $\boldsymbol{H} = 0$

Break;

End

Find $H_i = \min(\boldsymbol{H})$, and $H_i > 0$;

If $S_{\text{tol}} > S^i_{\text{req}}$

$N_i = S^i_{\text{req}}$;

$S_{\text{tol}} = S_{\text{tol}} - S^i_{\text{req}}$;

$H_i = 0$;

Else

Break;

End

End

Calculation Process:

For each user i with $H_i = 0$, apply the optimal power allocation in Table 4.2.

Calculate the i^{th} user's EC value J_i and define $\boldsymbol{J} = \begin{bmatrix} J_1, & J_2, & \dots & J_K \end{bmatrix}$.

Check Process:

While $S_{\text{tol}} > 0$

Find $J_i = \min(\boldsymbol{J})$, in which user i satisfies $H_i = 0$;

$N_i = N_i + 1$;

Apply the optimal power allocation algorithm to user i and update J_i.

End

Output: $\boldsymbol{N}$; E_{c} given in (4.6).

θ_i and χ^i_{EE} as the two influencing parameters on S^i_{req}, a small value of S^i_{req} may result from the following two possibilities: 1) user i has a small delay QoS exponent θ_i and the same χ^i_{EE} value, comparing to the other users; 2) user i has a small EE requirement factor χ^i_{EE}, and the same θ_i value, comparing to the others. For the first situation, a small delay QoS exponent θ_i means a loose requirement on delay QoS, which will offer a bigger EC value, when the allocated number of subcarriers and χ^i_{EE} are fixed. Meanwhile, for the second situation, a small value of χ^i_{EE} also provides a larger EC value, because now the EE requirement constraint is easy to be satisfied and the multi-carrier system will have more resource and flexibility to maximize the EC performance. Consequently, the design idea of the allocation process not only makes sure that as many users as possible can be served, but also intends to serve the user which can contribute a larger EC value.

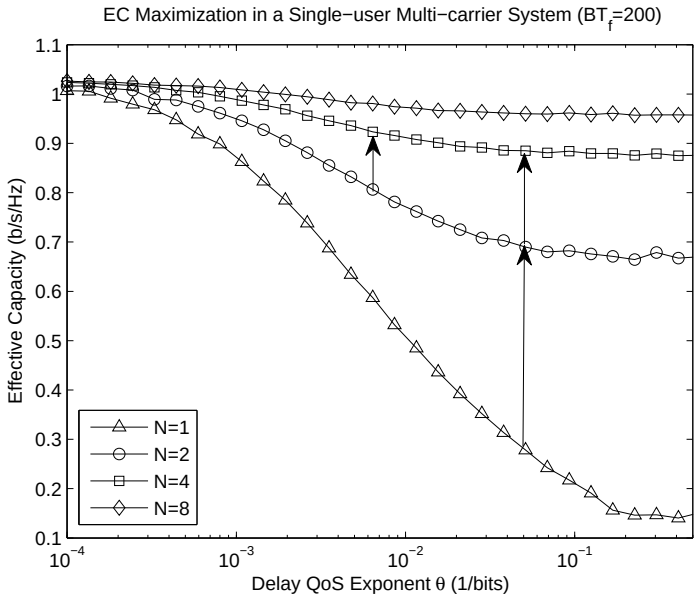


Figure 4.4: Effective capacity versus delay QoS exponent θ, for various values of N.

On the other hand, the design strategy of the check process, i.e., the user with current minimum EC value has the allocation priority, comes from Fig. 4.4, which describes the results of the maximum EC versus delay QoS exponent θ, for various values of N, in a single-user multi-carrier system. Specifically, Fig. 4.4 reveals that the user with current minimum EC value has a high possibility to offer the largest EC-increase, if given one more subcarrier. In more detail, from Fig. 4.4, it shows that for two users with the same values of θ, if we can allocate more subcarriers to them, the user with current smaller EC value, i.e., the one which has smaller number of subcarriers, will get a larger EC-increase. Furthermore, Fig. 4.4 shows that, for two users with the same number of subcarriers, if we allocate each user two more subcarriers, the user with relatively smaller EC value, namely, the one which has larger delay QoS exponent, will have a larger EC-increase. Simulation results in Section 4.4 confirm the effectiveness of this design method, and inform that the proposed heuristic algorithm offers very close performance with the fair-exhaustive algorithm.

4.3.2 Optimal Power Allocation for a Single-User Multi-Carrier System

Given a subcarrier assignment matrix $\boldsymbol{\phi}$, the multi-user OFDMA system can be viewed as a frequency-division multiple access (FDMA) system, where each user transmits data through a number of assigned subcarriers independently [108]. Therefore,

the original total EC maximization problem, subject to each user's link-layer EE requirement and maximum average power limit, can be transformed into a link-layer EE-EC tradeoff problem for each single-user multi-carrier system.

Specifically, for the k^{th} user, the problem can be expressed as

$$Q2: \max_{\substack{P_{k,n}\geq 0 \\ n\in\mathcal{N}_k}} \quad E_{\text{c}}^{k}\left(\theta_k, \boldsymbol{P}_k\right) \tag{4.9a}$$

$$\text{subject to:} \quad \text{EE}^{k}\left(\theta_k, \boldsymbol{P}_k\right) \geq \eta_{\text{req}}^{k}, \tag{4.9b}$$

$$\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k} P_{k,n}\right] \leq P_{\max}^{k}. \tag{4.9c}$$

By recalling that the total bandwidth allocated to the k^{th} user is B_k, the total instantaneous service rate of the k^{th} user is given by

$$R_k = \frac{B_k}{N_k} T_{\text{f}} \sum_{n=1}^{N_k} \log_2\left(1 + P_{k,n}\frac{\gamma_{k,n}}{P_{\mathcal{L}}^{k}\eta_0\left(\frac{B_k}{N_k}\right)}\right) \text{ (bits)}. \tag{4.10}$$

By inserting (4.10) into (4.4), we get the mathematical expression of EC for the k^{th} user. Correspondingly, the link-layer EE for the k^{th} user, as the ratio of EC to the total power expenditure, can be obtained. Therefore, problem $Q2$ can be expanded as

$$Q3: \max_{\substack{P_{k,n}^{\text{r}}\geq 0 \\ n\in\mathcal{N}_k}} \quad -\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{N_k}\left(1+N_k P_{k,n}^{\text{r}}\gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right) \tag{4.11a}$$

$$\text{subject to}: \quad \frac{-\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{N_k}\left(1+N_k P_{k,n}^{\text{r}}\gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right)}{K_{\ell}^{k}\left(P_{\text{c}_{\text{r}}}^{k} + \frac{1}{\epsilon}\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k} P_{k,n}^{\text{r}}\right]\right)} \geq \eta_{\text{req}}^{k}, \tag{4.11b}$$

$$K_{\ell}^{k}\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k} P_{k,n}^{\text{r}}\right] \leq P_{\max}^{k}, \tag{4.11c}$$

where $\alpha_k \equiv \dfrac{\theta_k T_{\text{f}} B_k}{\ln(2)}$, $P_{k,n}^{\text{r}} = \dfrac{P_{k,n}}{K_{\ell}^{k}}$, and $P_{\text{c}_{\text{r}}}^{k} = \dfrac{P_{\text{c}}^{k}}{K_{\ell}^{k}}$. Here $K_{\ell}^{k} = P_{\mathcal{L}}^{k}\eta_0 B_k$, which denotes the path loss factor, including both AWGN power and path loss power. Set $\hat{\eta}_{\text{req}}^{k} = K_{\ell}^{k}\eta_{\text{req}}^{k}$, and $\hat{P}_{\max}^{k} = P_{\max}^{k}/K_{\ell}^{k}$. Then, K_{ℓ}^{k} in (4.11a)-(4.11c) can be canceled to scale the system performance with respect to the path loss factor.

From (4.11a)-(4.11c), one can notice that the EC expression in a single-user multi-carrier system is not a linear summation of each subcarrier's achievable EC value.

Hence, the concavity and monotonicity of the EC function in a single-subcarrier system cannot be simply extended to the multi-carrier system. In order to find the joint energy and spectral efficient power allocation strategy in a single-user multi-carrier system, we start from analyzing the proposed problem $Q3$.

First, by referring to the scaled multi-carrier transmit power vector as $\boldsymbol{P}_k^{\mathrm{r}} = \left[P_{k,1}^{\mathrm{r}},\ P_{k,2}^{\mathrm{r}},\ \ldots\ P_{k,N}^{\mathrm{r}}\right]$, we note that the objective function (4.11a) is concave in $\boldsymbol{P}_k^{\mathrm{r}}$ [19]. Then, the link-layer EE, as the ratio of a concave function over a non-negative affine function in $\boldsymbol{P}_k^{\mathrm{r}}$, is a quasi-concave function in subcarrier power allocations [19]. Therefore, its upper contour set defined by (4.11b) is convex [59]. Hence, (4.11a)-(4.11c) is a concave optimization problem and the Karush-Kuhn-Tucker (KKT) conditions are both sufficient and necessary for the global optimum value. Furthermore, the proposed optimal power allocation strategy for the k^{th} user will be related to the joint PDF of the subcarrier power gains $\boldsymbol{\gamma}_k$, given by $\rho\left(\boldsymbol{\gamma}_k\right)$.

To solve the concave optimization problem (4.11a)-(4.11c), we start from analyzing the power-unconstrained problem (4.11a)-(4.11b), which paves the way for the power-constrained optimization problem. By transforming (4.11b) to

$$-\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{N_k}\left(1+N_k P_{k,n}^{\mathrm{r}}\gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right)-\hat{\eta}_{\mathrm{req}}^{k}\left(P_{\mathrm{c_r}}^{k}+\frac{1}{\epsilon}\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k}P_{k,n}^{\mathrm{r}}\right]\right)\geq 0, \tag{4.12}$$

the Lagrangian function can be given as follows

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{P}_k^{\mathrm{r}},\lambda\right)=&-\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{N_k}\left(1+N_k P_{k,n}^{\mathrm{r}}\gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right)\\&+\lambda\left(-\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{N_k}\left(1+N_k P_{k,n}^{\mathrm{r}}\gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right)\right.\\&\left.-\hat{\eta}_{\mathrm{req}}^{k}\left(P_{\mathrm{c_r}}^{k}+\frac{1}{\epsilon}\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k}P_{k,n}^{\mathrm{r}}\right]\right)\right)-\sum_{n=1}^{N_k}\mu_n P_{k,n}^{\mathrm{r}},\end{aligned} \tag{4.13}$$

where $\lambda\in R$ is the Lagrange multiplier associated to (4.12) and μ_n is the Lagrange multiplier associated to the constraint $P_{k,n}^{\mathrm{r}}\geq 0,\ \forall\ n\in\mathcal{N}_k$.

At the optimal power allocation, we have

$$\frac{\partial\mathcal{L}\left(\boldsymbol{P}_k^{\mathrm{r}},\lambda\right)}{\partial\boldsymbol{P}_k^{\mathrm{r}}}=0. \tag{4.14}$$

Because of the complementary slackness condition [59], if $P_{k,n}^{\mathrm{r}}>0$, then $\mu_n=0$, $\forall\ n\in\mathcal{N}_k$. On the other hand, if $P_{k,n}^{\mathrm{r}}=0,\ \exists\ n\in\mathcal{N}_k$, then $\mu_n\neq 0$. Thus, the following two cases need to be considered to find the optimal power allocation strategy.

4.3.2.1 Case 1: $P_{k,n}^{\mathrm{r}} > 0, \forall\ n \in \mathcal{N}_k$

In this case, all N_k subcarriers are allocated non-zero transmission power. Therefore, based on the complementary slackness, $\{\mu_n\}_{n=1}^{N_k} = 0$. Then, the KKT condition (4.14) can be simplified as

$$\prod_{i=1}^{N_k} \left(1 + N_k P_{k,i}^{\mathrm{r}} \gamma_{k,i}\right)^{-\frac{\alpha_k}{N_k}} = \frac{\beta}{\gamma_{k,n}} \left(1 + N_k P_{k,n}^{\mathrm{r}} \gamma_{k,n}\right),\ \forall\ n \in \mathcal{N}_k, \tag{4.15}$$

where $\beta = \dfrac{\lambda \hat{\eta}_{\mathrm{req}}^k}{\epsilon(\lambda+1)\log_2 \mathrm{e}} \mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{N_k} \left(1 + N_k P_{k,n}^{\mathrm{r}} \gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]$. By multiplying the right and left-hand sides of the N_k equations in (4.15), the optimal power allocation strategy for the k^{th} user on the n^{th} subcarrier can be obtained as

$$P_{k,n}^{\mathrm{r}} = \frac{1}{N_k}\left[\frac{1}{\beta^{\frac{1}{\alpha_k+1}} \prod_{i=1}^{N_k} \gamma_{k,i}^{\frac{\alpha_k}{(\alpha_k+1)N_k}}} - \frac{1}{\gamma_{k,n}}\right], \quad n \in \mathcal{N}_k. \tag{4.16}$$

The derived power allocation strategy (4.16) is optimal only when all subcarriers are assigned with positive powers. If there are one or more subcarriers which are allocated non-positive powers, then the second case needs to be taken into consideration.

4.3.2.2 Case 2: $P_{k,j}^{\mathrm{r}} = 0$, $\exists\ j \in \mathcal{N}_k$

If there exists $P_{k,j}^{\mathrm{r}} \leq 0$, then the set of subcarriers, which only positive powers are assigned, needs to be found.

Firstly, we define $\widehat{\mathcal{N}}_k = \left\{ n \in \mathcal{N}_k \middle| \frac{1}{N_k}\left[\frac{1}{\beta^{\frac{1}{\alpha_k+1}} \prod_{i=1}^{N_k} \gamma_{k,i}^{\frac{\alpha_k}{(\alpha_k+1)N_k}}} - \frac{1}{\gamma_{k,n}}\right] \geq 0 \right\}$. According to Lemma 1 in [20], the total power must be assigned to the subcarriers which belong to $\widehat{\mathcal{N}}_k$, while the subcarriers $n \notin \widehat{\mathcal{N}}_k$ should not be allocated any power. Therefore, a new power-unconstrained optimization problem could be expressed as

$$Q4: \max_{\substack{P_{k,n}^{\mathrm{r}} \geq 0 \\ n \in \widehat{\mathcal{N}}_k}} \quad -\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{\widehat{N}_k}\left(1 + N_k P_{k,n}^{\mathrm{r}} \gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right) \tag{4.17a}$$

$$\text{subject to}: \quad \frac{-\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{\widehat{N}_k}\left(1 + N_k P_{k,n}^{\mathrm{r}} \gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right)}{K_\ell^k\left(P_{\mathrm{c_r}}^k + \frac{1}{\epsilon}\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{\widehat{N}_k} P_{k,n}^{\mathrm{r}}\right]\right)} \geq \eta_{\mathrm{req}}^k, \tag{4.17b}$$

where $\widehat{N}_k = |\widehat{\mathcal{N}}_k|$ represents the cardinality of $\widehat{\mathcal{N}}_k$.

Therefore, if $P^{\mathrm{r}}_{k,n} > 0$, $\forall\ n \in \widehat{\mathcal{N}}_k$, then, the optimization problem can be solved exactly like Case 1. Otherwise, if there are subcarriers $n \in \widehat{\mathcal{N}}_k$ having $P^{\mathrm{r}}_{k,n} = 0$, then $\widehat{\mathcal{N}}_k$ must be further partitioned by recursively repeating the above process until a set $\mathcal{N}^*_k$ can be found, in which all subcarriers are allocated positive powers [19].

After obtaining $\mathcal{N}^*_k$, the optimal power allocations are computed as

$$P^{\mathrm{r}}_{k,n} = \begin{cases} \dfrac{1}{N_k}\left[\dfrac{1}{\beta^{\frac{N_k}{N_k+\alpha_k N^*_k}}\prod_{i\in\mathcal{N}^*_k}\gamma_{k,i}^{\frac{\alpha_k}{N_k+\alpha_k N^*_k}}} - \dfrac{1}{\gamma_{k,n}}\right], & n \in \mathcal{N}^*_k \\ 0, & \text{otherwise} \end{cases} \tag{4.18}$$

where $N^*_k = \mid \mathcal{N}^*_k \mid$.

The optimal value for β, referred to as β^*, can be found when the k^{th} user's EE constraint is satisfied with equality, yielding

$$-\frac{1}{\alpha_k}\log_2\left(\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\prod_{n=1}^{N_k}\left(1+N_k P^{\mathrm{r}}_{k,n}\gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}}\right]\right) - \hat{\eta}^k_{\mathrm{req}}\left(P^k_{\mathrm{c_r}} + \frac{1}{\epsilon}\mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k} P^{\mathrm{r}}_{k,n}\right]\right) = 0. \tag{4.19}$$

Further, the value of β^* which solves (4.19) will be found numerically, by applying Monte Carlo method and some root-finding algorithms. This is because that the closed-form expressions of the expectation values in (4.19) are too difficult to find, since the proposed optimal power allocation strategy on each subcarrier, given in (4.18), is dependent on the other subcarriers' CSI knowledge. Note that since EE versus EC is a bell shape curve, the required EE level, if possible, can be achieved at two different EC values, which means that there will be two solutions for β, i.e., β_1 and β_2, to satisfy (4.19). Assume that $\overline{P_{k1}} = \overline{P_k}\ |_{\beta=\beta_1}$, and $\overline{P_{k2}} = \overline{P_k}\ |_{\beta=\beta_2}$, where $\overline{P_k}$ stands for $K^k_\ell \mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k} P^{\mathrm{r}}_{k,n}\right]$. Therefore, the feasible set of the average input power level satisfying the EE constraint (4.11b) can be written as $\left[\overline{P_{k1}}, \overline{P_{k2}}\right]$[4]. Considering the intention to maximize EC and the fact that EC is a monotonically increasing function in $\overline{P_k}$ [19], the optimal average input power value $\overline{P^*_k}$, which solves the power-unconstrained problem (4.11a)-(4.11b), is chosen as the larger one which satisfies (4.19), i.e., $\overline{P^*_k} = \max\left(\overline{P_{k1}}, \overline{P_{k2}}\right)$. Based on the assumption that $\overline{P_{k2}}$ is larger than $\overline{P_{k1}}$, therefore $\overline{P^*_k} = \overline{P_{k2}}$, and correspondingly, $\beta^* = \beta_2$. Here, we complete the solving process of the optimal power allocation for the power-unconstrained problem (4.11a)-(4.11b).

[4]Without losing any generality, here we assume that $\overline{P_{k2}}$ is larger than $\overline{P_{k1}}$.

Table 4.2: Optimal Power Allocation Algorithm for a Single-User Multi-Carrier System

Input: $\left[\boldsymbol{\phi}, \theta_k, T_{\mathrm{f}}, B, N, N_k, P_{\mathrm{c}}^k, \epsilon, K_\ell^k, \boldsymbol{\gamma}_k, P_{\max}^k, \eta_{\mathrm{req}}^k\right]$

Step 1:

Have a initial guess of β.

Repeat

Create (4.19), using (4.16) or (4.18), which applies Monte Carlo method.

Update β using bisection method.

Until find β^* which solves (4.19).

Calculate $P_{k,n}^{\mathrm{r}}$, $n \in \mathcal{N}_k$, and $\overline{P_k^*} = K_\ell^k \mathbb{E}_{\boldsymbol{\gamma}_k}\left[\sum_{n=1}^{N_k} P_{k,n}^{\mathrm{r}}\right]\Big|_{\beta=\beta^*}$.

Step 2:

If $P_{\max}^k > \overline{P_k^*}$

Calculate E_{c}^k given in (4.5) and the link-layer EE^k value in (4.7).

Else

Create $\overline{P_k^*} = P_{\max}^k$ and update β^*, correspondingly.

Calculate $P_{k,n}^{\mathrm{r}}$, $n \in \mathcal{N}_k$, in (4.16) or (4.18).

Calculate E_{c}^k given in (4.5) and the link-layer EE^k value in (4.7).

End

Output: $\left[P_{k,n}^{\mathrm{r}}, \overline{P_k^*}, E_{\mathrm{c}}^k, \mathrm{EE}^k\right]$

By utilizing the above proposed optimal power allocation strategy, we start to analyze the optimization problem (4.11a)-(4.11c) with the average input power constraint. After the feasible set of the average power value for the EE constraint (4.11b) is found, the power-constrained EC maximization problem for the k^{th} user, subject to a link-layer EE constraint, can be simplified to

$$Q5: \max_{\substack{P_{k,n}^{\mathrm{r}} \geq 0 \\ n \in \mathcal{N}_k}} \quad -\frac{1}{\alpha_k} \log_2 \left(\mathbb{E}_{\boldsymbol{\gamma}_k} \left[\prod_{n=1}^{N_k} \left(1 + N_k P_{k,n}^{\mathrm{r}} \gamma_{k,n}\right)^{-\frac{\alpha_k}{N_k}} \right] \right) \tag{4.20a}$$

$$\text{subject to}: \quad \overline{P_k} \in \left[\overline{P_k}_1, \overline{P_k}_2\right], \tag{4.20b}$$

$$\overline{P_k} \leq P_{\max}^k. \tag{4.20c}$$

Since EC is a monotonically increasing function in $\overline{P_k}$ [19], therefore, the optimal average power value which solves the problem in (4.20a)-(4.20c) will be achieved at one of the three endpoint values, i.e., $\overline{P_k}_1$, $\overline{P_k}_2$, or $P_{\max}^k$. In more detail, if $\overline{P_k}_2 \leq P_{\max}^k$, it means that the k^{th} user has enough power to support the proposed optimal power allocation strategy given in (4.16). Therefore, in this case, the optimal power level $\overline{P_k^*}$, equals to $\overline{P_k}_2$, and the optimal power allocation strategy (4.16) will be applied and operated. On the other hand, if $\overline{P_k}_1 < P_{\max}^k < \overline{P_k}_2$, it means that $P_{\max}^k$ is too small to support the optimal situation. Hence, the system has to operate at $P_{\max}^k$ and

the optimal power allocation to solve (4.11a)-(4.11b) is according to (4.16), wherein, the optimal β^* is found such that $\overline{P_k^*}\,|_{\beta=\beta^*}= P_{\max}^k$. Moreover, if $P_{\max}^k < \overline{P_{k1}}$, it means that even the maximum available average power is too small to confirm the feasibility of the required EE value. Therefore, now the power-constrained problem $Q5$ has no feasible solution. To avoid this situation, we assume that each user's maximum available power is always sufficient to support the feasibility of its required EE value, i.e, $P_{\max}^k \geq \overline{P_{k1}}$.

To summarize, the Pseudocode of the optimal power allocation algorithm to solve the power-constrained link-layer EE-EC tradeoff problem for the k^{th} user, through multiple subcarriers, is illustrated in Table 4.2. After obtaining the optimal power allocation strategy and the optimal operational average power for problem $Q3$, further analysis is required to thoroughly understand and investigate the impact of the k^{th} user's circuit power value and EE requirement factor on its link-layer EE-EC tradeoff performance. Hence, we provide the following lemmas[5].

4.3.3 The Impact of P_{c}^k and χ_{EE}^k on the k^{th} User's EE-EC Tradeoff Performance

Lemma 6. *The k^{th} user's link-layer tradeoff EE value* $\text{EE}\left(\overline{P_k^*}\right)$ *decreases with* P_{c}^k.

Proof. The proof is provided in Appendix G. □

Lemma 6 indicates that if P_{c}^k becomes larger, the calculated link-layer EE for the k^{th} user decreases in this power-constrained EE-EC tradeoff problem, but its EC value will become larger, due to the conflict relationship between EE and EC. On the contrary, a smaller value of P_{c}^k will benefit its link-layer EE level, but the EC value will deteriorate.

Furthermore, the optimal tradeoff power value and the system performance can also be influenced by the introduced EE requirement factor. Specifically, when χ_{EE}^k increases, the required link-layer EE level increases. Therefore, the final operational link-layer EE value which satisfies the EE requirement equality increases. Since the proposed tradeoff average power operates at the EE-EC conflicting region, therefore the corresponding EC value will decrease due to the increase in EE level. Hence, we can obtain the following lemma.

Lemma 7. *The optimal average power value* $\overline{P_k^*}$ *monotonically decreases with* χ_{EE}^k, *but the corresponding link-layer tradeoff EE value* $\text{EE}\left(\overline{P_k^*}\right)$ *increases with* χ_{EE}^k.

[5]In these lemmas, the influence of $P_{\max}^k$ is ignored, by assuming that it is large enough to support the optimal power allocation strategy, i.e., $P_{\max}^k \geq \overline{P_k^*}$.

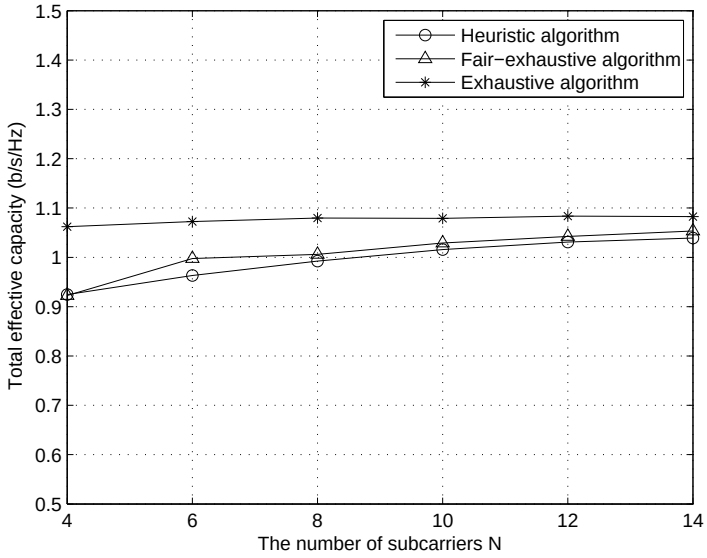


Figure 4.5: The total effective capacity versus the number of subcarriers N, for heuristic algorithm, exhaustive algorithm and fair-exhaustive algorithm.

Proof. The proof follows the above explanations and is omitted here. □

4.4 Simulation Results

In this section, we simulate the uplink transmission in a multi-user multi-subcarrier system, in which the fading statistics of different subcarriers are considered to be i.i.d. Rayleigh distributed such that the subcarrier power gains are realized as exponential random variables with unit mean. The performance of the exhaustive algorithm, the fair-exhaustive algorithm, and the heuristic algorithm on the total EC maximization problem, will be numerically evaluated and compared under the constraints of each user's link-layer EE requirement and average transmission power limit. To further analyze the problem and confirm the lemmas proposed in Section 4.3.3, the impact of delay QoS exponent, EE requirement factor and circuit-to-noise power ratio on each user's tradeoff EC value and the total EC performance is simulated and analyzed. In the following simulations, it is assumed that $B{\cdot}T_{\mathrm{f}} = 200$, the power amplifier efficiency $\epsilon = 1$, each user's individual average transmission power limit $P_{\max} = 10$dB, unless otherwise indicated.

In order to show the performance of the proposed heuristic algorithm, Fig. 4.5 shows the results of the total EC versus the number of subcarriers N, for the heuristic algorithm, the exhaustive algorithm and the fair-exhaustive algorithm. To get

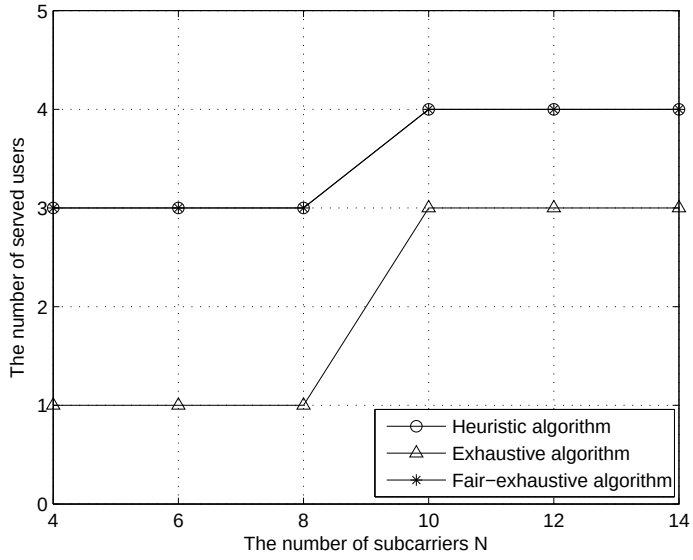


Figure 4.6: The number of served users versus the number of subcarriers N, for heuristic algorithm, exhaustive algorithm and fair-exhaustive algorithm.

Fig. 4.5, the number of users K is fixed, i.e., $K = 4$, in which all users have the same settings of EE requirement factor and circuit-to-noise power ratio, i.e., $\boldsymbol{\chi}_{\mathrm{EE}} = [0.7, 0.7, 0.7, 0.7]$, and $P_{\mathrm{cr}}^{k} = -10\mathrm{dB}$, $\forall k \in \mathcal{K}_0$. Here, $\boldsymbol{\chi}_{\mathrm{EE}}$ is the $1 \times K$ vector of the EE requirement factor for all K users. The delay QoS exponent vector $\boldsymbol{\theta}$ is given by $[10^{-3}, 10^{-3}, 10^{-2}, 10^{-2}]$. For the exhaustive algorithm, when the number of subcarriers N increases, the total EC does not change very much, due to the loose delay QoS requirements for all the users. For the fair-exhaustive algorithm and the heuristic algorithm, the total EC performance curves are very close. This indicates that the proposed heuristic algorithm not only has a low complexity and guarantees user fairness, but also offers a close-to-optimal performance.

To further compare the three algorithms, the plots for the number of served users versus the number of subcarriers N are included in Fig. 4.6. Although the exhaustive algorithm offers the best system performance in Fig. 4.5, Fig. 4.6 indicates that it serves the least number of users among all three algorithms. Especially, for the exhaustive algorithm, when $N \in [4, 8]$, it allocates all subcarriers to only one user, which shows a lack of fairness. On the contrary, for the heuristic algorithm and the fair-exhaustive algorithm, the number of served users shows an increasing trend until it equals to the total number of users K. This happens because the increase of N means more available frequency resources and the ability of supporting more users gradually increases.

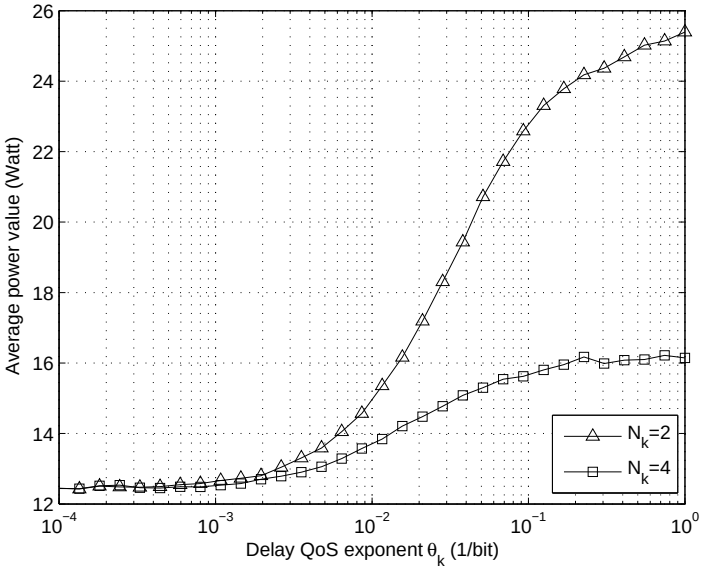


Figure 4.7: The optimal average tradeoff power value versus delay QoS exponent θ_k, for various values of N_k.

By considering the k^{th} user's multi-carrier system, Fig. 4.7 and Fig. 4.8 are plotted which respectively include the curves of the optimal average power[6] and the tradeoff EC value versus θ_k, for two different values of N_k, with $\chi^k_{\text{EE}} = 0.2$, and $P^k_{\text{c}_\text{r}} = -10\text{dB}$. Fig. 4.7 first shows that with a fixed N_k, when θ_k increases, the average power value increases. To explain this, we first recall that, the k^{th} user's EE requirement value is defined as a multiplication of χ^k_{EE} and $\eta^{k,N_k}_{\max}$, in which $\eta^{k,N_k}_{\max}$ is a function of θ_k and N_k. With the fixed values of N_k and χ^k_{EE}, $\eta^{k,N_k}_{\max}$ decreases with θ_k [19], and in turn, the EE requirement value decreases. Furthermore, the curve of link-layer EE versus average power becomes wider when the user's delay QoS exponent becomes more stringent [23]. Therefore, when θ_k increases, the optimal tradeoff average power obtained at a reduced EE requirement equality will become larger. Furthermore, Fig. 4.7 indicates that with a fixed value of θ_k, when N_k becomes larger, the average power value reduces. This is due to the fact that when the values of θ_k and χ^k_{EE} are fixed, $\eta^{k,N_k}_{\max}$ increases with N_k [19], as well as the required EE level. From Fig. 1 in [23], we note that a larger EE requirement will be satisfied at a smaller average power value. Hence, in this case, more available number of subcarriers lead to less average power consumption.

[6]Note that Fig. 4.7 only serves as a guideline, which shows the the trend of average power versus delay QoS exponent, for different settings of allocated subcarriers. The average power values given in Fig. 4.7 are not typical, as in reality these values need to be smaller.

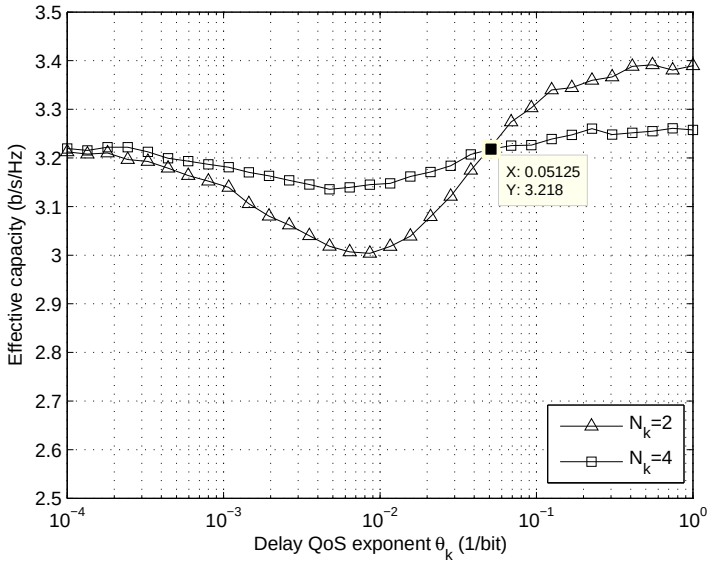


Figure 4.8: Effective capacity versus delay QoS exponent θ_k, for various values of N_k.

Fig. 4.8 shows the relationship between the k^{th} user's tradeoff EC value and θ_k for a single-user multi-carrier transmission system. This figure reveals two important conflicting situations and some insightful conclusions. Firstly, this figure indicates that one user's operational EC value will not show a monotonic trend with its delay QoS exponent, when there is a link-layer EE constraint. This phenomenon violates the monotonic trend of EC versus delay QoS exponent, in the EC-maximization situation provided in [20]. From [20], we note that for a fixed delay QoS exponent, the maximum EC increases monotonically with the transmission power. Also, for a fixed transmission power, the EC value monotonically decreases with the delay QoS exponent. However, in our case, when θ_k is small, the k^{th} user's link-layer EE requirement can be easily satisfied with a small value of transmission power. In contrast, when θ_k becomes stringent, the required EE value has to be satisfied with a very large power value, like the trend indicated in Fig. 4.7. In other words, the operational average power value will increase with θ_k. But, the increase of θ_k and the increase of the average power have a conflicting influence on the user's operational EC value. Therefore, with the inconsistent influence of these two parameters, EC will not show a monotonic trend, which can be confirmed from Fig. 4.8. Clearly, when θ_k is loose, the tradeoff EC value will be more influenced by θ_k. On the contrary, when θ_k becomes stringent, the average power dominates the situation, therefore the operational EC value shows an increasing trend, indicated from Fig. 4.8.

Secondly, Fig. 4.8 further reveals that one user's tradeoff EC value achieved at a

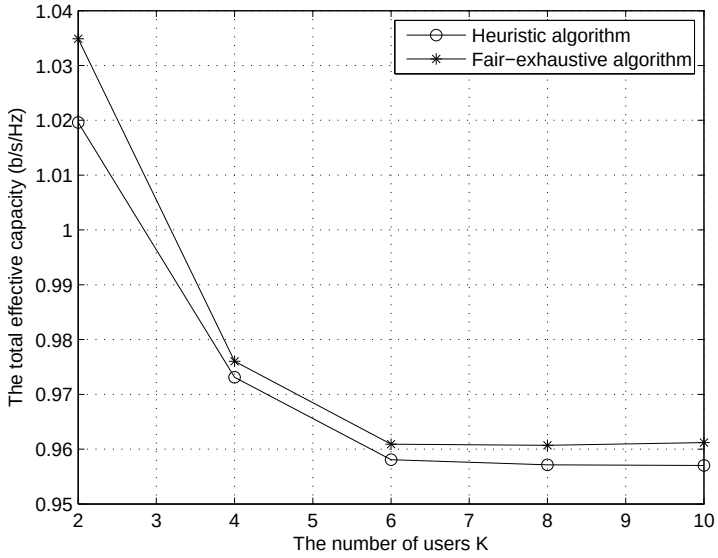


Figure 4.9: The total effective capacity versus the number of users K, for heuristic algorithm and exhaustive algorithm.

smaller number of subcarriers may be higher than the one obtained with relatively larger number of subcarriers, when there is a link-layer EE constraint. Specifically, when θ_k is loose, e.g., $\theta_k \in [10^{-4}, 0.05125]$, the tradeoff EC value with 4 subcarriers is higher than the one obtained with 2 subcarriers. When θ_k becomes stringent, e.g., $\theta_k \in [0.05125, 10^0]$, the tradeoff EC value achieved with 4 subcarriers is lower than the one obtained with 2 subcarriers. This phenomenon also violates the monotonic trend of EC versus the number of subcarriers in EC-maximization situation analyzed in [20]. This is due to the fact that with a link-layer EE requirement, when N_k increase, the average power value required to satisfy the EE constraint decreases. Since the increase of N_k and the corresponding decrease of the average power will have a conflicting influence on the user's operational EC value, then EC will not show a monotonic trend. Apparently, Fig. 4.8 indicates that when θ_k is loose, the tradeoff EC value will be more influenced by N_k. When θ_k becomes stringent, the average power dominates the situation, therefore the operational EC value follows the same trend with the average power. In conclusion, Fig. 4.7 and Fig. 4.8 indicate that when there is a link-layer EE requirement, each user's operational tradeoff EC value may not show a monotonic trend with its delay QoS exponent value or its available number of subcarriers.

To examine the effect of the number of users on a multi-user multi-carrier system with limited resources, Fig. 4.9 includes the plots for the total EC versus the number

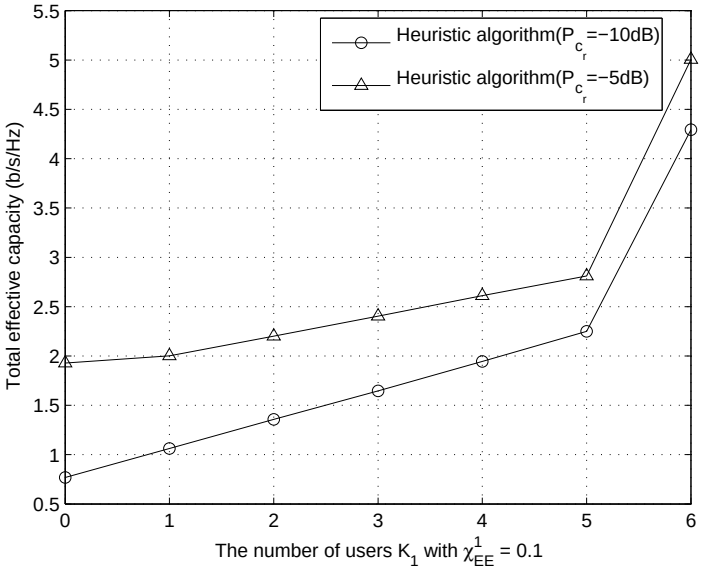


Figure 4.10: The total effective capacity versus the number of users K_1 in group $\mathcal{K}_1$, for various values of $P_{\mathrm{c_r}}$.

of users K, for the heuristic algorithm and the fair-exhaustive algorithm. Specifically, the total number of available subcarriers is fixed at $N = 10$. All users are assumed to have the same settings of circuit-to-noise power ratio, delay QoS exponent, and EE requirement factor, i.e., $P^k_{\mathrm{c_r}} = -10\mathrm{dB}$, $\theta_k = 10^{-2}$, and $\chi^k_{\mathrm{EE}} = 0.7$, $\forall k \in \mathcal{K}_0$. When the number of users K increases, the total EC values calculated from the two algorithms decrease and then stabilize when $K \geq 6$. This happens because, when K increases from 2 to 6, the number of served users increases and correspondingly, the number of subcarriers allocated to each served user decreases. Henceforth, the achievable EC for each served user reduces and the total EC value calculated from (4.6) decreases. When $K \geq 6$, the number of served users remains the same, due to the limited number of available subcarriers. Hence, the total EC value stays stable when K becomes greater than 6.

Assume that all K users, having the same delay QoS exponent and circuit-to-noise power ratio $P_{\mathrm{c_r}}$, are split into two groups, i.e., $\mathcal{K}_1$ and $\mathcal{K}_2$. In group $\mathcal{K}_1$, all K_1 users are required to have the same settings of the EE requirement factor, i.e., $\chi^1_{\mathrm{EE}} = 0.1$. Meanwhile, in group $\mathcal{K}_2$, all $K - K_1$ users are assumed to have a larger EE requirement factor values, i.e., $\chi^2_{\mathrm{EE}} = 0.8$. This indicates that the users in group $\mathcal{K}_1$ have looser EE requirements compared to the users in group $\mathcal{K}_2$. Set the total number of users $K = 6$, and the total number of subcarriers $N = 12$. Fig. 4.10 includes the plots for the results of the total EC versus the number of users K_1 in group $\mathcal{K}_1$, for

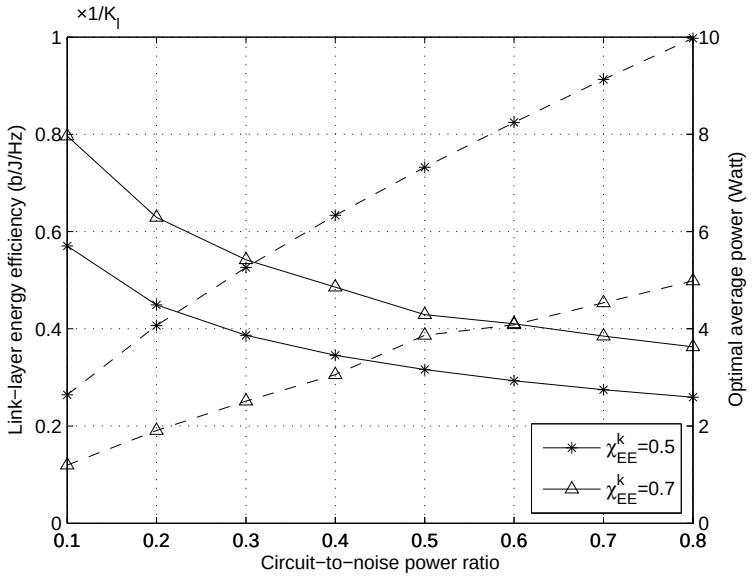


Figure 4.11: Link-layer energy efficiency and the optimal average power versus circuit-to-noise power ratio $P_{\mathrm{c_r}}^k$, for various values of χ_{EE}^k.

various values of circuit-to-noise power ratio $P_{\mathrm{c_r}}$. With a fixed $P_{\mathrm{c_r}}$, when K_1 increases from 0 to 6, the total EC value, in b/s/Hz, gradually increases. This is because when K_1 increases, the number of users with $\chi_{\mathrm{EE}}^1 = 0.1$ increases and correspondingly, the number of users with $\chi_{\mathrm{EE}}^2 = 0.8$ reduces. Note that for each user, a large value of EE requirement factor means that the user has a strict requirement on its link-layer EE value and will end up with a relatively small EC value. Therefore, when the number of users in group $\mathcal{K}_2$ reduces, the system can save more resource to benefit the total EC value, rather than sacrifice the system performance to support the strict EE requirements. When K_1 increases from 5 to 6, the number of users in group $\mathcal{K}_2$ reduces from 1 to 0 and the total EC value grows dramatically. This is due to the fact that in the check process of heuristic algorithm, the user having current minimum EC value will get the priority, which corresponds to the single user in group $\mathcal{K}_2$, when $K_1 = 5$. Therefore, in this case, the heuristic algorithm spends many resources on the user with $\chi_{\mathrm{EE}}^2 = 0.8$. When $K_1 = 6$, all users are in group $\mathcal{K}_1$ and they have the same loose EE requirements, i.e., $\chi_{\mathrm{EE}}^1 = 0.1$. Therefore, the system resources can be arranged evenly, which results in a great growth in the total EC value. Furthermore, from Fig. 4.10, we note that when $P_{\mathrm{c_r}}$ becomes larger, the system total EC value increases. Since a bigger value of $P_{\mathrm{c_r}}$ for all users will not change their relative difference, and correspondingly, will not change the subcarrier assignment solution, this phenomenon indicates that given a fixed subcarrier assignment, when one user's

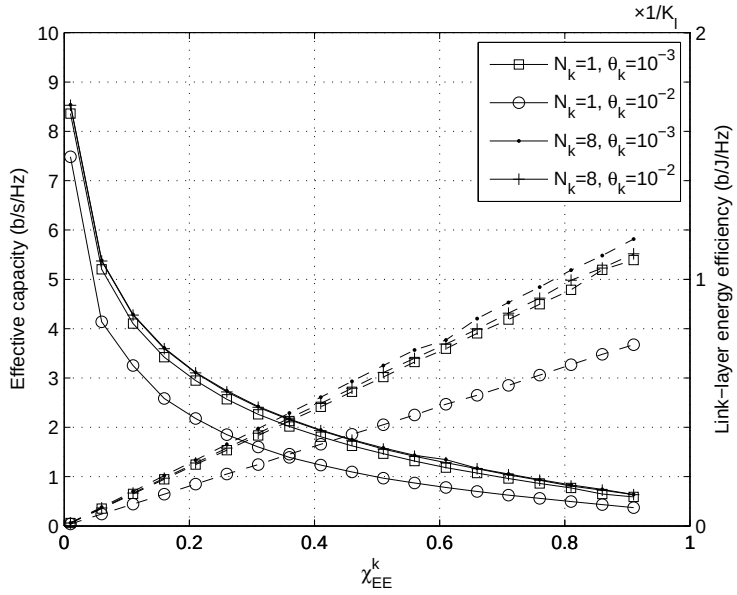


Figure 4.12: Effective capacity and link-layer energy efficiency versus $\chi^k_{\rm EE}$, for various values of θ_k and N_k.

circuit power increases, the total EC value will increase, as well as its own EC value.

To analyze the impact of the circuit-to-noise power ratio $P^k_{\rm c_r}$ and the EE requirement factor $\chi^k_{\rm EE}$ on the $k^{\rm th}$ user's multi-carrier system, Fig. 4.11 plots the results of the link-layer EE (on the left hand side (LHS) y-Axis, in solid lines) and the optimal tradeoff average power (on the right hand side (RHS) y-Axis, in dash lines) versus $P^k_{\rm c_r}$, for two different values of $\chi^k_{\rm EE}$, considering $N_k = 4$ and $\theta_k = 10^{-2}$. When $\chi^k_{\rm EE}$ is fixed, the link-layer EE value decreases and the optimal average power increases with $P^k_{\rm c_r}$, which confirms the proved Lemma 6. Furthermore, for a fixed $P^k_{\rm c_r}$, when $\chi^k_{\rm EE}$ becomes larger, the $k^{\rm th}$ user's link-layer EE value increases, but the optimal average power decreases, which confirms the proposed Lemma 7 in Section 4.3.3.

The plots of link-layer EE (on the RHS y-Axis, in dash lines) and EC (on the LHS y-Axis, in solid lines) versus $\chi^k_{\rm EE}$, for various values of delay QoS exponent θ_k and N_k, are included in Fig. 4.12. From this figure, we note that with fixed number of subcarriers N_k, when $\chi^k_{\rm EE}$ increases, EE increases. This confirms the proposed Lemma 7 in Section 4.3.3. Furthermore, with a fixed N_k, EC decreases with $\chi^k_{\rm EE}$. This is due to the fact that the tradeoff system operates in the conflicting region of EE and EC, therefore the EE-increases result from EC-reductions. Moreover, when $\chi^k_{\rm EE}$ and θ_k are fixed, as the number of subcarriers increases, both EC and EE increase. For a fixed N_k, when the delay QoS exponent θ_k increases from 10^{-3} to 10^{-2}, both EE and EC decrease. Especially, when $N_k = 1$, the decreases of EE

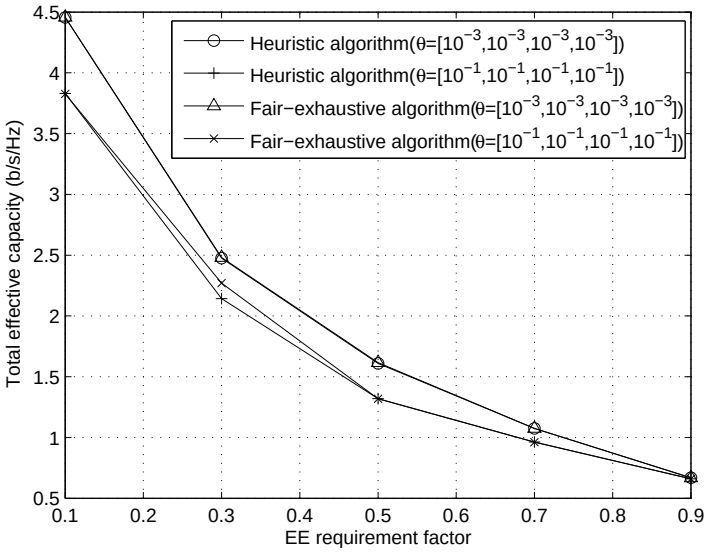


Figure 4.13: The total effective capacity versus EE requirement factor for different values of delay QoS exponent in heuristic algorithm, and fair-exhaustive algorithm.

and EC, as a result of the increase in θ_k, are significant. However, when N_k is larger, e.g., $N_k = 8$, the decreases of EE and EC are minor. This indicates that the multi-carrier communication system is more robust against delay requirements, in comparison with single-carrier communication systems. In other words, when the delay QoS requirement becomes more stringent, the multi-carrier system would sacrifice less EE and EC to guarantee the required delay constraint.

Assume that the total number of users $K = 4$ and the total number of available subcarriers $N = 8$. Specifically, all K users are assumed to have the same settings of delay QoS exponent and EE requirement factor. To further analyze and investigate the effect of EE requirement factor on the multi-user multi-carrier system, Fig. 4.13 includes the plots for the total EC versus EE requirement factor, with $P_{\mathrm{cr}} = -10\mathrm{dB}$ and two different values of $\boldsymbol{\theta}$, for the heuristic algorithm and the fair-exhaustive algorithm. When the EE requirement factor increases, the total EC value of the multi-user multi-carrier system decreases for the heuristic algorithm and the fair-exhaustive algorithm. Furthermore, when all users' delay requirements are loose, i.e., $\boldsymbol{\theta} = [10^{-3}, 10^{-3}, 10^{-3}, 10^{-3}]$, the EC curves of the two algorithms are exactly the same. This indicates that for a system having loose delay requirements, the difference of the total EC values calculated from the two algorithms is very small. When the delay QoS exponent values become larger, the total EC values become smaller, for both of the two algorithms. This is because, a larger value of delay QoS exponent represents a

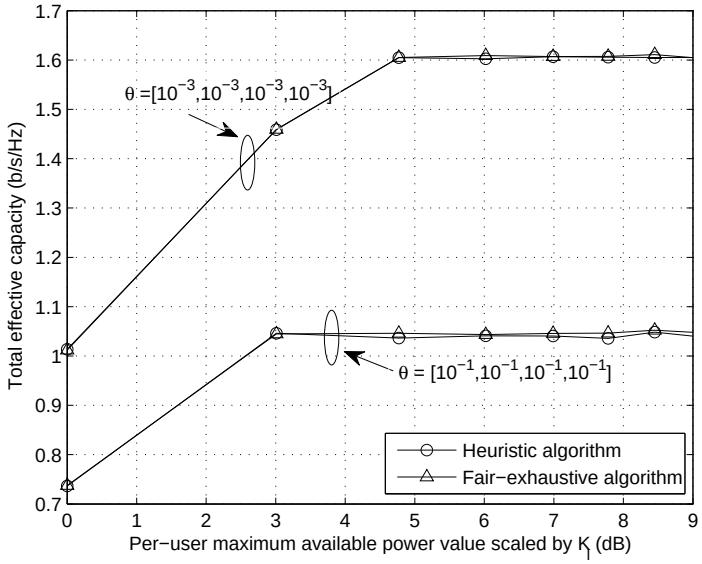


Figure 4.14: The total effective capacity versus maximum power value, for different values of $\boldsymbol{\theta}$.

more stringent delay requirement, therefore each user's maximum achievable arrival rate that it can support to maintain the target delay requirement, becomes small. Henceforth, the total system link-layer achievable rate reduces, correspondingly.

Considering that the per-user maximum available power value can influence the optimal results, Fig. 4.14 is included to show the performance of the heuristic algorithm and the fair-exhaustive algorithm, when the maximum available power value varies. The total number of subcarriers is fixed at $N = 6$, and the number of users is $K = 4$. All users are assumed to have the same settings of EE requirement factor, i.e., $\boldsymbol{\chi}_{\mathrm{EE}} = [0.5, 0.5, 0.5, 0.5]$, and the same values of circuit-to-noise power ratio, i.e., $P_{\mathrm{c_r}}^{k} = -10\mathrm{dB}$, $\forall k \in \mathcal{K}_0$. In addition, two different scenarios of delay QoS exponent vector $\boldsymbol{\theta}$ are included in Fig. 4.14, i.e., all elements in $\boldsymbol{\theta}$ are either 10^{-3} or 10^{-1}. Firstly, Fig. 4.14 shows that in both scenarios, the calculated EC values from the two algorithms are close, only with very little difference which makes the two curves difficult to distinguish. This confirms that the proposed heuristic algorithm indeed guarantees a close-to-optimal performance. Furthermore, when all users have more stringent delay QoS requirements, the total EC value reduces, which means that the value of EC needs to be sacrificed in this situation. More importantly, from Fig. 4.14, one can notice that for a fixed $\boldsymbol{\theta}$, the curves first increase, and then stabilize. This is because when the maximum available power value is too small to support the proposed optimal power value, the system has to operate at $P_{\max}$. Therefore, the

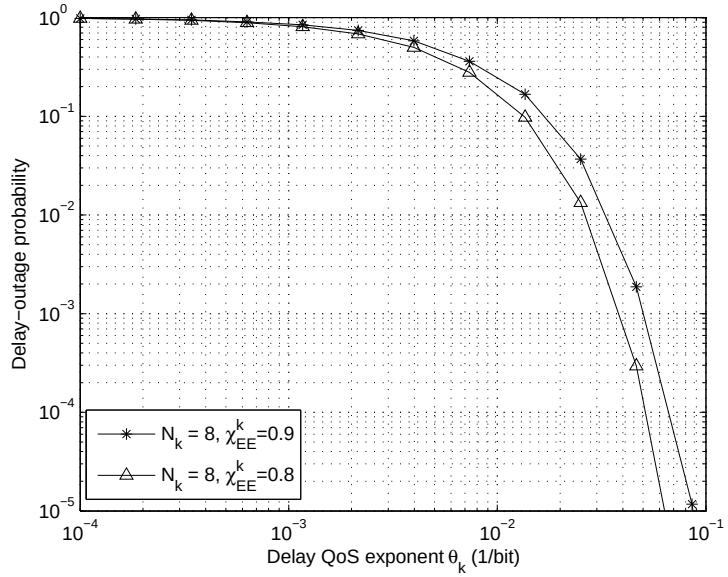


Figure 4.15: Delay-outage probability versus delay QoS exponent θ_k, for different values of χ_{EE}.

final calculated EC value is smaller in this case, since it is obtained at $P_{\max}$, rather than at the optimal power value. On the other hand, when the value of $P_{\max}$ becomes larger than the proposed optimal power value, then the system will operate at the optimal average power, which gives the optimal EC value, satisfying each user's EE constraint. To find detailed analysis, please refer to Section 4.3.2.

Fig. 4.15 plots the delay-outage probability for the k^{th} user, $P_{\mathrm{delay}}^{\mathrm{out}}$, versus delay QoS exponent θ_k, for various values of χ_{EE}^k with a maximum tolerable delay threshold $D_{\max} = 200$ and the circuit-to-noise power ratio $P_{\mathrm{c_r}}^k = -10$dB. This figure reveals that for the loose delay-constrained situations, e.g., $\theta_k = 10^{-4}$, the delay-outage probability values stay the same with different values of χ_{EE}^k. For more stringent delay-constrained situation, e.g., $\theta_k = 10^{-2}$, a smaller χ_{EE}^k ends up with lower delay-outage probability. This happens because a smaller χ_{EE}^k value means more sacrifice of EE from its maximum value, and in turn, results in more increase in its EC value. Therefore, the probability that the buffer length exceeds $D_{\max}$ decreases, henceforth, the delay-outage probability reduces.

4.5 Summary

A total EC maximization problem for the uplink transmission in a multi-user multi-carrier OFDMA system, was formulated as a combinatorial integer programming

problem, subject to each user's link-layer EE requirement as well as the individual's average transmission power limit. To solve this challenging resource allocation problem, it was first decoupled into a frequency provisioning problem and an independent multi-carrier link-layer EE-EC tradeoff problem for each user. A low-complexity heuristic algorithm was proposed, which not only offers close-to-optimal solutions, while serving as many users as possible, but also has a complexity linearly relating to the size of the problem. After obtaining the subcarrier assignment matrix, the multi-carrier link-layer EE-EC tradeoff problem for each user was formulated and solved by using KKT conditions. The per-user optimal power allocation strategy, across both frequency and time domains, was then derived. Further, the impact of the circuit power and the EE requirement factor on each user's tradeoff EE level and optimal average power value was theoretically investigated. Simulation results confirmed the proofs and design intentions, and further revealed that when there is a link-layer EE constraint, each user's tradeoff EC value may not monotonically decrease with its delay QoS provisioning, and the tradeoff EC value obtained with less subcarriers may be higher than the one achieved with more subcarriers.

In this chapter, a complex resource allocation problem was proposed and solved, including the subcarrier allocation and the optimal power allocation strategy, for the uplink transmission in a multi-user multi-carrier OFDMA system. In the next chapter, we then start to analyze the performance of the individual achievable link-layer rate and the total EC, under the per-user statistical delay QoS requirement, for a downlink non-orthogonal multiple access network with multiple users.

Chapter 5

Link-Layer Rate in a Downlink NOMA Network

5.1 Introduction

In this chapter, the achievable link-layer rate, namely, effective capacity (EC), is studied and investigated for a downlink non-orthogonal multiple access (NOMA) network with M users, under the per-user statistical delay quality-of-service (QoS) requirements. Specifically, the M users are assumed to be divided into multiple NOMA pairs. Conventional orthogonal multiple access (OMA) then is applied for inter-NOMA-pairs multiple access. Focusing on the total link-layer rate for a downlink M-user network, it is proved that OMA outperforms NOMA when the transmit signal-to-noise ratio (SNR) is small. On the contrary, simulation results show that NOMA prevails over OMA at high values of SNR. Aware of the importance of a two-user NOMA network, the impact of the transmit SNR and the delay QoS requirement on the individual EC performance and the total link-layer rate are theoretically investigated for a two-user network. Specifically, for delay-constrained and delay-unconstrained users, it is proved that for the user with the stronger channel condition in a two-user network, NOMA prevails over OMA when the transmit SNR is large. On the other hand, for the user with the weaker channel condition in a two-user network, NOMA outperforms OMA when the transmit SNR is small. Furthermore, for the user with the weaker channel condition, the individual EC in NOMA is limited to a maximum value, even if the transmit SNR goes to infinity. To confirm these insightful conclusions, the closed-form expressions for the individual EC in a two-user network, by applying NOMA or OMA, are derived for both users and then confirmed using Monte Carlo simulations.

The remainder of this chapter is organized as follows. The system model is given in Section 5.2. In Section 5.3, the theory of effective capacity is briefly introduced. Then, we start to analyze and investigate the individual EC and the total link-layer rate for a downlink NOMA network in Section 5.4, which includes the closed-form expressions for the link-layer rates in a two-user network, in NOMA and OMA scenarios, and the theoretical conclusions for a two-user network and a downlink NOMA network with mutiple NOMA pairs. Simulation results are given in Section 5.5, followed by conclusions in Section 5.6.

5.2 System Model

We consider a cellular downlink transmission with one base station (BS) and M single-antenna users. At the BS, the upper layer packets are organized into frames, which are then stored at the transmit buffer[1], in the link layer. After split into bit streams, these frames will be transmitted through the allocated channel. According to the NOMA principle, the BS will send $\sum_{k=1}^{M} \sqrt{\alpha_k P} s_k$ to the destinations, where s_k is the message for the k^{th} user, P is the total transmission power, and α_k denotes the power allocation coefficient for the k^{th} user.

As for each wireless channel from the BS to an individual user, it is assumed to be block fading with a bandwidth of B, i.e., the channel gain is invariant during each fading-block, but independently varies from one fading-block to another. The length of each fading-block, denoted by T_{f}, is assumed to be an integer multiple of the symbol duration T_{s}. Meanwhile, the duration of one frame size is assumed to be equal to the length of the fading-block, i.e., T_{f}. The channel gain between the BS and the k^{th} user is denoted by ${h_k}^2$, which is modeled according to Rayleigh fading distribution. Without loss of generality, the users' channels are assumed to be sorted so that $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_M|^2$, which indicates that the k^{th} user always holds the k^{th} weakest channel. Henceforth, based on the NOMA protocol, the power coefficients can be ordered as $\alpha_1 \geq \cdots \geq \alpha_M$, and $\sum_{k=1}^{M} \alpha_k = 1$ [27].

The received signal at the k^{th} user is given by $y_k = h_k \sum_{l=1}^{M} \sqrt{\alpha_l P} s_l + n_k$, where n_k denotes the additive white Gaussian noise. By applying the successive interference cancellation (SIC) technique, the k^{th} user will detect the i^{th} user's message, when

[1]Here, it is assumed that the BS offers one virtual buffer for every served user.

[2]The time index t is omitted because the channel gains are assumed to be stationary and ergodic random processes.

$i < k$, and then remove the i^{th} user's message from its received signal, in a successive manner [27]. The message for the j^{th} user, for $j > k$, however, will be treated as noise at the k^{th} user. Note that the condition under which the k^{th} user can successfully decode the i^{th} user's message is to satisfy $R_{i\to k} \geq \tilde{R}_i$ [109]. Here, $\tilde{R}_i$ is the i^{th} user's target data rate, and $R_{i\to k}$ denotes the k^{th} user's data rate to detect the i^{th} user's message, i.e., $R_{i\to k} = \log_2\left(1 + \frac{\rho|h_k|^2\alpha_i}{\rho|h_k|^2\sum_{l=i+1}^{M}\alpha_l + 1}\right)$, where ρ denotes the transmit SNR, i.e., $\rho = \frac{P}{N_0 B}$, with $N_0 B$ indicating the noise power. Assume that $\tilde{R}_i$ is determined opportunistically by the i^{th} user's channel condition [109], i.e., $\tilde{R}_i = R_i = \log_2\left(1 + \frac{\rho|h_i|^2\alpha_i}{\rho|h_i|^2\sum_{l=i+1}^{M}\alpha_l + 1}\right)$, which means that its target rate equals to the data rate achieved when it decodes its own message. Hence, it is easy to verify that the condition $R_{i\to k} \geq \tilde{R}_i$ always holds since $|h_k|^2 \geq |h_i|^2$, for $k > i$.

Consequently, the achievable data rate[3], in b/s/Hz, for the k^{th} user in a downlink NOMA network, can be formulated as

$$R_k = \log_2\left(1 + \frac{\rho|h_k|^2\alpha_k}{\rho|h_k|^2\sum_{l=k+1}^{M}\alpha_l + 1}\right). \tag{5.1}$$

5.3 Effective Capacity

Let us take the k^{th} user as an example. At the BS, considering the dynamic queueing system for the k^{th} user, we assume that the buffer size is infinite and the link can serve $R_k(t)$ packets per unit of time, which means that the capacity of the link at time t is $R_k(t)$. From Chapter 2.1, we note that by using the large deviation theory, the buffer overflow probability satisfies (2.16). When the focus is on the delay experienced by a source packet arriving at time t, defined by $D(t)$, the delay-outage probability $P_{\text{delay}}^{\text{out}}$ can be given in (2.27). Then, by assuming that Gärtner-Ellis limit exists, effective capacity represents the maximum arrival rate that a link can support, on the condition that a required delay QoS metric is satisfied [10].

By recalling that the wireless channel from the BS to the k^{th} user follows a block fading distribution, hence, the EC of the k^{th} user can be formulated as [10, 86]

$$E_{\text{c}}^{k} = -\frac{1}{\theta_k T_{\text{f}} B}\ln\left(\mathbb{E}\left[e^{-\theta_k T_{\text{f}} B R_k}\right]\right), \quad \text{(b/s/Hz)}. \tag{5.2}$$

[3]The distance-based path-loss is assumed to be uniform for each user.

Here, $\mathbb{E}[\cdot]$ indicates the expectation over the probability density function (PDF) of the allocated channel. Then, by inserting (5.1) into (5.2), the achievable link-layer rate for the k^{th} user in a downlink NOMA network can be obtained, yielding

$$E_{\text{c}}^{k} = -\frac{1}{\theta_k T_{\text{f}} B} \ln \left(\mathbb{E} \left[\left(1 + \frac{\rho |h_k|^2 \alpha_k}{\rho |h_k|^2 \sum_{l=k+1}^{M} \alpha_l + 1} \right)^{-\frac{\theta_k T_{\text{f}} B}{\ln 2}} \right] \right). \tag{5.3}$$

Furthermore, we note that the parameter θ_k ($\theta_k > 0$) denotes the exponential decay rate of the delay-outage probability, for the k^{th} user. A smaller θ_k represents a slower decay rate, which indicates that the user can tolerate a loose delay QoS guarantee, while a larger θ_k means that a more stringent delay QoS guarantee is required [10, 86]. Specifically, when $\theta_k \to 0$, it indicates that the k^{th} user has no delay requirement. When $\theta_k \to \infty$, it means that the k^{th} user has an extremely stringent delay requirement [110].

5.4 Effective Capacity in a Downlink NOMA Network

Aware of the difficulty of deriving the closed-form expression for the individual EC in (5.3) when all M users transmit on the same channel, we start to investigate the situation when there are multiple NOMA pairs in a M-user network. Specifically, the M users are assumed to be divided into $\frac{M}{2}$ groups[4], so that within each group, NOMA will be implemented for only two users, and the conventional OMA can be used for inter-NOMA-pairs multiple access [27]. Furthermore, a two-user downlink version of NOMA, called the multiuser superposition transmission (MUST), has been proposed for the Third Generation Partnership Project Long Term Evolution Advanced (3GPP-LTE-A) networks [111]. Inspired by this, we first focus on the link-layer rate performance of a two-user downlink NOMA network, which itself is of great importance, and also paves the way for the performance analysis of multiple NOMA pairs. Closed-form expressions and insightful theoretical conclusions are first provided. Finally, based on the proposed derivations and theoretical insights, the total EC for the multiple NOMA pairs is derived and investigated, in comparison with the total EC for M OMA users.

[4]To achieve this, M is assumed to be an even positive number.

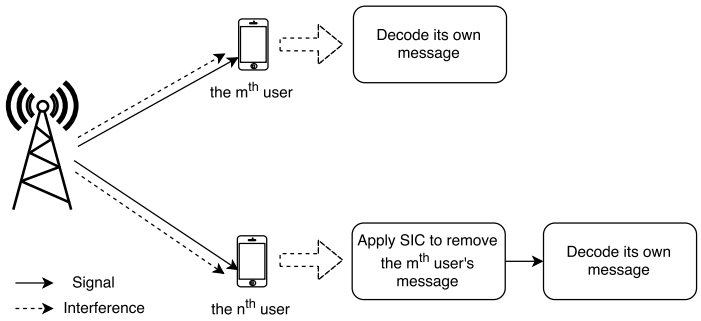


Figure 5.1: Two-user downlink NOMA network.

5.4.1 Effective Capacity in a Two-user NOMA Network

Without loss of generality, the m^{th} user and the n^{th} user, $m < n$, are assumed to be paired together as a two-user NOMA network, as depicted in Fig. 5.1. By applying the SIC strategy, the n^{th} user, which has the relatively stronger channel condition, will first decode the message of the user with the weaker channel condition, i.e., the m^{th} user, and then decode its own message by removing the m^{th} user's message. On the other hand, the m^{th} user with the weaker channel condition, will decode its own message by treating the n^{th} user's information as noise. In order to make sure that SIC can be correctly carried out at the n^{th} user, it is required that $R_{m\to n} \geq R_m$, i.e., $\log_2\left(1+\frac{\rho\alpha_m|h_n|^2}{\rho\alpha_n|h_n|^2+1}\right) \geq R_m$. According to the analysis in Section 5.2, it is noted that this always holds since $|h_n|^2 \geq |h_m|^2$, for $n > m$.

By applying the fixed power allocation, the power allocation coefficients for the m^{th} user and the n^{th} user are denoted by α_m and α_n, respectively, where $\alpha_m \geq \alpha_n$, and $\alpha_m + \alpha_n = 1$, according to the NOMA principle. By assuming that both users experience the same strength of additive white Gaussian noise, then the achievable data rates, in b/s/Hz, for the m^{th} user and the n^{th} user in a two-user NOMA network, are respectively formulated as

$$R_m = \log_2\left(1+\frac{\rho\alpha_m|h_m|^2}{\rho\alpha_n|h_m|^2+1}\right), \tag{5.4a}$$

$$R_n = \log_2\left(1+\rho\alpha_n|h_n|^2\right). \tag{5.4b}$$

On the other hand, if the m^{th} user and the n^{th} user each have their message transmitted using OMA scheduling, e.g., time division multiple access (TDMA), with total transmit SNR ρ, the achievable data rate of each user can then be given by

$$\bar{R}_i = \frac{1}{2}\log_2\left(1+\rho|h_i|^2\right),\ i \in \{m,n\} \tag{5.5}$$

where $\frac{1}{2}$ denotes that each user has only half of the available radio resources in OMA networks. Considering the duration of one frame as one time slot, (5.5) implies that in TDMA networks, each user can only occupy half of the time slot to transmit, while in the other half time slot, it will stay silent[5].

Assuming that the Gärtner-Ellis theorem [52, Pages 34-36] is satisfied, the expressions of EC for the m^{th} user and the n^{th} user in a block fading channel can be respectively given as [10]

$$E_{\text{c}}^{m} = -\frac{1}{\theta_m T_{\text{f}} B} \ln\left(\mathbb{E}\left[e^{-\theta_m T_{\text{f}} B R_m}\right]\right) \quad \text{(b/s/Hz)}, \tag{5.6a}$$

$$E_{\text{c}}^{n} = -\frac{1}{\theta_n T_{\text{f}} B} \ln\left(\mathbb{E}\left[e^{-\theta_n T_{\text{f}} B R_n}\right]\right) \quad \text{(b/s/Hz)}. \tag{5.6b}$$

By inserting (5.4a) into (5.6a) and inserting (5.4b) into (5.6b), we then get that

$$E_{\text{c}}^{m} = -\frac{1}{\theta_m T_{\text{f}} B} \ln\left(\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]\right), \tag{5.7a}$$

$$E_{\text{c}}^{n} = -\frac{1}{\theta_n T_{\text{f}} B} \ln\left(\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]\right), \tag{5.7b}$$

where $\beta_m = -\frac{\theta_m T_{\text{f}} B}{2\ln 2}$, and $\beta_n = -\frac{\theta_n T_{\text{f}} B}{2\ln 2}$.

For an OMA scheme, such as TDMA, the EC expressions for both users can be calculated by inserting (5.5) into (5.6a) and (5.6b), which yield to

$$\bar{E}_{\text{c}}^{m} = -\frac{1}{\theta_m T_{\text{f}} B} \ln\left(\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]\right), \tag{5.8a}$$

$$\bar{E}_{\text{c}}^{n} = -\frac{1}{\theta_n T_{\text{f}} B} \ln\left(\mathbb{E}\left[\left(1+\rho|h_n|^2\right)^{\beta_n}\right]\right). \tag{5.8b}$$

In the following subsection, the closed-form expressions for the link-layer rates are first derived for both users, in NOMA and OMA, i.e., E_{c}^{m}, $\bar{E}_{\text{c}}^{m}$, E_{c}^{n}, and $\bar{E}_{\text{c}}^{n}$. Further, the impact of the transmit SNR ρ and the per-user delay QoS exponent, on the individual EC performance and the total link-layer rates, in both NOMA and OMA scenarios, will be investigated and analyzed for the two-user network.

5.4.1.1 The Closed-Form Expressions for the Individual EC in a Two-user System

Suppose that $h_1, \ldots, h_M$ are M unordered independent channel gains, modeled according to the unit-variance Rayleigh fading distribution. Set $\gamma_m = \rho|h_m|^2$ and

[5]Note that the way of equally allocating resource is a typical and special case. However, the influence of different resource allocation strategies is beyond the scope of this chapter.

$\gamma_n = \rho|h_n|^2$. When γ_m and γ_n are unordered, the PDF of γ_m and γ_n is denoted by $f(\gamma_m)$ and $f(\gamma_n)$, respectively. Correspondingly, the cumulative distribution function (CDF) of the unordered γ_m and γ_n can be denoted by $F(\gamma_m)$, and $F(\gamma_n)$. Since the unordered channel gains are assumed to be statistically independent and identically distributed, hence, one can notice that $f(\gamma_m) = f(\gamma_n)$, and $F(\gamma_m) = F(\gamma_n)$, $\forall\ m,\ n \in \{1, \ldots, M\}$. However, when the users' channels are assumed to be sorted so that $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_M|^2$, the order statistics of different channel power gains will not be the same. In NOMA networks, the users are ordered first according to their channel conditions, therefore the statistical features of the ordered channel power gains fall into the scope of the order statistics [112]. The PDF of the ordered γ_m and γ_n, where $\gamma_m \leq \gamma_n$, are denoted by $f_{(m)}(\gamma_m)$, and $f_{(n)}(\gamma_n)$, respectively. From order statistics [112], $f_{(m)}(\gamma_m)$ and $f_{(n)}(\gamma_n)$ are given by

$$f_{(m)}(\gamma_m) = \psi_m f(\gamma_m) F(\gamma_m)^{m-1} \left(1 - F(\gamma_m)\right)^{M-m}, \tag{5.9a}$$

$$f_{(n)}(\gamma_n) = \psi_n f(\gamma_n) F(\gamma_n)^{n-1} \left(1 - F(\gamma_n)\right)^{M-n}, \tag{5.9b}$$

where $\psi_m = \dfrac{1}{B(m, M-m+1)}$, $\psi_n = \dfrac{1}{B(n, M-n+1)}$, in which $B(a,b)$ denotes the beta function, according to $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ [100]. Here, $\Gamma(a) = a!$, as a is a positive integer.

Theorem 7. *For the m^{th} user, the closed-form expression for the EC in NOMA, E_c^m, is given in* (5.10)*. Meanwhile, the EC in OMA, $\bar{E}_c^m$, can be expressed in closed-form, given in* (5.11)*.*

Proof. The proof is provided in Appendix H. □

$$E_{\mathrm{c}}^{m} = -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\alpha_n^{-2\beta_m} \psi_m}{\rho} \left(\sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k \frac{\rho}{M-m+1+k} \right.\right.$$
$$+ \frac{\theta_m(\alpha_n - 1)}{\alpha_n \ln 2} \sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k e^{\frac{M-m+1+k}{\rho\alpha_n}} E_i\left(-\frac{M-m+1+k}{\rho\alpha_n}\right)$$
$$+ \sum_{j=2}^{\infty} \binom{2\beta_m}{j} \left(\frac{\alpha_n - 1}{\alpha_n}\right)^j \sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k \left(\frac{\sum_{i=1}^{j-1} \frac{(i-1)!}{\alpha_n^{-i}} \left(-\frac{M-m+1+k}{\rho}\right)^{j-i-1}}{(j-1)!} \right.$$
$$\left.\left.\left.\left. - \frac{\left(-\frac{M-m+1+k}{\rho}\right)^{j-1}}{(j-1)!} e^{\frac{M-m+1+k}{\rho\alpha_n}} E_i\left(-\frac{M-m+1+k}{\rho\alpha_n}\right) \right) \right) \right) \right), \quad (5.10)$$

$$\bar{E}_{\mathrm{c}}^{m} = -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\psi_m}{\rho} \sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k U\left(1, 2+\beta_m, \frac{M-m+1+k}{\rho}\right) \right), \quad (5.11)$$

where $E_i(\cdot)$ is the exponential integral, and $U(a, b, z)$ is the confluent hypergeometric function of the second kind [100].

Theorem 8. *For the n^{th} user, the closed-form expression for the EC in NOMA, E_c^n, is given in* (5.12)*. Meanwhile, the EC in OMA, $\bar{E}_c^n$, can be expressed in closed-form, given in* (5.13)*.*

Proof. The proof is omitted here due to the page limit, but can be found by following similar steps as in Appendix H. □

$$E_{\mathrm{c}}^{n} = -\frac{1}{\theta_n T_{\mathrm{f}} B} \ln \left(\frac{\psi_n}{\rho\alpha_n} \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k U\left(1, 2+2\beta_n, \frac{M-n+1+k}{\rho\alpha_n}\right) \right), \quad (5.12)$$

$$\bar{E}_{\mathrm{c}}^{n} = -\frac{1}{\theta_n T_{\mathrm{f}} B} \ln \left(\frac{\psi_n}{\rho} \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k U\left(1, 2+\beta_n, \frac{M-n+1+k}{\rho}\right) \right). \quad (5.13)$$

The accuracy of the above closed-form expressions will be confirmed by comparing with Monte Carlo simulations in Section 5.5. Then, we start to investigate the impact of the transmit SNR ρ and the per-user delay QoS exponents θ_m, θ_n, on the individual

EC performance and the total link-layer rate for a two-user network, in both NOMA and OMA scenarios. Two cases are deliberately analyzed in the following subsections, i.e., Case 1: consider delay-constrained users[6]; Case 2: consider delay-unconstrained users. Note that Case 2 is an extreme case of no delay, in which the individual EC is proved to be equivalent to ergodic capacity[7]. Interestingly, the theoretical and simulation results obtained for this case are indeed novel and not found in the current literature. Further, by including Case 1 and Case 2, the performance of a two-user downlink NOMA network, either delay-constrained or delay-unconstrained, can be comprehensively analyzed and investigated.

5.4.1.2 Case 1: Consider Delay-Constrained Users

Lemma 8. *Considering the individual EC in NOMA and OMA, for both users, we prove that*

(a) When $\rho \to 0$, $E_c^m \to 0$, $\bar{E}_c^m \to 0$, $E_c^m - \bar{E}_c^m \to 0$, $E_c^n \to 0$, $\bar{E}_c^n \to 0$, *and* $E_c^n - \bar{E}_c^n \to 0$.

(b) When $\rho \to \infty$[8], $\lim\limits_{\rho\to\infty} E_c^m = \log_2\left(\dfrac{1}{\alpha_n}\right)$, $\lim\limits_{\rho\to\infty} \bar{E}_c^m \to \infty$, *and* $\lim\limits_{\rho\to\infty} \left(E_c^m - \bar{E}_c^m\right) \to -\infty$.

(c) When $\rho \to \infty$, $\lim\limits_{\rho\to\infty} E_c^n \to \infty$, $\lim\limits_{\rho\to\infty} \bar{E}_c^n \to \infty$, *and* $\lim\limits_{\rho\to\infty} \left(E_c^n - \bar{E}_c^n\right) \to \infty$.

Proof. The proof is provided in Appendix I. □

From Lemma 8.(a), it shows that, for both users, either in NOMA or OMA, their individual rates start at the same initial value of 0, at small values of ρ. Lemma 8.(b), on the other hand, indicates that for the weaker user[9], when $\rho \to \infty$, its EC achieved by applying NOMA is limited by $\log_2\left(\dfrac{1}{\alpha_n}\right)$. This means that in a two-user NOMA network, the weaker user can only achieve a limited EC, no matter how large the transmit SNR can be. On the contrary, for the stronger user[10], Lemma 8.(c) indicates that when $\rho \to \infty$, its achievable EC in NOMA approaches infinity. Furthermore, Lemma 8.(b) and Lemma 8.(c) reveal that when $\rho \to \infty$, the EC values achieved by applying OMA approach infinity, for both of the two users.

[6]In this case, finite values of the delay QoS exponents θ_m, θ_n are considered.

[7]The proof and further explanations can be found in Lemma 13 in Section 5.4.1.3.

[8]Note that $\rho \to \infty$ is not practical, but this is only to provide a guideline. In the simulation results provided in Section 5.5, it shows that the conclusions proved for the case of $\rho \to \infty$, are valid for values of ρ as big as $\rho = 30$dB.

[9]Hereafter, the user with the weaker channel condition is referred to as the weaker user.

[10]Hereafter, the user with the stronger channel condition is referred to as the stronger user.

Apparently, Lemma 8 only considers two extreme cases of ρ for both users. Henceforth, from Lemma 8, one cannot know how the individual EC will change with respect to ρ on general terms. Will NOMA be always better than OMA for the n^{th} user, at any positive values of ρ? Will OMA be always better than NOMA for the m^{th} user, for any settings of ρ? To answer these questions and to further analyze the impact of ρ on the individual EC, in a two-user NOMA network and in a two-user OMA network, we provide the following lemmas.

Lemma 9. *Considering the m^{th} user's EC, in NOMA and OMA, we prove that*

(a) At any values of ρ, $\dfrac{\partial E_c^m}{\partial \rho} \geq 0$, and $\dfrac{\partial \bar{E}_c^m}{\partial \rho} \geq 0$.

(b) When $\rho \to 0$, $\lim\limits_{\rho \to 0} \dfrac{\partial \left(E_c^m - \bar{E}_c^m\right)}{\partial \rho} = \left(\dfrac{\frac{1}{2} - \alpha_n}{\ln 2}\right) \mathbb{E}\left[|h_m|^2\right] \geq 0$.

(c) When ρ is very large, $\dfrac{\partial \left(E_c^m - \bar{E}_c^m\right)}{\partial \rho} \leq 0$, and it approaches 0 when $\rho \to \infty$.

Proof. The proof is provided in Appendix J. □

From Lemma 9.(a), it shows that for the weaker user, its achievable EC, in NOMA or OMA, is always non-decreasing with the transmit SNR. Furthermore, Lemma 9.(b) indicates that, for the weaker user, when the transmit SNR is very small, the EC in NOMA has a faster increasing speed than that in OMA. On the contrary, Lemma 9.(c) shows that for the weaker user, when the transmit SNR is very large, the EC in OMA increases faster than that in NOMA.

To further explain the above theoretical conclusions, the focus lies on analyzing the EC difference between NOMA and OMA, for the weaker user in a two-user system. From Lemma 8 and Lemma 9, one can conclude that, $E_{\text{c}}^m - \bar{E}_{\text{c}}^m$ starts at the initial value of 0, first increases, and at the end decreases to $-\infty$ with a gradually diminishing speed. This means that, for the weaker user, NOMA can achieve higher EC than OMA, at small values of ρ. When the transmit SNR becomes extremely large, OMA is more beneficial than NOMA, for the weaker user. Finally, when $\rho \to \infty$, the performance gain of OMA over NOMA becomes stable.

Lemma 10. *Considering the n^{th} user's EC, in NOMA and OMA, we prove that*

(a) At any values of ρ, $\dfrac{\partial E_c^n}{\partial \rho} \geq 0$, and $\dfrac{\partial \bar{E}_c^n}{\partial \rho} \geq 0$.

(b) When $\rho \to 0$, $\lim_{\rho \to 0} \frac{\partial \left(E_c^n - \bar{E}_c^n\right)}{\partial \rho} = \left(\frac{\alpha_n - \frac{1}{2}}{\ln 2}\right) \mathbb{E}\left[|h_n|^2\right] \leq 0.$

(c) When ρ *is very large,* $\frac{\partial \left(E_c^n - \bar{E}_c^n\right)}{\partial \rho} \geq 0$, *and it approaches 0 when* $\rho \to \infty$.

Proof. The proof is provided in Appendix K. □

From Lemma 10.(a), it shows that for the stronger user, its achievable EC, in NOMA or OMA, has a non-decreasing trend with the transmit SNR. Furthermore, Lemma 10.(b) indicates that, for the stronger user, when the transmit SNR is very small, the EC in OMA increases faster than that in NOMA. On the contrary, Lemma 10.(c) shows that when the transmit SNR becomes very large, the EC in NOMA increases faster than the one in OMA, for the stronger user.

Then we start to analyze the range of ρ, in which NOMA is more beneficial than OMA, for the stronger user in a two-user system. From Lemma 8 and Lemma 10, one can conclude that, $E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n$ starts at the initial value of 0, first decreases, and finally increases to ∞ with a gradually reducing speed. This means that, for the stronger user, OMA achieves higher EC than NOMA, when the transmit SNR is small. At high values of ρ, NOMA becomes more beneficial than OMA, for the stronger user. Finally, when $\rho \to \infty$, the performance gain of NOMA over OMA becomes stable, for the stronger user.

In order to investigate the impact of the transmit SNR ρ on the performance of the total link-layer achievable rate, we define $\mathrm{T}_N = E_{\mathrm{c}}^m + E_{\mathrm{c}}^n$, which indicates the total EC for the two-user NOMA network. Meanwhile, we define $\mathrm{T}_O = \bar{E}_{\mathrm{c}}^m + \bar{E}_{\mathrm{c}}^n$, which denotes the total achievable link-layer rate for the two-user OMA system. Note that the total link-layer achievable rate for a two-user system is defined as a linear summation of the two users' EC values. This is due to the reason that for each user, there is a dynamic queueing system with an infinite queue size. Then, for each user, we can get its EC, which specifies the maximum arrival rate that its link can support, on the condition that this user's delay QoS requirement is satisfied. Therefore, when we consider the two users as a system, the total maximum achievable arrival rate can be defined as a linear summation of the two users' EC values.

Lemma 11. *Considering the total EC in NOMA,* T_N, *for the two-user system, we prove that*

(a) At any values of ρ, $\frac{\partial T_N}{\partial \rho} \geq 0$.

(b) When $\rho \to 0$, $T_N \to 0$, $\lim\limits_{\rho\to 0} \dfrac{\partial T_N}{\partial \rho} = \dfrac{1-\alpha_n}{\ln 2}\mathbb{E}\left[|h_m|^2\right] + \dfrac{\alpha_n}{\ln 2}\mathbb{E}\left[|h_n|^2\right] \geq 0.$

(c) When $\rho \to \infty$, $T_N \to \infty$, $\lim\limits_{\rho\to \infty} \dfrac{\partial T_N}{\partial \rho} = 0.$

Considering the total EC in OMA, T_O*, for the two-user system, we prove that*

(d) At any values of ρ, $\dfrac{\partial T_O}{\partial \rho} \geq 0.$

(e) When $\rho \to 0$, $T_O \to 0$, $\lim\limits_{\rho\to 0} \dfrac{\partial T_O}{\partial \rho} = \dfrac{1}{2\ln 2}\mathbb{E}\left[|h_m|^2\right] + \dfrac{1}{2\ln 2}\mathbb{E}\left[|h_n|^2\right] \geq 0.$

(f) When $\rho \to \infty$, $T_O \to \infty$, $\lim\limits_{\rho\to \infty} \dfrac{\partial T_O}{\partial \rho} = 0.$

Proof. The proof is provided in Appendix L. □

From Lemma 11.(a) and Lemma 11.(d), one can note that the total link-layer rate for the two-user system, either in NOMA or OMA, shows a non-decreasing trend with the transmit SNR. Furthermore, Lemma 11.(b) indicates that when the NOMA scheme is applied, the total EC has a constant slope at small values of ρ, in which the constant depends on the average of the channel power gains and the allocated power coefficients. On the contrary, from Lemma 11.(e), one can find that the total EC obtained in OMA scheme also shows a constant increasing speed at small values of ρ, in which the constant only depends on the average of the channel power gains. Finally, when $\rho \to \infty$, Lemma 11.(c) and Lemma 11.(f) show that the increasing speed of the total EC, either in NOMA or OMA, gradually diminishes.

To thoroughly investigate the region of ρ, in which NOMA is more advantageous than OMA, in terms of the total link-layer rate for the complete two-user system, we analyze $\mathrm{T}_N - \mathrm{T}_O$ in the following lemma.

Lemma 12. *Considering the difference of the total EC between NOMA and OMA, for a two-user system, we prove that*

(a) When $\rho \to 0$, $T_N - T_O \to 0$, $\lim\limits_{\rho\to 0} \dfrac{\partial\left(T_N - T_O\right)}{\partial \rho} = \dfrac{1-2\alpha_n}{2\ln 2}\mathbb{E}[|h_m|^2] + \dfrac{2\alpha_n - 1}{2\ln 2}\mathbb{E}[|h_n|^2] \leq 0.$

(b) When $\rho \to \infty$, $T_N - T_O$ *approaches a constant, given in* (5.14)*, and we get* $\lim\limits_{\rho\to \infty} \dfrac{\partial\left(T_N - T_O\right)}{\partial \rho} = 0.$

Proof. The proof is provided in Appendix M. □

$$\lim_{\rho\to\infty}(\mathrm{T}_N-\mathrm{T}_O)=-\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{\alpha_n^{-2\beta_m}}{\mathbb{E}\Big[\big(|h_m|^2\big)^{\beta_m}\Big]}\right)-\frac{1}{\theta_n T_{\mathrm{f}} B}\ln\left(\frac{{\alpha_n}^{2\beta_n}\mathbb{E}\Big[\big(|h_n|^2\big)^{2\beta_n}\Big]}{\mathbb{E}\Big[\big(|h_n|^2\big)^{\beta_n}\Big]}\right). \quad (5.14)$$

From Lemma 12.(a) and Lemma 12.(b), one can conclude that $\mathrm{T}_N - \mathrm{T}_O$ starts at the initial value of 0, first decreases and finally approaches a constant when $\rho \to \infty$, given in (5.14). This reveals that at extremely high SNRs, the difference of the total link-layer rate, between NOMA and OMA, stabilizes at a constant value, which is irrelevant with the transmit SNR, but depends on the two users' average channel power gains, power coefficients, and delay QoS exponents. Together with Lemma 11, one can conclude that OMA can achieve a higher value of the total EC, when the transmit SNR is small, in comparison with the NOMA scheme.

5.4.1.3 Case 2: Consider Delay-Unconstrained Users

In this subsection, the delay-unconstrained EC is investigated, in a two-user NOMA network and a two-user OMA network, when $\theta_m \to 0$, $\theta_n \to 0$, i.e., $\lim_{\theta_m\to 0} E_{\mathrm{c}}^m$, $\lim_{\theta_m\to 0} \bar{E}_{\mathrm{c}}^m$, $\lim_{\theta_n\to 0} E_{\mathrm{c}}^n$, $\lim_{\theta_n\to 0} \bar{E}_{\mathrm{c}}^n$, as well as the EC difference between NOMA and OMA, for both users, i.e., $\lim_{\theta_m\to 0}\left(E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m\right)$ and $\lim_{\theta_n\to 0}\left(E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n\right)$. Further, the impact of ρ in this delay-unconstrained situation is also analyzed and investigated.

Lemma 13. *Considering the EC for the m^{th} user with $\theta_m \to 0$, in NOMA and OMA, we prove that*

(a) When $\theta_m \to 0$, $\lim_{\theta_m\to 0} E_c^m = \mathbb{E}[R_m]$, $\lim_{\theta_m\to 0} \bar{E}_c^m = \mathbb{E}\left[\bar{R}_m\right]$, $\lim_{\theta_m\to 0}\left(E_c^m - \bar{E}_c^m\right) = \mathbb{E}[R_m] - \mathbb{E}\left[\bar{R}_m\right]$.

(b) When $\theta_m \to 0$, $\rho\to\infty$, $\lim_{\substack{\theta_m\to 0\\ \rho\to\infty}} E_c^m = \log_2\left(\frac{1}{\alpha_n}\right)$, $\lim_{\substack{\theta_m\to 0\\ \rho\to\infty}} \bar{E}_c^m \to \infty$, $\lim_{\substack{\theta_m\to 0\\ \rho\to\infty}}\left(E_c^m - \bar{E}_c^m\right) \to -\infty$.

Considering the EC for the n^{th} user with $\theta_n \to 0$, in NOMA and OMA, we prove that

(c) When $\theta_n \to 0$, $\lim_{\theta_n\to 0} E_c^n = \mathbb{E}[R_n]$, $\lim_{\theta_n\to 0} \bar{E}_c^n = \mathbb{E}\left[\bar{R}_n\right]$, $\lim_{\theta_n\to 0}\left(E_c^n - \bar{E}_c^n\right) = \mathbb{E}[R_n] - \mathbb{E}\left[\bar{R}_n\right]$.

(d) When $\theta_n \to 0$, $\rho \to \infty$, $\lim_{\substack{\theta_n\to 0\\ \rho\to\infty}} E_c^n \to \infty$, $\lim_{\substack{\theta_n\to 0\\ \rho\to\infty}} \bar{E}_c^n \to \infty$, $\lim_{\substack{\theta_n\to 0\\ \rho\to\infty}}\left(E_c^n - \bar{E}_c^n\right) \to \infty$.

Proof. The proof is provided in Appendix N. □

From Lemma 13.(a) and Lemma 13.(c), it is noted that for both users, no matter in NOMA or OMA, when there is no delay requirement, i.e., $\theta_m \to 0$, and $\theta_n \to 0$, the individual achievable link-layer rate is equivalent to the ergodic capacity. Furthermore, from Lemma 8 and Lemma 13, one can find that, the same conclusions regarding to the performance of the system at high SNRs apply to the delay-unconstrained and the delay-constrained users. For example, from Lemma 8.(b) and Lemma 13.(b), it indicates that the weaker user in a two-user NOMA system can only achieve a limited EC, no matter how large the transmit SNR can be, or how strict or loose the delay exponent is. Further, one can also conclude that, for the weaker user, either with or without delay constraint, OMA offers higher EC than NOMA, when $\rho \to \infty$. On the contrary, for the stronger user, either with or without delay constraint, NOMA achieves higher EC than OMA at high SNRs.

Note that in Section 5.4.1.2 and Section 5.4.1.3, we have comprehensively investigated the individual link-layer rate and the total EC for a two-user NOMA system, in comparison with the conventional OMA scheme. For delay-constrained and delay-unconstrained users, we have characterized the region of ρ, in which NOMA is more beneficial than OMA, in terms of the individual and the total EC. These insightful conclusions, mathematically derived and theoretically proved, can provide valuable guidelines for the further research, such as the resource allocation design, user pairing/clustering technique and delay analysis in NOMA. Further, the above theoretical conclusions will be confirmed using simulation results in Section 5.5.

5.4.2 Effective Capacity of Multiple NOMA Pairs

After analyzing the two-user NOMA network and deriving the closed-form expressions, the total achievable link-layer rate for multiple NOMA pairs can be investigated. By considering that the M users are divided into $\frac{M}{2}$ groups, we define $\mathbb{I} = \{1, 2, \ldots, \frac{M}{2}\}$, which contains the group index. Then, all NOMA pairs can be included in Φ, $\Phi = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{M/2}\}$, satisfying $\boldsymbol{\phi}_i \cap \boldsymbol{\phi}_j = \varnothing, i \neq j, \forall i, j \in \mathbb{I}$, where $\boldsymbol{\phi}_i = \{(m_i, n_i) \mid m_i \neq n_i, |h_{m_i}|^2 \leq |h_{n_i}|^2, \forall i \in \mathbb{I}\}$ denotes the i^{th} NOMA pair with two users, i.e., m_i and n_i.

Assume that for the i^{th} NOMA pair, $\forall i \in \mathbb{I}$, NOMA will be implemented for the two users, i.e., m_i and n_i. Meanwhile, for the inter-group multiple access, it is assumed that TDMA will be applied. Hence, for the two users in the i^{th} NOMA pair,

the achievable data rates, in b/s/Hz, can be respectively formulated as

$$R_{m_i} = \frac{2}{M}\log_2\left(1+\frac{\rho\alpha_{m_i}|h_{m_i}|^2}{\rho\alpha_{n_i}|h_{m_i}|^2+1}\right), \tag{5.15a}$$

$$R_{n_i} = \frac{2}{M}\log_2\left(1+\rho\alpha_{n_i}|h_{n_i}|^2\right). \tag{5.15b}$$

On the other hand, if the users m_i and n_i each have their message transmitted using TDMA, the achievable data rate for each user can be given by

$$\bar{R}_j = \frac{1}{M}\log_2\left(1+\rho|h_j|^2\right),\ j\in\{m_i,n_i\}, \tag{5.16}$$

where $\frac{1}{M}$ denotes that each user has only $\frac{1}{M}$ of the time slot to transmit, while in the other fractions of the time slot, it will stay silent.

Assuming that the Gärtner-Ellis theorem is satisfied, the EC formulations for the users m_i and n_i in the i^{th} NOMA pair can be obtained, yielding

$$E_{\text{c}}^{m_i} = -\frac{1}{\theta_{m_i}T_{\text{f}}B}\ln\left(\mathbb{E}\left[\left(\frac{\rho|h_{m_i}|^2+1}{\rho\alpha_{n_i}|h_{m_i}|^2+1}\right)^{\frac{4}{M}\beta_{m_i}}\right]\right), \tag{5.17a}$$

$$E_{\text{c}}^{n_i} = -\frac{1}{\theta_{n_i}T_{\text{f}}B}\ln\left(\mathbb{E}\left[\left(1+\rho\alpha_{n_i}|h_{n_i}|^2\right)^{\frac{4}{M}\beta_{n_i}}\right]\right), \tag{5.17b}$$

where $\beta_{m_i} = -\frac{\theta_{m_i}T_{\text{f}}B}{2\ln 2}$, and $\beta_{n_i} = -\frac{\theta_{n_i}T_{\text{f}}B}{2\ln 2}$. On the contrary, for the TDMA scheme, the EC expressions for both users can also be obtained, which respectively yield to

$$\bar{E}_{\text{c}}^{m_i} = -\frac{1}{\theta_{m_i}T_{\text{f}}B}\ln\left(\mathbb{E}\left[\left(1+\rho|h_{m_i}|^2\right)^{\frac{2}{M}\beta_{m_i}}\right]\right), \tag{5.18a}$$

$$\bar{E}_{\text{c}}^{n_i} = -\frac{1}{\theta_{n_i}T_{\text{f}}B}\ln\left(\mathbb{E}\left[\left(1+\rho|h_{n_i}|^2\right)^{\frac{2}{M}\beta_{n_i}}\right]\right). \tag{5.18b}$$

Comparing (5.17a)-(5.18b) with (5.7a)-(5.8b), one can notice that the EC formulations for the two users in the i^{th} NOMA pair, have similar expressions with those proposed for a two-user NOMA network in Section 5.4.1. Hence, by following similar steps in Appendix H, the closed-form expressions for $E_{\text{c}}^{m_i}$, $E_{\text{c}}^{n_i}$, $\bar{E}_{\text{c}}^{m_i}$, and $\bar{E}_{\text{c}}^{n_i}$ can be easily obtained, which are omitted here for simplicity. Our focus lies on analyzing the total EC of multiple NOMA pairs, denoted by M_N, in comparison with the total EC for the M OMA users, i.e., M_O. Note that M_N can be defined as $\sum_{i=1}^{M/2}(E_{\text{c}}^{m_i}+E_{\text{c}}^{n_i})$, and correspondingly, M_O equals to $\sum_{i=1}^{M/2}\left(\bar{E}_{\text{c}}^{m_i}+\bar{E}_{\text{c}}^{n_i}\right)$. To investigate the region of ρ, in which NOMA can offer a higher value of the total link-layer rate for multiple NOMA pairs, in comparison with the OMA scheme, we provide the following lemma.

Lemma 14. *Considering the difference of the total EC, between multiple NOMA pairs and M OMA users, we prove that*

(a) When $\rho \to 0$, $M_N - M_O \to 0$, $\lim_{\rho\to 0} \frac{\partial(M_N - M_O)}{\partial \rho} = \sum_{i=1}^{M/2} \frac{1-2\alpha_{n_i}}{M \ln 2} (\mathbb{E}[|h_{m_i}|^2] - \mathbb{E}[|h_{n_i}|^2]) \leq 0$.

(b) When $\rho \to \infty$, $M_N - M_O$ *approaches a constant, given in* (5.19), *and we get* $\lim_{\rho\to\infty} \frac{\partial (M_N - M_O)}{\partial \rho} = 0$.

Proof. The proof follows similar steps in Appendix M, and is omitted here for simplicity. □

$$\lim_{\rho\to\infty}(\mathrm{M}_N - \mathrm{M}_O) = \sum_{i=1}^{M/2} -\frac{1}{\theta_{m_i} T_{\mathrm{f}} B} \ln\left(\frac{\alpha_{n_i}^{-\frac{4}{M}\beta_{m_i}}}{\mathbb{E}\left[(|h_{m_i}|^2)^{\frac{2}{M}\beta_{m_i}}\right]}\right) - \frac{1}{\theta_{n_i} T_{\mathrm{f}} B} \ln\left(\frac{\alpha_{n_i}^{\frac{4}{M}\beta_{n_i}} \mathbb{E}\left[(|h_{n_i}|^2)^{\frac{4}{M}\beta_{n_i}}\right]}{\mathbb{E}\left[(|h_{n_i}|^2)^{\frac{2}{M}\beta_{n_i}}\right]}\right). \tag{5.19}$$

From Lemma 14, one can conclude that $\mathrm{M}_N - \mathrm{M}_O$ starts at the initial value of 0, first decreases at small values of ρ, and finally approaches a constant when $\rho \to \infty$, given in (5.19). This indicates that OMA outperforms NOMA on the total link-layer rate performance for a M-user network, at small SNRs. Simulation results in the next section further show that NOMA achieves higher total EC than OMA at high values of SNR. Finally, Lemma 14.(b) indicates that the performance gain of NOMA over OMA becomes stable when the transmit SNR becomes extremely high.

5.5 Numerical Results

In this section, all the theorems and the lemmas proposed in Section 5.4 will be numerically confirmed. Further, the impact of the per-user delay QoS exponent, and the transmit SNR ρ on the individual EC performance and the total link-layer rate, in NOMA and OMA scenarios, is numerically analyzed and investigated in this section. Specifically, we start from showing the simulation results for the two-user system, in NOMA and OMA. To consider a two-user NOMA system, the total number of users $M = 10$, and the users with the 2^{rd} and the 8^{th} weakest channels are assumed to be paired together, i.e., $m = 2$, $n = 8$. The corresponding power coefficients for the two users are set as, $\alpha_m = 0.8$, $\alpha_n = 0.2$, unless otherwise indicated. The fading-block duration $T_{\mathrm{f}} = 0.01$ ms, and the bandwidth $B = 100$kHz.

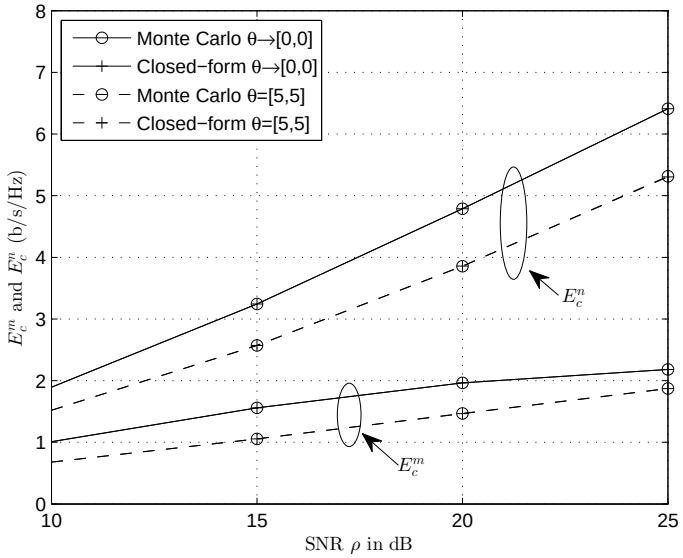


Figure 5.2: E_{c}^{m} and E_{c}^{n}, in NOMA, versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

To confirm the accuracy of the proposed closed-form expressions for EC in NOMA scheme for both users, Fig. 5.2 plots the curves of E_{c}^{m} and E_{c}^{n} versus the transmit SNR ρ, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = [\theta_m, \theta_n]$. This figure shows the results calculated in two ways, i.e., by using Monte Carlo simulation method and the proposed closed-form expressions in this chapter. From Fig. 5.2, the accuracy of the closed-form expressions for EC in NOMA scheme for both users can be confirmed. For the m^{th} user and the n^{th} user, E_{c}^{m} and E_{c}^{n} gradually increase with the transmit SNR ρ, which confirms the proposed Lemma 9.(a) and Lemma 10.(a). Further, when the delay QoS exponent vector becomes more stringent, i.e., changing from $\boldsymbol{\theta} \to [0, 0]$ to $\boldsymbol{\theta} = [5, 5]$, the individual link-layer rates in NOMA, for both users, decrease. This phenomenon will be further investigated in Fig. 5.10.

Fig. 5.3 includes the plots for E_{c}^{m} and $\bar{E}_{\mathrm{c}}^{m}$ versus the transmit SNR ρ, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$. This figure first shows that when ρ increases, the link-layer rate for the m^{th} user, either in NOMA or OMA, shows a non-decreasing trend. This confirms the proved Lemma 9.(a). For the E_{c}^{m} in NOMA scheme, it first increases when ρ is relatively small, then reaches a limit when ρ becomes very large. This observation confirms Lemma 8.(b), since it is proved that when $\rho \to \infty$, E_{c}^{m} approaches a maximum limit which is independent from the transmit SNR and the user's delay QoS requirement. Further, from Fig. 5.3, one can notice that E_{c}^{m} saturates as soon as $\rho \geq 30$dB, although in Lemma 8.(b), the

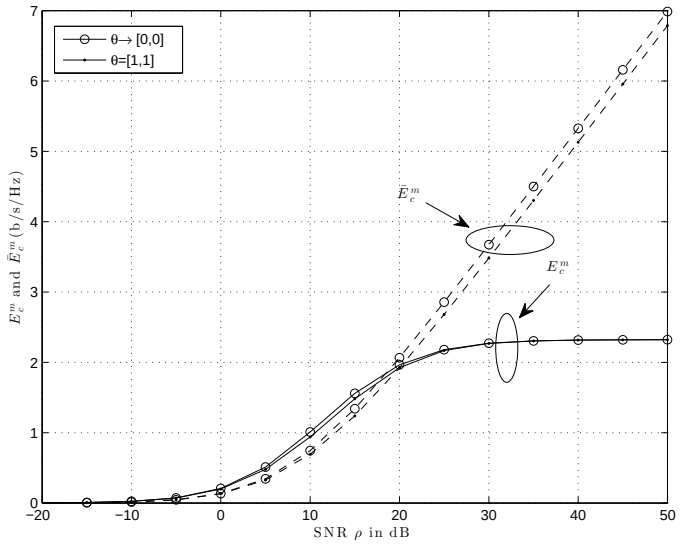


Figure 5.3: E_{c}^{m}, in NOMA, and $\bar{E}_{\mathrm{c}}^{m}$, in OMA, versus the transmit SNR ρ for various values of $\boldsymbol{\theta}$.

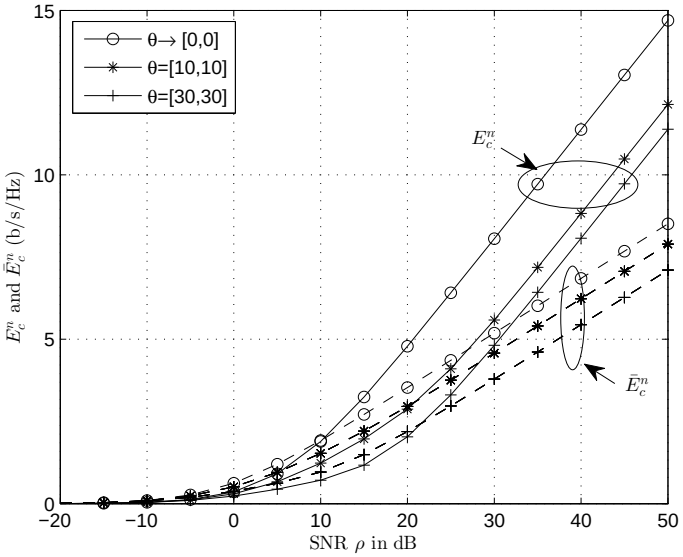


Figure 5.4: E_{c}^{n}, in NOMA, and $\bar{E}_{\mathrm{c}}^{n}$, in OMA, versus the transmit SNR ρ for various values of $\boldsymbol{\theta}$.

maximum limit of E_{c}^{m} achieves when $\rho \to \infty$. Finally, Fig. 5.3 shows that E_{c}^{m} in NOMA prevails over $\bar{E}_{\mathrm{c}}^{m}$ in OMA, when ρ is small, but with the increase of ρ, OMA outperforms NOMA on the link-layer rate performance, for the m^{th} user, which confirms the analysis and explanations in Lemma 9 and Lemma 13.

Considering the n^{th} user, Fig. 5.4 plots the curves of E_{c}^{n} and $\bar{E}_{\mathrm{c}}^{n}$ versus the

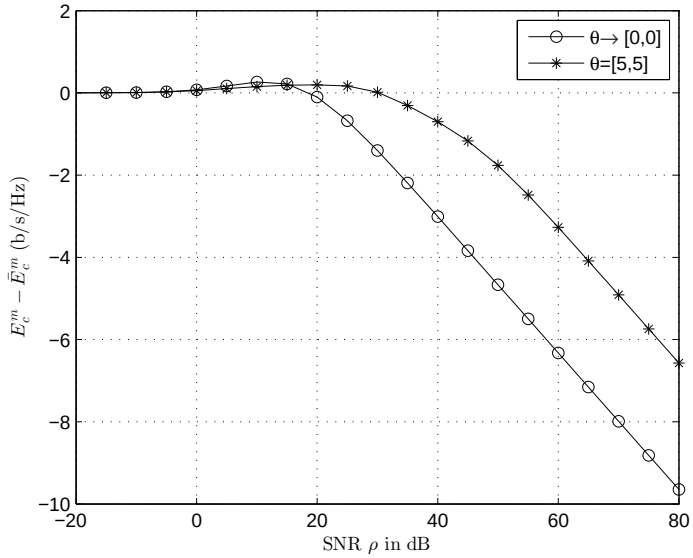


Figure 5.5: $E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m}$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

transmit SNR ρ, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$. From this figure, we note that E_{c}^{n} and $\bar{E}_{\mathrm{c}}^{n}$ start at the same value of 0, then monotonically increase with respect to the transmit SNR ρ. This confirms Lemma 8.(a) and Lemma 10.(a). Furthermore, for a fixed value of $\boldsymbol{\theta}$, when ρ is small, $\bar{E}_{\mathrm{c}}^{n}$ in OMA is larger than E_{c}^{n} in NOMA, but with the increase of the transmit SNR, NOMA becomes more beneficial, in terms of the link-layer rate, which is analytically explained in Lemma 10 and Lemma 13. In addition, when the delay QoS exponent vector becomes more stringent, i.e., changing from $\boldsymbol{\theta} \to [0, 0]$ to $\boldsymbol{\theta} = [30, 30]$, the link-layer rate for the n^{th} user, either in NOMA or OMA, decreases, considering a fixed value of ρ.

In order to investigate the advantage of NOMA over OMA, for the m^{th} user and the n^{th} user, Fig. 5.5 and Fig. 5.6 are provided, which include the plots for $E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m}$ and $E_{\mathrm{c}}^{n} - \bar{E}_{\mathrm{c}}^{n}$ versus the transmit SNR ρ, respectively, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$. Fig. 5.5 indicates that for the m^{th} user, $E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m}$ starts at the initial value of 0, increases slightly at small values of ρ, and then decreases when the transmit SNR ρ further increases. This confirms Lemma 9.(b) and Lemma 9.(c). When the transmit SNR is high and fixed, Fig. 5.5 further shows that a more stringent delay requirement with $\boldsymbol{\theta} = [5, 5]$, results in a larger value of $E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m}$ than the delay-unconstrained situation with $\boldsymbol{\theta} \to [0, 0]$. Specifically, in comparison with the delay-unconstrained system, the delay-constrained system with $\boldsymbol{\theta} = [5, 5]$ allows a longer range of ρ, in which NOMA prevails over OMA. On the other hand, for the n^{th} user, Fig. 5.6 shows that $E_{\mathrm{c}}^{n} - \bar{E}_{\mathrm{c}}^{n}$ first starts at the initial value of 0,

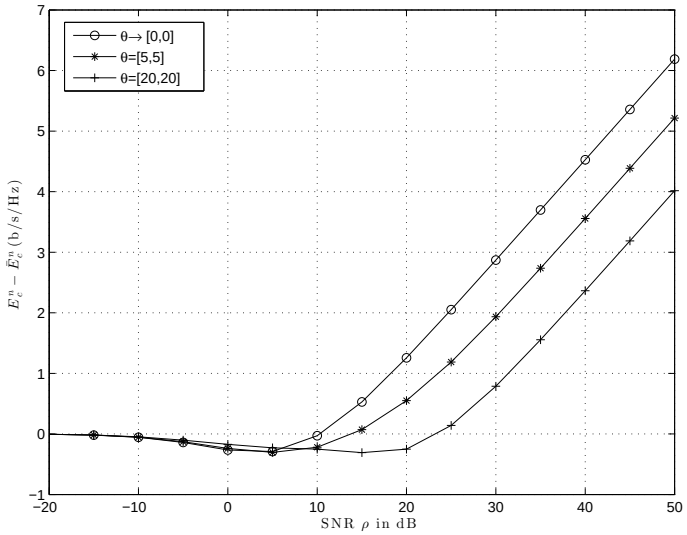


Figure 5.6: $E_{\rm c}^n - \bar{E}_{\rm c}^n$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

slightly decreases when ρ is small, and with the further increase of ρ, it increases. This confirms Lemma 10.(b) and Lemma 10.(c). Furthermore, when the transmit SNR is high and fixed, a more stringent delay requirement with $\boldsymbol{\theta} = [20, 20]$ leads to a smaller value of $E_{\rm c}^n - \bar{E}_{\rm c}^n$, than the delay-unconstrained situation with $\boldsymbol{\theta} \to [0, 0]$.

To investigate the impact of ρ on the performance of the total link-layer rate for the two-user system, Fig. 5.7 is included which plots the curves of T_N in NOMA and T_O in OMA, versus the transmit SNR ρ, for various values of $\boldsymbol{\theta}$. Fig. 5.7 first indicates that the total EC for the two-user network, either in NOMA or OMA, starts at the initial value of 0, and then gradually increases with the transmit SNR ρ. This confirms Lemma 11.(a) and Lemma 11.(d). Specifically, when ρ is very small, the total rate for the two-user network in OMA, T_O, has a faster increasing speed than that in NOMA, T_N, which has been proved and explained in Lemma 12.(a). With the increase of ρ, from Fig. 5.7, one can note that T_N in NOMA gradually becomes higher than T_O in OMA, for both of the delay-constrained situation with $\boldsymbol{\theta} = [1, 1]$ and the delay-unconstrained situation with $\boldsymbol{\theta} \to [0, 0]$. Furthermore, at high values of ρ, the gap of the total EC between NOMA and OMA, for this two-user network, becomes steady, which confirms Lemma 12.(b).

To further investigate and analyze the impact of the transmit SNR ρ and the delay QoS exponent vector $\boldsymbol{\theta}$ on the total EC difference, between a two-user NOMA network and a two-user OMA network, Fig. 5.8 and Fig. 5.9 are included which show the plots for $\mathrm{T}_N - \mathrm{T}_O$ versus the transmit SNR ρ, for various settings of the

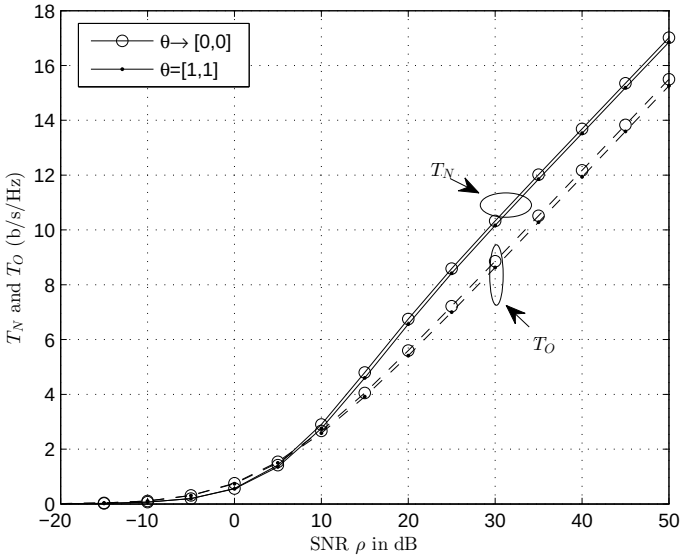


Figure 5.7: T_N and T_O versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

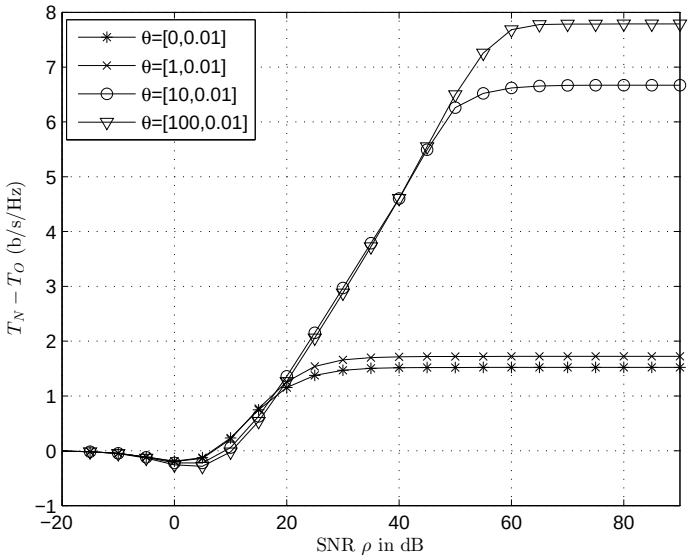


Figure 5.8: $\mathrm{T}_N - \mathrm{T}_O$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

delay QoS exponent vector $\boldsymbol{\theta}$. Specifically, to plot Fig. 5.8, the delay QoS exponent of the n^{th} user is fixed at $\theta_n = 0.01$. Meanwhile, in Fig. 5.9, all curves are plotted by fixing the value of θ_m at 0.01. From Fig. 5.8, one can note that for a fixed value of $\boldsymbol{\theta}$, $\mathrm{T}_N - \mathrm{T}_O$ starts at the initial value of 0, first decreases, then increases with the transmit SNR ρ, finally reaches a maximum limit and stabilizes. This confirms the

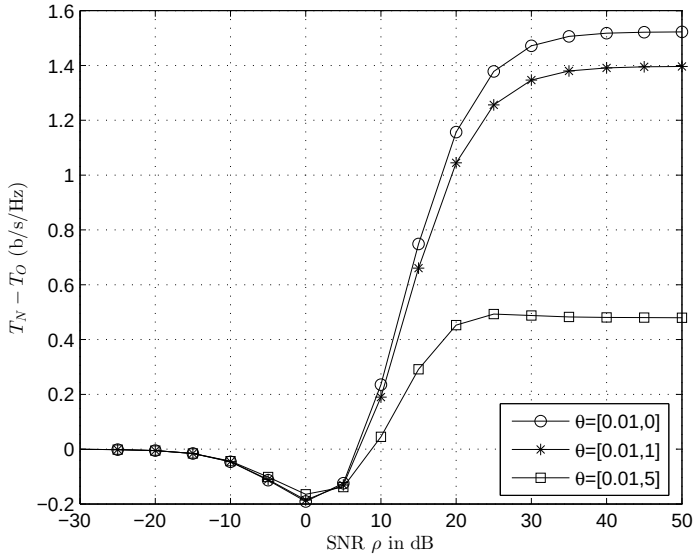


Figure 5.9: $\mathrm{T}_N - \mathrm{T}_O$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

analysis and explanations proposed in Lemma 12. Further, Fig. 5.8 indicates that when θ_n is fixed at 0.01, a larger θ_m leads to a higher value of $\mathrm{T}_N - \mathrm{T}_O$ at high SNRs. Correspondingly, Fig. 5.9 shows that when θ_m is fixed at 0.01, a smaller θ_n results in a higher level of $\mathrm{T}_N - \mathrm{T}_O$ at high SNRs.

To investigate the impact of the delay QoS exponent θ_m on the link-layer rate performance for the m^{th} user, Fig. 5.10 plots the results of E_{c}^m in NOMA (in solid lines) and $\mathbb{E}[R_m]$ (in dash lines) versus the delay QoS exponent θ_m, for various values of ρ. This figure first indicates that, when the m^{th} user has a loose delay requirement, i.e., $\theta_m \leq 10^{-1}$, the link-layer rate in NOMA, E_{c}^m, is equivalent to the physical-layer rate $\mathbb{E}[R_m]$, which confirms Lemma 13.(a). When the delay requirement becomes more stringent, E_{c}^m gradually decreases to the minimum value of 0, for various values of ρ. On the contrary, the curves of $\mathbb{E}[R_m]$ versus θ_m always stay high and stable, but this is due to the reason that there is no delay requirement guaranteed when the physical-layer rate is considered. Furthermore, considering a fixed θ_m, when ρ increases from 10 dB to 30 dB, E_{c}^m becomes larger, which indicates that a higher value of ρ will result in a larger value of EC in NOMA, for the m^{th} user.

Finally, the focus lies on the comparison of NOMA and OMA, in terms of the total link-layer rate difference, between multiple NOMA pairs and M OMA users. To investigate the impact of the transmit SNR ρ and the user pairing set Φ on the total EC difference, Fig. 5.11 includes the plots for $\mathrm{M}_N - \mathrm{M}_O$ versus the transmit SNR ρ, for various settings of the user pairing set Φ. Specifically, the total number of users

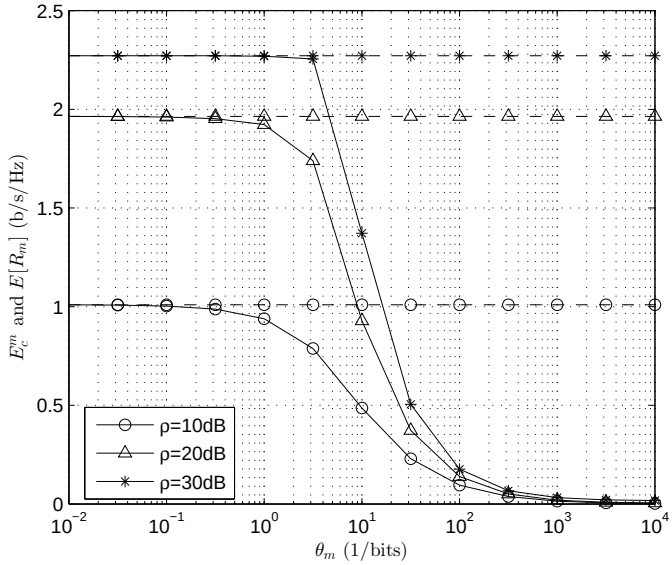


Figure 5.10: E_{c}^{m}, in NOMA, versus θ_m for various values of the transmit SNR ρ.

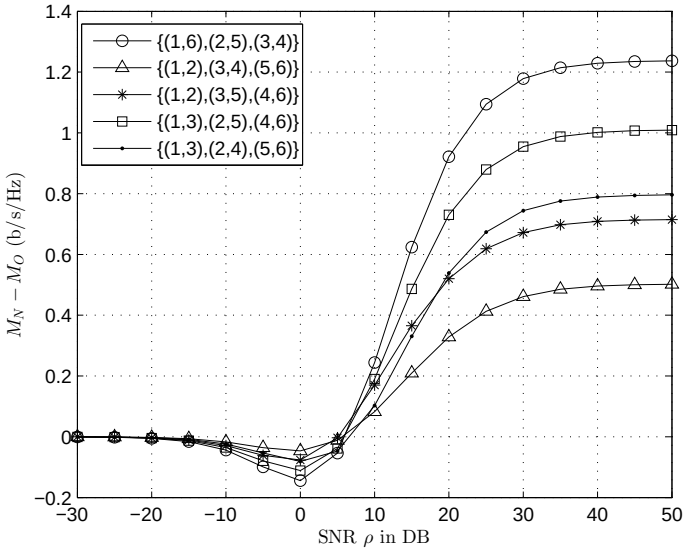


Figure 5.11: $\mathrm{M}_N - \mathrm{M}_O$, versus the transmit SNR ρ for various settings of user pairing set Φ.

$M = 6$, the power coefficients allocated to both users in a NOMA pair are given as $\alpha_{m_i} = 0.8$, $\alpha_{n_i} = 0.2$, $\forall i \in \mathbb{I}$, $(m_i, n_i) \in \Phi$[11], and the delay QoS exponents of all users are assumed to be approaching 0. From Fig. 5.11, it is noted that for a fixed setting of Φ, $\mathrm{M}_N - \mathrm{M}_O$ starts at the initial value of 0, first decreases, then increases until it

[11]Different settings of power coefficients can influence the simulation results, but this is beyond the scope of this chapter, and can be kept as a future research topic.

reaches a maximum value. This confirms the proposed Lemma 14 in Section 5.4.2, which reveals that OMA achieves higher total EC than NOMA at small values of ρ. Fig. 5.11 also indicates that NOMA is more beneficial than OMA, on the total EC performance for a M-user network, when the transmit SNR becomes extremely high. Furthermore, from Fig. 5.11, it shows that at high SNRs, the user pairing setting of $\Phi = \{(1,6),(2,5),(3,4)\}$ provides the largest level of $\mathrm{M}_N - \mathrm{M}_O$, which means that among all the simulated settings, this case is the best user pairing solution.

5.6 Summary

The individual achievable link-layer rate and the total EC, under the per-user statistical delay QoS requirement, were investigated and analyzed for a downlink NOMA network with M users. Assuming that the M users are divided into multiple NOMA pairs, it was proved that NOMA offers higher total EC than OMA at high SNRs. Furthermore, the performance gain of NOMA over OMA was proved to become stable when the transmit SNR is extremely high. This indicates that once above a high level, the increase of transmit SNR cannot guarantee more performance gain. Aware of the importance of a two-user downlink NOMA network, the impact of the transmit SNR and the delay QoS requirement on the individual EC performance and the total link-layer rate, in NOMA and OMA, was studied for a two-user NOMA network. Two cases were respectively analyzed. Case 1: Consider delay-constrained users. Case 2: Consider delay-unconstrained users. For the stronger user, either delay-constrained or delay-unconstrained, it was proved that NOMA prevails over OMA, when the transmit SNR is large. On the contrary, for the weaker user in a two-user network, it was proved that NOMA offers higher EC than OMA at small values of SNR. Furthermore, for the weaker user, either delay-constrained or delay-unconstrained, the EC in NOMA was proved to be limited to a maximum value, even if the transmit SNR goes to infinity. To confirm these theoretical conclusions, the closed-form expressions for the individual EC in a two-user network were derived and confirmed by using simulation results. Finally, simulation results also revealed that the user pairing settings and the allocated power coefficients can influence the throughput performance, which can be reserved as potential research topics.

Chapter 6

Conclusions and Future Work

6.1 Summary

Due to the imperative need of satisfying statistical delay QoS guarantees, this thesis mainly discussed the delay-constrained resource allocation and the link-layer throughput analysis, focusing on different wireless communication networks. Based on the link-layer channel model proposed in [10], the maximum arrival rate that a given service process can support was analyzed and investigated, while satisfying a required statistical delay QoS constraint.

Considering the compromise between the achievable rate, energy savings, and delay QoS provisioning, the focus first lay on designing an efficient resource allocation strategy to balance the three important metrics. Note that EC denotes the maximum arrival rate with a guaranteed delay violation probability and the link-layer EE can be formulated as the ratio of EC to the total power expenditure [12]. Hence, the focus then lay on jointly maximizing EC and the link-layer EE, so that the three QoS metrics can be balanced. It has been proved that the delay-constrained link-layer rate, namely EC, conflicts with the link-layer EE, just like the inconsistent property of EE and SE, from the physical-layer channel model [22]. Hence, to jointly maximize two incompatible metrics, it falls into the scope of the multi-objective optimization problem. Focusing on a point-to-point single-user single-carrier communication system, a multi-objective optimization problem of link-layer EE and EC was proposed and investigated, under a delay-outage probability constraint and an average transmit power limit. This problem was then solved and the proposed optimal power allocation strategy was proved to be sufficient for the Pareto optimal set of the original formulated multi-objective optimization problem.

To further balance the three QoS metrics in a more practical scenario, the delay-QoS driven resource allocation problem was studied for the uplink transmission in

a multi-user multi-carrier OFDMA system. A total EC maximization problem was proposed and formulated, subject to each user's link-layer EE requirement and the per-user average transmit power limit. To solve this problem and provide the resource allocation solution, a low-complexity heuristic algorithm was proposed, which first allocates each served user the exact number of its required subcarriers, and then implements the per-user optimal power allocation strategy. Finally, the remaining subcarriers will be allocated by adopting the strategy that the user with current minimum EC value has the allocation priority.

Finally, based on the link-layer channel model, the performance of an achievable link-layer rate was analyzed for a downlink NOMA network with M users, under the per-user statistical delay QoS requirement. The advantage of NOMA over OMA was investigated by analyzing the impact of the transmit SNR on the total link-layer rate performance. Considering a two-user network as a special case, the performance gain of NOMA over OMA was also analyzed and investigated, for delay-constrained and delay-unconstrained users. To confirm the proposed theoretical conclusions, the closed-form expressions for the achievable EC in a two-user network, in NOMA and OMA, were derived for both users and then confirmed using Monte Carlo simulations.

In more detail, the main contributions of this thesis can be summarized as follows.

In Chapter 2, the EC theory and convex optimization theory were briefly introduced, followed by the literature review. This chapter provides a comprehensive overview, which helps the readers to thoroughly understand the background knowledge in regard to the link-layer channel model, convex optimization and the related existing literature.

In Chapter 3, a normalized link-layer EE-EC MOP in a Nakagami-m fading channel was formulated and then transformed into a weighted SOP, under a delay-outage probability constraint and an average transmit power limit. Specifically, two normalization values were introduced to make the two objectives comparable. To solve the power-constrained SOP, the unconstrained SOP was analyzed first. By proving that the unconstrained EE-EC tradeoff problem is continuously differentiable, strictly quasiconvex in the average power and follows a cup shape curve, it can be concluded that the global optimum is unique and can be achieved at a finite value. Finally, for the power-unconstrained EE-EC tradeoff problem, a closed-form expression for the optimal power allocation strategy was derived, which paves the way for the power-constrained problem. However, for a formulated MOP, the optimal solutions are a set of points which all fit Pareto optimality, rather than a single global solution obtained after solving an SOP. The question then arises, "does the proposed optimal solution

belong to the Pareto optimal set of the original MOP?" According to Theorem 5, Theorem 6, and Lemma 1, one can confirm that the proposed optimal power, calculated for every pre-determined weight value, is sufficient for the Pareto optimal set of the original EE-EC MOP. Further, it was shown that the proposed optimal solution includes the optimal power allocation strategy for the link-layer EE-maximization problem and also the one for the link-layer EC-maximization problem, as extreme cases. This means that the formulated tradeoff problem is general, and the proposed optimal solution is flexible. Finally, the power-constrained link-layer EE-EC tradeoff problem was solved, and the Pseudocode of the optimal power allocation algorithm was provided. Since the calculated optimal power value can be influenced by the system parameters, hence, the impact of these factors on the tradeoff performance was thoroughly studied. It was proved that the average optimal power level monotonically decreases with the importance weight, but strictly increases with the normalization factor, circuit power and power amplifier efficiency. Based on these conclusions, one can find that, if the system prefers a higher EC, a larger value of the normalization factor as well as a smaller importance weight should be chosen to offer a larger optimal transmit power, and correspondingly, a larger EC. In contrast, if a user prefers a higher EE, a smaller value of the normalization factor as well as a larger importance weight are more beneficial. Simulation results confirmed the analytical derivations and further showed the impact of fading severeness and transmission power limit on the tradeoff performance.

In Chapter 4, a total EC maximization problem for the uplink transmission, in a multi-user multi-carrier OFDMA system, was formulated as a combinatorial integer programming problem, subject to each user's link-layer EE requirement as well as the per-user average transmission power limit. Specifically, an adjustable EE requirement factor was introduced to further tune each user's EE requirement value, which transformed the formulated problem into a tradeoff problem between the total EC and all users' individual EE achievements. In order to make the formulated problem tractable, the solving process was divided into two steps: frequency provisioning which decided the number of allocated subcarriers for each user, and then optimal power allocation for each single-user multi-carrier system. In order to obtain the subcarrier assignment solution, a low-complexity heuristic algorithm was proposed and compared with the traditional exhaustive algorithm and a fair-exhaustive algorithm. From the design intention and the simulation results, one can see that the proposed heuristic algorithm cares about user fairness, offers a close-to-optimal performance,

and also has a complexity linearly relating to the size of the problem. Given a subcarrier assignment matrix, the multi-user OFDMA system can then be viewed as a FDMA system, where each user transmits data through a number of assigned subcarriers independently. Hence, the formulated tradeoff problem between the total EC and all users' individual EE achievements can be transformed into a link-layer EE-EC tradeoff problem for each single-user multi-carrier system. The per-user optimal power allocation strategy, across both frequency and time domains, was then derived. Further, to thoroughly investigate the impact of system parameters on the tradeoff performance, it was proved that each user's achieved link-layer EE value monotonically decreases with its circuit power, but increases with its EE requirement factor. Furthermore, each user's optimal average power was proved to be monotonically decreasing with its EE requirement factor. Simulation results confirmed the proofs and design intentions, and further revealed that when there is a link-layer EE constraint, each user's tradeoff EC level will not show a monotonic trend with its delay QoS exponent. This phenonmenon differs from the monotonic trend of the maximum EC versus delay QoS exponent for the unconstrained EC-maximization problem proposed in [20]. On the other hand, simulation results also indicated that when there is a link-layer EE constraint, the tradeoff EC value achieved with a smaller number of available subcarrier may be higher than the one obtained with more subcarriers. This also differs from the monotonic trend of the maximum EC versus the number of subcarriers for the unconstrained EC-maximization problem in [20].

In Chapter 5, the achievable link-layer rate was studied for a downlink NOMA network with M users, under the per-user statistical delay QoS requirement. Specifically, the M users were assumed to be divided into multiple NOMA pairs, with conventional OMA applied for inter-NOMA-pairs multiple access. Focusing on the total link-layer rate for a downlink M-user network, it was proved that OMA outperforms NOMA when the transmit SNR is small. On the contrary, NOMA was proved to be more beneficial than OMA at high values of SNR. Furthermore, the advantage of NOMA over OMA was found to be stable when the transmit SNR is extremely high. This indicates that once above a high level, the increase of transmit SNR cannot guarantee any more performance gain. Note that a two-user downlink NOMA, called as the MUST, has been proposed for the 3GPP-LTE-A networks. Hence, aware of the importance of a two-user NOMA network, the impact of the transmit SNR and the delay QoS requirement on the individual EC performance and the total link-layer rate was also analyzed and investigated for a two-user network. Specifically, for delay-constrained and delay-unconstrained users, it was proved that for the user with the

stronger channel condition in a two-user network, NOMA prevails over OMA when the transmit SNR is large. On the other hand, for the user with the weaker channel condition, it was proved that NOMA outperforms OMA at small values of transmit SNR. Furthermore, for the user with the weaker channel condition, the achievable EC in NOMA was proved to be limited to a maximum value, even if the transmit SNR goes to infinity. To confirm these theoretical conclusions, the closed-form expressions for the achievable EC in a two-user network, by applying NOMA or OMA, were derived for both users and then confirmed using Monte Carlo simulations.

6.2 Future Work

Based on the present results, further work can be carried out in the following areas:

1. Note that the delay-constrained resource allocation has been studied in a single-user single-carrier system and a multi-user multi-carrier communication system, in Chapter 3 and Chapter 4, respectively. However, single-antenna transmitters and receivers were assumed in the two chapters, with perfect CSI considered at the transmitter. More practical and complicated scenarios can be considered, such as other multiple access techniques, imperfect CSI, and multi-antenna users. Firstly, when other multiple access techniques are considered, e.g., power-domain NOMA, the resource allocation problem would be more challenging, due to the interference from other NOMA users. Hence, it is of great importance if an optimal delay-constrained power allocation strategy can be found for a downlink NOMA network. On the other hand, for multi-antenna users with perfect and imperfect CSI assumed, the delay-constrained resource allocation also deserves elaborate study.

2. In Chapter 5, the performance of an achievable link-layer rate, was studied and investigated for a M-user downlink NOMA network, under the per-user statistical delay QoS requirement. However, the M users were assumed to be divided into multiple NOMA pairs in Chapter 5. Firstly, assuming that all M users transmit on the same channel, the closed-form expression for the achievable EC deserves elaborate study. On the other hand, the impact of the transmit SNR and the per-user delay QoS requirement on the individual EC and the total link-layer rate requires further research, for a downlink NOMA network with M users transmitting on the same channel.

3. According to Chapter 2, the theory of EB was introduced by providing a statistical envelope process, which gives an upper bound on the traffic flows in a probabilistic manner. Then, the minimum envelope rate was proved to be the EB satisfying the required buffer overflow probability, under some certain conditions. As a simple and efficient link-layer channel model, the theory of EB and EC is easy to apply. However, different link-layer bounds with respect to the service rate and the arrival rate can be found and applied, by using stochastic process, to satisfy different delay limitations.

Appendix A

Proof of Lemma 1

Proof. Suppose the point $\overline{P_{\mathrm{t}}^{*}} \in [0, P_{\max}]$, is a Pareto optimal solution for problem $Q2$ and it is not a Pareto optimal solution for problem $Q1$. Hence, there must exist $\overline{P_{\mathrm{t}}'}$ with $\frac{\mathrm{SE}(\overline{P_{\mathrm{t}}'})}{\Psi_{\mathrm{SE}}} \geq \frac{\mathrm{SE}(\overline{P_{\mathrm{t}}^{*}})}{\Psi_{\mathrm{SE}}}$, $\frac{\mathrm{EE}(\overline{P_{\mathrm{t}}'})}{\Psi_{\mathrm{EE}}} \geq \frac{\mathrm{EE}(\overline{P_{\mathrm{t}}^{*}})}{\Psi_{\mathrm{EE}}}$, and also at least one of the two following conditions happens: 1) $\frac{\mathrm{SE}(\overline{P_{\mathrm{t}}'})}{\Psi_{\mathrm{SE}}} > \frac{\mathrm{SE}(\overline{P_{\mathrm{t}}^{*}})}{\Psi_{\mathrm{SE}}}$, 2) $\frac{\mathrm{EE}(\overline{P_{\mathrm{t}}'})}{\Psi_{\mathrm{EE}}} > \frac{\mathrm{EE}(\overline{P_{\mathrm{t}}^{*}})}{\Psi_{\mathrm{EE}}}$[1]. Note that $\mathrm{SE}(\overline{P_{\mathrm{t}}})$, $\mathrm{EE}(\overline{P_{\mathrm{t}}})$, for $\overline{P_{\mathrm{t}}} \in [0, P_{\max}]$, are always positive, therefore, there exists $\overline{P_{\mathrm{t}}'}$ which guarantees that $\frac{\Psi_{\mathrm{SE}}}{\mathrm{SE}(\overline{P_{\mathrm{t}}'})} \leq \frac{\Psi_{\mathrm{SE}}}{\mathrm{SE}(\overline{P_{\mathrm{t}}^{*}})}$, $\frac{\Psi_{\mathrm{EE}}}{\mathrm{EE}(\overline{P_{\mathrm{t}}'})} \leq \frac{\Psi_{\mathrm{EE}}}{\mathrm{EE}(\overline{P_{\mathrm{t}}^{*}})}$, and at least one of the two following conditions happens: 1) $\frac{\Psi_{\mathrm{SE}}}{\mathrm{SE}(\overline{P_{\mathrm{t}}'})} < \frac{\Psi_{\mathrm{SE}}}{\mathrm{SE}(\overline{P_{\mathrm{t}}^{*}})}$, 2) $\frac{\Psi_{\mathrm{EE}}}{\mathrm{EE}(\overline{P_{\mathrm{t}}'})} < \frac{\Psi_{\mathrm{EE}}}{\mathrm{EE}(\overline{P_{\mathrm{t}}^{*}})}$. This contradicts the assumption that $\overline{P_{\mathrm{t}}^{*}}$ is a Pareto optimal solution for problem $Q2$. This concludes the proof of Lemma 1. □

[1]Here, $\mathrm{SE}(\overline{P})$ and $\mathrm{EE}(\overline{P})$ are defined as the SE and EE values achieved at certain average power $\overline{P}$.

Appendix B

Proof of Theorem 5

Proof. [1]Since $\hat{P}$ is unique optimal solution for the weighted SOP, then $\sum_{i=1}^{q} w_i f_i(\hat{P}) < \sum_{i=1}^{q} w_i f_i(P)$, $w_i \in [0,1]$, $\sum_{i=1}^{q} w_i = 1$, for all $P \in [0, P_{\max}]$. Suppose $\hat{P}$ is not a Pareto optimal solution for the MOP. Hence, there must exist $P' \in [0, P_{\max}]$ with $f_i(P') \leq f_i(\hat{P})$ for all $i = 1, \ldots, q$, and there is at least one j, such that $f_j(P') < f_j(\hat{P})$, $j = 1, \ldots, q$. Multiplying by the weights, we have $w_i f_i(P') \leq w_i f_i(\hat{P})$ for all $i = 1, \ldots, q$, and $\sum_{i=1}^{q} w_i f(P') \leq \sum_{i=1}^{q} w_i f(\hat{P})$. This contradicts the uniqueness assumption. Therefore, the theorem is proved. □

[1]A similar theorem was mentioned in [113], but the proof was not provided.

Appendix C

Proof of Theorem 6

Proof. Denote the sublevel set of U7 by $\mathcal{S}_\beta = \left\{ \overline{P_\mathrm{r}} \in \left[0, \frac{P_{\max}}{K_\ell}\right] \middle| U7 \leq \beta \right\}$. According to [59], $U7$ is strictly quasiconvex in $\overline{P_\mathrm{r}}$ if $\mathcal{S}_\beta$ is strictly convex for any real number β. In more detail, a set is strictly convex if any line (without the endpoints) connecting two points in the set is inside the interior of the set. In other words, the set C is strictly convex if every point $c = \lambda a + (1-\lambda)b$, $\lambda \in (0,1)$, $\lambda \in \mathbb{R}$, for any two points $a, b \in C$, $a \neq b$, is inside the interior of C.

Firstly, when $\beta < 0$, no points exist for $U7 = \beta$. When $\beta \geq 0$, $\mathcal{S}_\beta$ is equivalent to

$$\mathcal{S}_\beta = \left\{ P_\mathrm{r} \in \left[0, \frac{P_{\max}}{K_\ell}\right] \middle| \ln\left(\mathbb{E}_\gamma\left[(1+P_\mathrm{r}\gamma)^{-\alpha(\theta)}\right]\right) - \beta\left(w_1 \Psi_{\mathrm{EE_r}}\left(P_{\mathrm{c_r}} + \frac{1}{\epsilon}\overline{P_\mathrm{r}}\right) + (1-w_1)\,\Psi_{\mathrm{EC}}\right) \leq 0 \right\}.$$

Since $\ln\left(\mathbb{E}_\gamma\left[(1+P_\mathrm{r}\gamma)^{-\alpha(\theta)}\right]\right)$ is strictly convex [98], and $\beta\left(w_1 \Psi_{\mathrm{EE_r}}\left(P_{\mathrm{c_r}} + \frac{1}{\epsilon}\overline{P_\mathrm{r}}\right) + (1-w_1)\,\Psi_{\mathrm{EC}}\right)$ is affine in $\overline{P_\mathrm{r}}$, therefore, $\mathcal{S}_\beta$ is strictly convex for any real number β and $U7$ is strictly quasiconvex in $\overline{P_\mathrm{r}}$. This proves Property 1).

We now take the first derivative of (3.7a) with respect to $\overline{P_\mathrm{r}}$, yielding

$$U5' = \frac{\frac{w_1 \Psi_{\mathrm{EE_r}}}{\epsilon} E_\mathrm{c} - J(\overline{P_\mathrm{r}}) E_\mathrm{c}'}{{E_\mathrm{c}}^2},$$

where $J(\overline{P_\mathrm{r}}) = w_1 \Psi_{\mathrm{EE_r}}\left(P_{\mathrm{c_r}} + \frac{1}{\epsilon}\overline{P_\mathrm{r}}\right) + (1-w_1)\Psi_{\mathrm{EC}}$ and $E_\mathrm{c}' = \frac{\mathrm{d}E_\mathrm{c}}{\mathrm{d}\overline{P_\mathrm{r}}}$. When $\overline{P_\mathrm{r}} \to 0$, $E_\mathrm{c} \to 0$, $J(\overline{P_\mathrm{r}}) > 0$ and $E_\mathrm{c}' > 0$, therefore, $U5'\big|_{\overline{P_\mathrm{r}} \to 0} < 0$. On the other hand, when

$\overline{P_{\rm r}} \to \infty$, we have

$$\lim_{\overline{P_{\rm r}}\to\infty} \frac{\frac{w_1\Psi_{\rm EE_r}}{\epsilon}E_{\rm c}}{J(\overline{P_{\rm r}})E'_{\rm c}} = \lim_{\overline{P_{\rm r}}\to\infty} \frac{\frac{w_1\Psi_{\rm EE_r}}{\epsilon}E'_{\rm c}}{\frac{w_1\Psi_{\rm EE_r}}{\epsilon}E'_{\rm c} + J(\overline{P_{\rm r}})E''_{\rm c}}, \tag{C.1}$$

where $E''_{\rm c} = \dfrac{{\rm d}^2E_{\rm c}}{{\rm d}\overline{P_{\rm r}}^2}$. We note that $E''_{\rm c} < 0$, due to the fact that EC is strictly concave in $\overline{P_{\rm r}}$ [98]. Now, by using the fact that $J(\overline{P_{\rm r}}) > 0$, one can show that the RHS of (C.1) is bigger than 1, which means that $U5' \mid_{\overline{P_{\rm r}}\to\infty} > 0$. Hence, when $\overline{P_{\rm r}} \to \infty$, $U5$ is an increasing function in $\overline{P_{\rm r}}$. We note that $U7$ is derived by canceling the negative multiplied constant in $U5$ and then inverting the objective function. Therefore, when $\overline{P_{\rm r}} \to 0$, $U7$ monotonically decreases and when $\overline{P_{\rm r}} \to \infty$, $U7$ monotonically increases. This proves that $U7$ has a cup shape curve in $\overline{P_{\rm r}}$, which completes the proof for Property 2).

Now, we set $f(\overline{P_{\rm r}}) = \ln\left(\mathbb{E}_\gamma\left[(1+P_{\rm r}\gamma)^{-\alpha(\theta)}\right]\right)$ and take the first derivative of $U7$ with respect to $\overline{P_{\rm r}}$ to get

$$\begin{aligned} U7' &= \lim_{\Delta\overline{P_{\rm r}}\to 0} \frac{\dfrac{f\left(\overline{P_{\rm r}}+\Delta\overline{P_{\rm r}}\right)}{J\left(\overline{P_{\rm r}}+\Delta\overline{P_{\rm r}}\right)} - \dfrac{f\left(\overline{P_{\rm r}}\right)}{J\left(\overline{P_{\rm r}}\right)}}{\Delta\overline{P_{\rm r}}} \\ &= \lim_{\Delta\overline{P_{\rm r}}\to 0} \frac{\dfrac{f\left(\overline{P_{\rm r}}+\Delta\overline{P_{\rm r}}\right)-f\left(\overline{P_{\rm r}}\right)}{\Delta\overline{P_{\rm r}}} - \dfrac{w_1\Psi_{\rm EE_r}}{\epsilon}U7}{J\left(\overline{P_{\rm r}}+\Delta\overline{P_{\rm r}}\right)} \\ &= \lim_{\Delta\overline{P_{\rm r}}\to 0} \frac{f(\overline{P_{\rm r}})' - \dfrac{w_1\Psi_{\rm EE_r}}{\epsilon}U7}{J\left(\overline{P_{\rm r}}+\Delta\overline{P_{\rm r}}\right)}. \end{aligned}$$

Therefore, ${\rm sgn}(U7') = {\rm sgn}\left(f(\overline{P_{\rm r}})' - \dfrac{w_1\Psi_{\rm EE_r}}{\epsilon}U7\right)$. This completes the proof of Property 3). □

Appendix D

Proof of Lemma 2

Proof. Here, we briefly prove that problem (P') is a convex program in (y, ϕ), and if (y^*, ϕ^*) is an optimal solution of (P'), then $x^* = y^*/\phi^*$ is an optimal solution of (P).

Since f is a convex function, therefore, for the objective function of problem (P'), we have

$$\begin{aligned}
&(\lambda\phi_1 + (1-\lambda)\phi_2)\, f\left(\frac{\lambda y_1 + (1-\lambda)y_2}{\lambda\phi_1 + (1-\lambda)\phi_2}\right) \\
=&(\lambda\phi_1 + (1-\lambda)\phi_2)\, f\left(\frac{\lambda\phi_1}{\lambda\phi_1 + (1-\lambda)\phi_2}\frac{y_1}{\phi_1} + \frac{(1-\lambda)\phi_2}{\lambda\phi_1 + (1-\lambda)\phi_2}\frac{y_2}{\phi_2}\right) \\
\leq&\lambda\phi_1 f\left(\frac{y_1}{\phi_1}\right) + (1-\lambda)\phi_2 f\left(\frac{y_2}{\phi_2}\right)
\end{aligned}$$

for any (y_1, ϕ_1), $(y_2, \phi_2) \in R^n \times R_+$, and $\lambda \in [0, 1]$. Hence, the objective function of problem (P') is convex in (y, ϕ).

Now, since g is affine, which is also convex, $\phi g(y/\phi)$ can be proved to be convex, by following similar steps. Therefore, the feasible constraint set is a convex set and one can conclude that problem (P') is a convex program if f is convex and g is an affine function on S.

Henceforth, from the Charnes-Cooper transformation, we note that if the optimal solution (y^*, ϕ^*) of problem (P') is found, then $x^* = y^*/\phi^*$ is optimal for problem (P). □

Appendix E

Proof of Lemma 3

Proof. Recall that the proposed optimal power allocation strategy is given as

$$P_{\mathrm{r}}^{*}=\left[\frac{\alpha(\theta)^{\frac{1}{1+\alpha(\theta)}}}{(w_1\nu)^{\frac{1}{1+\alpha(\theta)}}\gamma^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}-\frac{1}{\gamma}\right]^{+}, \tag{E.1}$$

where $\nu=\dfrac{\lambda\Psi_{\mathrm{EE_r}}}{\epsilon}\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}^{*}\gamma)^{-\alpha(\theta)}\right]$ and $[x]^{+}=\max\{0,x\}$.

Let us first calculate the closed-form for $\overline{P_{\mathrm{r}}^{*}}$, considering the unit-variance Nakagami-m block fading channel.

$$\overline{P_{\mathrm{r}}^{*}}=\int_{\gamma_0}^{\infty}\left(\frac{\alpha(\theta)^{\frac{1}{1+\alpha(\theta)}}}{(w_1\nu)^{\frac{1}{1+\alpha(\theta)}}\gamma^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}-\frac{1}{\gamma}\right)\frac{m^m\gamma^{m-1}}{\Gamma(m)}\,\mathrm{e}^{-m\gamma}d\gamma, \tag{E.2}$$

where γ_0 is the threshold calculated by setting P_{r}^{*} as 0, i.e., $\gamma_0=\dfrac{w_1\nu}{\alpha(\theta)}$. Then, we get

$$\overline{P_{\mathrm{r}}^{*}}=\frac{\left(\dfrac{\alpha(\theta)}{w_1\nu}\right)^{\frac{1}{1+\alpha(\theta)}}m^m}{\Gamma(m)}\int_{\gamma_0}^{\infty}\gamma^{m-1-\frac{\alpha(\theta)}{1+\alpha(\theta)}}e^{-m\gamma}d\gamma-\frac{m^m}{\Gamma(m)}\int_{\gamma_0}^{\infty}\gamma^{m-2}e^{-m\gamma}d\gamma. \tag{E.3}$$

By setting a new variable $t = m\gamma$, (E.3) can be rewritten as

$$\overline{P_{\mathrm{r}}^{*}} = \frac{\left(\frac{\alpha(\theta)}{w_1\nu}\right)^{\frac{1}{1+\alpha(\theta)}} m^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma(m)} \int_{m\gamma_0}^{\infty} t^{m-1-\frac{\alpha(\theta)}{1+\alpha(\theta)}} e^{-t} dt - \frac{m}{\Gamma(m)} \int_{m\gamma_0}^{\infty} t^{m-2} e^{-t} dt \quad \text{(E.4)}$$

$$= \frac{\left(\frac{\alpha(\theta)}{w_1\nu}\right)^{\frac{1}{1+\alpha(\theta)}} m^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma(m)\left(m - \frac{\alpha(\theta)}{1+\alpha(\theta)}\right)} \int_{m\gamma_0}^{\infty} e^{-t} dt^{m-\frac{\alpha(\theta)}{1+\alpha(\theta)}} - \frac{m}{\Gamma(m)(m-1)} \int_{m\gamma_0}^{\infty} e^{-t} dt^{m-1}$$

$$= \frac{\left(\frac{\alpha(\theta)}{w_1\nu}\right)^{\frac{1}{1+\alpha(\theta)}} m^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma(m)\left(m - \frac{\alpha(\theta)}{1+\alpha(\theta)}\right)} \left(-e^{-m\gamma_0} (m\gamma_0)^{m-\frac{\alpha(\theta)}{1+\alpha(\theta)}} + \int_{m\gamma_0}^{\infty} t^{m+\frac{1}{1+\alpha(\theta)}-1} e^{-t} dt \right)$$

$$- \frac{m}{\Gamma(m)(m-1)} \left(-e^{-m\gamma_0} (m\gamma_0)^{m-1} + \int_{m\gamma_0}^{\infty} t^{m-1} e^{-t} dt \right). \quad \text{(E.5)}$$

Note that $\Gamma(a,x) = \int_x^{\infty} t^{a-1} \mathrm{e}^{-t}\, dt$ [100]. Therefore, one can get that $\int_{m\gamma_0}^{\infty} t^{m-1} e^{-t} dt = \Gamma(m, m\gamma_0)$ and $\int_{m\gamma_0}^{\infty} t^{m+\frac{1}{1+\alpha(\theta)}-1} e^{-t} dt = \Gamma(m+\frac{1}{1+\alpha(\theta)}, m\gamma_0)$. Henceforth, by applying the upper incomplete gamma function and inserting $\gamma_0 = \frac{w_1\nu^*}{\alpha(\theta)}$[1] into (E.5), we can finally get that

$$\overline{P_{\mathrm{r}}^{*}} = \frac{\left(\frac{\alpha(\theta)}{w_1\nu^*}\right)^{\frac{1}{1+\alpha(\theta)}} m^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma(m)\left(m-\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)} \left[-\left(\frac{w_1\nu^* m}{\alpha(\theta)}\right)^{\left(m-\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)} \mathrm{e}^{-\frac{w_1\nu^* m}{\alpha(\theta)}} + \Gamma\left(m+\frac{1}{1+\alpha(\theta)}, \frac{w_1\nu^* m}{\alpha(\theta)}\right) \right]$$

$$- \frac{m}{\Gamma(m)(m-1)} \left[-\left(\frac{w_1\nu^* m}{\alpha(\theta)}\right)^{m-1} \mathrm{e}^{-\frac{w_1\nu^* m}{\alpha(\theta)}} + \Gamma\left(m, \frac{w_1\nu^* m}{\alpha(\theta)}\right) \right], \text{when } m \neq 1, m \neq \frac{\alpha(\theta)}{\alpha(\theta)+1}. \quad \text{(E.6)}$$

To make (E.6) feasible, the necessary conditions are $m \neq 1$, and $m \neq \frac{\alpha(\theta)}{\alpha(\theta)+1}$. Now, we calculate the closed-form expression for $\overline{P_{\mathrm{r}}^{*}}$ when $m = 1$.

When $m = 1$, the channel distribution becomes Rayleigh fading and $f_\gamma(\gamma) = e^{-\gamma}$.

[1]Here, ν^* is the optimal value for v.

Hence, we can get that

$$\overline{P_{\mathrm{r}}^{*}} = \int_{\gamma_0}^{\infty} \left(\frac{\alpha(\theta)^{\frac{1}{1+\alpha(\theta)}}}{(w_1\nu)^{\frac{1}{1+\alpha(\theta)}}\gamma^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}} - \frac{1}{\gamma} \right) e^{-\gamma} d\gamma$$

$$= \left(\frac{\alpha(\theta)}{w_1\nu}\right)^{\frac{1}{1+\alpha(\theta)}} \int_{\gamma_0}^{\infty} \gamma^{\frac{1}{1+\alpha(\theta)}-1} e^{-\gamma} d\gamma - \int_{\gamma_0}^{\infty} \frac{e^{-\gamma}}{\gamma} d\gamma. \tag{E.7}$$

Note that $\mathrm{E}_1(x) = \int_x^{\infty} \frac{\mathrm{e}^{-z}}{z}\, dz$. By applying $\mathrm{E}_1(x)$ and inserting $\gamma_0 = \frac{w_1\nu^*}{\alpha(\theta)}$, (E.7) can be expressed as

$$\overline{P_{\mathrm{r}}^{*}} = \left(\frac{\alpha(\theta)}{w_1\nu^*}\right)^{\frac{1}{1+\alpha(\theta)}} \Gamma\left(\frac{1}{1+\alpha(\theta)}, \frac{w_1\nu^*}{\alpha(\theta)}\right) - \mathrm{E}_1\left(\frac{w_1\nu^*}{\alpha(\theta)}\right), \text{ when } m = 1. \tag{E.8}$$

Finally, let us calculate the closed-form expression for $\overline{P_{\mathrm{r}}^{*}}$, when $m = \frac{\alpha(\theta)}{\alpha(\theta)+1}$. By inserting $m = \frac{\alpha(\theta)}{\alpha(\theta)+1}$ into (E.4), one can get that

$$\overline{P_{\mathrm{r}}^{*}} = \frac{\left(\frac{\alpha(\theta)}{w_1\nu}\right)^{\frac{1}{1+\alpha(\theta)}} \left(\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma\left(\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)} \mathrm{E}_1\left(\frac{\alpha(\theta)\gamma_0}{1+\alpha(\theta)}\right) - \frac{\frac{\alpha(\theta)}{1+\alpha(\theta)}}{\Gamma(\frac{\alpha(\theta)}{1+\alpha(\theta)})} \int_{\frac{\alpha(\theta)\gamma_0}{1+\alpha(\theta)}}^{\infty} t^{\frac{\alpha(\theta)}{1+\alpha(\theta)}-2} e^{-t} dt \tag{E.9}$$

$$= \frac{\left(\frac{\alpha(\theta)}{w_1\nu}\right)^{\frac{1}{1+\alpha(\theta)}} \left(\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)^{\frac{\alpha(\theta)}{1+\alpha(\theta)}}}{\Gamma\left(\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)} \mathrm{E}_1\left(\frac{\alpha(\theta)\gamma_0}{1+\alpha(\theta)}\right)$$

$$- \frac{\frac{\alpha(\theta)}{1+\alpha(\theta)}}{\Gamma\left(\frac{\alpha(\theta)}{1+\alpha(\theta)}\right)\left(\frac{\alpha(\theta)}{1+\alpha(\theta)}-1\right)} \int_{\frac{\alpha(\theta)\gamma_0}{1+\alpha(\theta)}}^{\infty} e^{-t} dt^{\frac{\alpha(\theta)}{1+\alpha(\theta)}-1}. \tag{E.10}$$

By applying the upper incomplete gamma function and $\mathrm{E}_1(x)$, (E.10) can be expressed

as

$$\overline{P_{\mathrm{r}}^{*}}=\frac{\left(\frac{\alpha(\theta)}{w_1\nu^{*}}\right)^{\frac{1}{\alpha(\theta)+1}}\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)^{\frac{\alpha(\theta)}{\alpha(\theta)+1}}}{\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)}\mathrm{E}_1\left(\frac{w_1\nu^{*}}{1+\alpha(\theta)}\right)+\frac{\alpha(\theta)}{\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1}\right)}$$
$$\times\left[-\mathrm{e}^{-\frac{w_1\nu^{*}}{\alpha(\theta)+1}}\left(\frac{w_1\nu^{*}}{\alpha(\theta)+1}\right)^{-\frac{1}{\alpha(\theta)+1}}+\Gamma\left(\frac{\alpha(\theta)}{\alpha(\theta)+1},\frac{w_1\nu^{*}}{\alpha(\theta)+1}\right)\right],\ \text{when } m=\frac{\alpha(\theta)}{\alpha(\theta)+1}. \tag{E.11}$$

Hence, the closed-form expressions for $\overline{P_{\mathrm{r}}^{*}}$ have been given in (E.6), (E.8), (E.11), for $m=1$, $m=\frac{\alpha(\theta)}{\alpha(\theta)+1}$, and the other cases. The closed-form expressions for $\mathbb{E}_{\gamma}\left[(1+P_{\mathrm{r}}^{*}\gamma)^{-\alpha(\theta)}\right]$ can be derived by following similar steps and the proof is omitted here for simplicity. □

Appendix F

Proof of Lemma 4

Proof. For a system with normalization values $\Psi_{\mathrm{EE},1} = \mathrm{EE}\,|_{\overline{P_\mathrm{t}}=P_{\mathrm{norm},1}}$ and $\Psi_{\mathrm{EC},1} = E_\mathrm{c}\,|_{\overline{P_\mathrm{t}}=P_{\mathrm{norm},1}}$, the optimal average transmit power is assumed to be P_1^*. Taking the first derivative of the function $U5$ with respect to the average transmit power, it yields

$$
\begin{aligned}
&U5'\Big|_{\substack{\overline{P_\mathrm{t}}=P_1^* \\ P_\mathrm{norm}=P_{\mathrm{norm},1}}} \\
&= \frac{\dfrac{w_1\Psi_{\mathrm{EE},1}}{\epsilon}E_\mathrm{c} - \left(w_1\Psi_{\mathrm{EE},1}\left(P_\mathrm{c}+\dfrac{1}{\epsilon}\overline{P_\mathrm{t}}\right)+(1-w_1)\,\Psi_{\mathrm{EC},1}\right)E_\mathrm{c}'}{E_\mathrm{c}^2} && \text{(F.1a)} \\
&= \Psi_{\mathrm{EE},1}\left(\frac{w_1 E_\mathrm{c}}{\epsilon} - \left(\frac{w_1}{\epsilon}P_1^* + P_\mathrm{c} + \frac{(1-w_1)}{\epsilon}P_{\mathrm{norm},1}\right)E_\mathrm{c}'\right), && \text{(F.1b)}
\end{aligned}
$$

which equals to 0 at the optimal point P_1^*.

Then, let us consider a system with a larger P_norm, i.e., $P_{\mathrm{norm},2} = P_{\mathrm{norm},1} + \Delta P_\mathrm{norm}$, $\Delta P_\mathrm{norm} > 0$. Correspondingly, the normalization values are denoted by $\Psi_{\mathrm{EE},2}$ and $\Psi_{\mathrm{EC},2}$, where $\Psi_{\mathrm{EE},2} = \mathrm{EE}\,|_{\overline{P_\mathrm{t}}=P_{\mathrm{norm},2}}$ and $\Psi_{\mathrm{EC},2} = E_\mathrm{c}\,|_{\overline{P_\mathrm{t}}=P_{\mathrm{norm},2}}$. In this case, the optimal average transmit power at which the tradeoff formulation can be maximized is denoted by P_2^*. Replacing $P_{\mathrm{norm},1}$ in (F.1a) with $P_{\mathrm{norm},2}$, we have

$$
\begin{aligned}
&U5'\Big|_{\substack{\overline{P_\mathrm{t}}=P_1^* \\ P_\mathrm{norm}=P_{\mathrm{norm},1}+\Delta P_\mathrm{norm}}} \\
&= \frac{\Psi_{\mathrm{EE},2}\left[\dfrac{w_1 E_\mathrm{c}}{\epsilon} - \left(\dfrac{w_1}{\epsilon}P_1^* + P_\mathrm{c} + \dfrac{(1-w_1)}{\epsilon}\left(P_{\mathrm{norm},1}+\Delta P_\mathrm{norm}\right)\right)E_\mathrm{c}'\right]}{E_\mathrm{c}^2}. && \text{(F.2a)}
\end{aligned}
$$

By using (F.1b), (F.2a) reduces to

$$
U5'\Big|_{\substack{\overline{P_\mathrm{t}}=P_1^* \\ P_\mathrm{norm}=P_{\mathrm{norm},1}+\Delta P_\mathrm{norm}}} = -\frac{\Psi_{\mathrm{EE},2}\dfrac{(1-w_1)}{\epsilon}\Delta P_\mathrm{norm}E_\mathrm{c}'}{E_\mathrm{c}^2} < 0. \qquad \text{(F.3)}
$$

From Theorem 6, it is noted that $U5$ strictly decreases with the average transmit power until reaching the minimum, then it becomes a monotonically increasing function. Therefore, (F.3) means that $U5$ with a larger P_{norm} decreases at P_1^* and has not reached its minimum yet, which means P_2^* must be larger than P_1^*. Hence, we complete the proof which shows that when the normalization factor becomes larger, the optimal average power value increases.

Further, it is easy to prove that the optimal average power monotonically decreases with w_1, therefore the proof is omitted here. This completes the proof of Lemma 4. □

Appendix G

Proof of Lemma 6

Proof. Assume that the final calculated tradeoff EE for the k^{th} user, obtained with a circuit power value $P_{\text{c},1}^k$ and N_k allocated subcarriers, is achieved at the average power $\overline{P_{k,1}^*}$, i.e., $\text{EE}^k\Big|_{\substack{P_{\text{c}}^k=P_{\text{c},1}^k \\ \overline{P_k}=\overline{P_{k,1}^*}}}$. Meanwhile, the maximum achievable EE value for the k^{th} user, which is calculated by assuming all N subcarriers in the system are allocated to it, is assumed to be achieved at $\overline{P_{\text{EE},1}^{k*}}$, i.e., $\eta_{\text{max},1}^{k,N} = \text{EE}^k\Big|_{\substack{N_k=N \\ P_{\text{c}}^k=P_{\text{c},1}^k \\ \overline{P_k}=\overline{P_{\text{EE},1}^{k*}}}}$.

If the k^{th} user has a higher circuit power, i.e, $P_{\text{c},2}^k = P_{\text{c},1}^k + \Delta P_{\text{c}}^k$, $\Delta P_{\text{c}}^k > 0$, the calculated tradeoff EE for the k^{th} user, obtained with N_k allocated subcarriers, is assumed to be achieved at $\overline{P_{k,2}^*}$, i.e., $\text{EE}^k\Big|_{\substack{P_{\text{c}}^k=P_{\text{c},2}^k \\ \overline{P_k}=\overline{P_{k,2}^*}}}$. Meanwhile, the maximum achievable EE value for the k^{th} user, which is calculated by assuming all N subcarriers are allocated to it, is assumed to be achieved at $\overline{P_{\text{EE},2}^{k*}}$, i.e., $\eta_{\text{max},2}^{k,N} = \text{EE}^k\Big|_{\substack{N_k=N \\ P_{\text{c}}^k=P_{\text{c},2}^k \\ \overline{P_k}=\overline{P_{\text{EE},2}^{k*}}}}$.

Since we note that the optimal average power value is found when the k^{th} user's EE constraint is satisfied with equality, hence, from (4.19), we have the following equations:

$$\text{EE}^k\Big|_{\substack{P_{\text{c}}^k=P_{\text{c},1}^k \\ \overline{P_k}=\overline{P_{k,1}^*}}} = \chi_{\text{EE}}^k \times \eta_{\text{max},1}^{k,N} = \chi_{\text{EE}}^k \times \text{EE}^k\Big|_{\substack{N_k=N \\ P_{\text{c}}^k=P_{\text{c},1}^k \\ \overline{P_k}=\overline{P_{\text{EE},1}^{k*}}}}, \tag{G.1a}$$

$$\text{EE}^k\Big|_{\substack{P_{\text{c}}^k=P_{\text{c},2}^k \\ \overline{P_k}=\overline{P_{k,2}^*}}} = \chi_{\text{EE}}^k \times \eta_{\text{max},2}^{k,N} = \chi_{\text{EE}}^k \times \text{EE}^k\Big|_{\substack{N_k=N \\ P_{\text{c}}^k=P_{\text{c},2}^k \\ \overline{P_k}=\overline{P_{\text{EE},2}^{k*}}}}. \tag{G.1b}$$

In order to investigate the influence of circuit power P_{c}^k on the tradeoff EE value, we start from analyzing the effect of P_{c}^k on the maximum EE value $\eta_{\text{max}}^{k,N}$. For the system with $P_{\text{c},1}^k$, if we assume the operational average input power is $\overline{P_{\text{EE},2}^{k*}}$, the

corresponding calculated link-layer EE value becomes

$$\mathrm{EE}^k\Big|_{\substack{N_k=N\\ P_{\mathrm{c}}^k=P_{\mathrm{c},1}^k\\ \overline{P_k}=\overline{P_{\mathrm{EE},2}^{k*}}}} = \frac{E_{\mathrm{c}}^k\Big|_{\substack{N_k=N\\ \overline{P_k}=\overline{P_{\mathrm{EE},2}^{k*}}}}}{P_{\mathrm{c},1}^k + \frac{1}{\epsilon}\overline{P_{\mathrm{EE},2}^{k*}}}. \tag{G.2}$$

Meanwhile, for the system with $P_{\mathrm{c},2}^k$, the maximum achievable EE value $\eta_{\mathrm{max},2}^{k,N}$ can be expanded as

$$\mathrm{EE}^k\Big|_{\substack{N_k=N\\ P_{\mathrm{c}}^k=P_{\mathrm{c},2}^k\\ \overline{P_k}=\overline{P_{\mathrm{EE},2}^{k*}}}} = \frac{E_{\mathrm{c}}^k\Big|_{\substack{N_k=N\\ \overline{P_k}=\overline{P_{\mathrm{EE},2}^{k*}}}}}{P_{\mathrm{c},2}^k + \frac{1}{\epsilon}\overline{P_{\mathrm{EE},2}^{k*}}}. \tag{G.3}$$

Apparently, we can notice that the link-layer EE value in (G.2) is larger than the one in (G.3), because $P_{\mathrm{c},1}^k < P_{\mathrm{c},2}^k$. This indicates that the calculated link-layer EE value in (G.2), for the system with $P_{\mathrm{c},1}^k$, is larger than the maximum achievable EE value $\eta_{\mathrm{max},2}^{k,N}$, for the system with $P_{\mathrm{c},2}^k$. Furthermore, we note that $\eta_{\mathrm{max},1}^{k,N}$ is larger than the calculated EE value in (G.2), because $\eta_{\mathrm{max},1}^{k,N}$ is the maximum achievable link-layer EE value for the system with the circuit power $P_{\mathrm{c},1}^k$. Henceforth, one can derive that the maximum achievable link-layer EE value $\eta_{\mathrm{max},1}^{k,N}$ for the system with circuit power $P_{\mathrm{c},1}^k$ is larger than the one obtained at a larger circuit power $P_{\mathrm{c},2}^k$, i.e., $\eta_{\mathrm{max},2}^{k,N}$. This means that with a fixed number of allocated subcarriers, when one user's circuit power becomes larger, its maximum achievable EE value reduces.

Since the k^{th} user's link-layer tradeoff EE value, achieved at its EE requirement equality, is basically a certain ratio of its maximum achievable EE value, hence, from (G.1a)-(G.1b), we can finally conclude that one user's tradeoff EE level also decreases with its circuit power. □

Appendix H

Proof of Theorem 7

Proof. By applying the order statistics, the EC in NOMA for the m^{th} user, E_{c}^{m}, can be expanded as

$$E_{\text{c}}^{m} = -\frac{1}{\theta_m T_{\text{f}} B} \ln \left(\int_0^{\infty} \left(\frac{\gamma_m + 1}{\alpha_n \gamma_m + 1} \right)^{2\beta_m} \psi_m f(\gamma_m) F(\gamma_m)^{m-1} \left(1 - F(\gamma_m)\right)^{M-m} d\gamma_m \right). \tag{H.1}$$

By inserting $f(\gamma_m) = \frac{1}{\rho} e^{-\frac{\gamma_m}{\rho}}$, and $F(\gamma_m) = 1 - e^{-\frac{\gamma_m}{\rho}}$ into (H.1), we have

$$E_{\text{c}}^{m} = -\frac{1}{\theta_m T_{\text{f}} B} \ln \left(\frac{\psi_m}{\rho} \int_0^{\infty} \left(\frac{\gamma_m + 1}{\alpha_n \gamma_m + 1} \right)^{2\beta_m} e^{-\frac{(M-m+1)\gamma_m}{\rho}} \left(1 - e^{-\frac{\gamma_m}{\rho}} \right)^{m-1} d\gamma_m \right). \tag{H.2}$$

According to the generalized binomial expansion, we first get the transformation

$$\left(\frac{\gamma_m + 1}{\gamma_m \alpha_n + 1} \right)^{2\beta_m} = \left(\frac{1}{\alpha_n} \right)^{2\beta_m} \left(1 + \frac{\alpha_n - 1}{\gamma_m \alpha_n + 1} \right)^{2\beta_m}, \tag{H.3}$$

where $\left(1 + \frac{\alpha_n - 1}{\gamma_m \alpha_n + 1} \right)^{2\beta_m}$ can then be expanded as $\sum_{j=0}^{\infty} \binom{2\beta_m}{j} \left(\frac{\alpha_n - 1}{\gamma_m \alpha_n + 1} \right)^{j}$, due to the fact that $(1+x)^s = \sum_{j=0}^{\infty} \binom{s}{j} x^j$, for $|x| < 1$, where $\binom{s}{j}$ is defined as follows [100]:

$$\binom{s}{j} = \frac{s(s-1)\ldots(s-j+1)}{j!} = \frac{(s)_j}{j!}, \quad \text{if } j \geq 1, \tag{H.4}$$

where $(\cdot)_j$ is the Pochhammer symbol, and $\binom{s}{0} = 1$ [100].

Then, $\left(1-e^{-\frac{\gamma_m}{\rho}}\right)^{m-1}$ can be replaced with $\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k e^{-\frac{\gamma_m}{\rho}k}$, by using the binomial expansion [100]. Therefore, (H.2) can be transformed into

$$E_{\mathrm{c}}^{m}=-\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{(\alpha_n)^{-2\beta_m}\,\psi_m}{\rho}\int_0^{\infty}\left(\sum_{j=0}^{\infty}\binom{2\beta_m}{j}\left(\frac{\alpha_n-1}{\gamma_m\alpha_n+1}\right)^j\right)\right.$$
$$\left.\times\left(\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}\right)d\gamma_m\right) \tag{H.5a}$$

$$=-\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{(\alpha_n)^{-2\beta_m}\,\psi_m}{\rho}\int_0^{\infty}\left(\underbrace{1}_{\text{when j=0}}+\underbrace{(2\beta_m)\,\frac{\alpha_n-1}{\gamma_m\alpha_n+1}}_{\text{when j =1}}+\underbrace{\sum_{j=2}^{\infty}\binom{2\beta_m}{j}\left(\frac{\alpha_n-1}{\gamma_m\alpha_n+1}\right)^j}_{\text{when j}\geq 2}\right)\right.$$
$$\left.\times\left(\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}\right)d\gamma_m\right) \tag{H.5b}$$

$$=-\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{(\alpha_n)^{-2\beta_m}\,\psi_m}{\rho}\left(\int_0^{\infty}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}\,d\gamma_m\right.\right.$$
$$+(2\beta_m)\,(\alpha_n-1)\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k\int_0^{\infty}\frac{e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}}{\gamma_m\alpha_n+1}d\gamma_m$$
$$\left.\left.+\sum_{j=2}^{\infty}\binom{2\beta_m}{j}(\alpha_n-1)^j\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k\int_0^{\infty}\frac{e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}}{(\gamma_m\alpha_n+1)^j}d\gamma_m\right)\right). \tag{H.5c}$$

We now use the following equation from [114], namely, (3.353.2) and (3.352.4).

$$\int_0^{\infty}\frac{e^{-\mu x}}{(x+\beta)^n}dx=\frac{1}{(n-1)!}\sum_{k=1}^{n-1}(k-1)!(-\mu)^{n-k-1}\beta^{-k}-\frac{(-\mu)^{n-1}}{(n-1)!}e^{\beta\mu}E_i(-\beta\mu),$$
$$[n\geq 2,\ |\arg\,\beta|<\pi,\ \mathrm{Re}\,\mu>0] \tag{H.6a}$$

$$\int_0^{\infty}\frac{e^{-\mu x}}{x+\beta}dx=-e^{\beta\mu}E_i(-\beta\mu),\quad[|\arg\,\beta|<\pi,\ \mathrm{Re}\,\mu>0]\,, \tag{H.6b}$$

where $E_i(\cdot)$ is the exponential integral.

By applying (H.6a) and (H.6b) into (H.5c), the closed-form expression for E_{c}^{m} can

be finally expressed as

$$E_{\mathrm{c}}^{m} = -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{(\alpha_n)^{-2\beta_m}\, \psi_m}{\rho} \left(\sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k \frac{\rho}{M-m+1+k} \right. \right.$$
$$+ \frac{\theta_m(\alpha_n - 1)}{\alpha_n \ln 2} \sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k e^{\frac{M-m+1+k}{\rho\alpha_n}} E_i\left(-\frac{M-m+1+k}{\rho\alpha_n}\right)$$
$$+ \sum_{j=2}^{\infty} \binom{2\beta_m}{j} \left(\frac{\alpha_n - 1}{\alpha_n}\right)^j \sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k \left(\frac{1}{(j-1)!} \sum_{i=1}^{j-1} \frac{(i-1)!}{\left(\frac{1}{\alpha_n}\right)^i} \left(-\frac{M-m+1+k}{\rho}\right)^{j-i-1} \right.$$
$$\left. \left. \left. \left. - \frac{\left(-\frac{M-m+1+k}{\rho}\right)^{j-1}}{(j-1)!}\, e^{\frac{M-m+1+k}{\rho\alpha_n}} E_i\left(-\frac{M-m+1+k}{\rho\alpha_n}\right) \right) \right) \right) \right). \tag{H.7}$$

Now, let us consider the closed-form expression for the EC in OMA scheme for the m^{th} user. By applying the order statistics, the EC in OMA for the m^{th} user, $\bar{E}_{\mathrm{c}}^{m}$, can be expanded as

$$\bar{E}_{\mathrm{c}}^{m} = -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\psi_m}{\rho} \int_0^{\infty} (1+\gamma_m)^{\beta_m}\, e^{-\frac{(M-m+1)\gamma_m}{\rho}} \left(1 - e^{-\frac{\gamma_m}{\rho}}\right)^{m-1} d\gamma_m \right). \tag{H.8}$$

After applying the binomial expansion for $\left(1 - e^{-\frac{\gamma_m}{\rho}}\right)^{m-1}$, we have

$$\bar{E}_{\mathrm{c}}^{m} = -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\psi_m}{\rho} \sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k \int_0^{\infty} (1+\gamma_m)^{\beta_m}\, e^{-\frac{(M-m+1+k)\gamma_m}{\rho}} d\gamma_m \right). \tag{H.9}$$

From (13.2.5) in [100], we note that

$$U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^{\infty} e^{-zt} t^{a-1} (1+t)^{b-a-1} dt, \quad \text{for Re } a,\ \text{Re } z > 0, \tag{H.10}$$

where $U(\cdot)$ is the confluent hypergeometric function of the second kind [100]. By applying (H.10) to (H.9), $\bar{E}_{\mathrm{c}}^{m}$ can be finally expressed as

$$\bar{E}_{\mathrm{c}}^{m} = -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\psi_m}{\rho} \sum_{k=0}^{m-1} \binom{m-1}{k} (-1)^k U\left(1, 2+\beta_m, \frac{M-m+1+k}{\rho}\right) \right). \tag{H.11}$$

□

Appendix I

Proof of Lemma 8

Proof. Inserting $\rho \to 0$ into (5.7a), (5.8a), (5.7b), and (5.8b), one can prove that $E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m} \to 0$, and $E_{\mathrm{c}}^{n} - \bar{E}_{\mathrm{c}}^{n} \to 0$. When $\rho \to \infty$, the EC in NOMA for the m^{th} user, i.e., E_{c}^{m}, can be expressed as

$$\lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln\left(\mathbb{E}\left[\left(\frac{|h_m|^2 + \frac{1}{\rho}}{\alpha_n |h_m|^2 + \frac{1}{\rho}}\right)^{2\beta_m}\right]\right) = \log_2\left(\frac{1}{\alpha_n}\right). \tag{I.1}$$

Considering finite value of θ_m, we can prove that $\lim_{\rho\to\infty} \bar{E}_{\mathrm{c}}^{m}$ becomes

$$\begin{aligned}
&\lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln\left(\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]\right) \\
&= \lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln\left(\rho^{\beta_m}\mathbb{E}\left[\left(\frac{1}{\rho} + |h_m|^2\right)^{\beta_m}\right]\right) \\
&= \lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B} \left(\ln\left(\rho^{\beta_m}\right) + \ln\left(\mathbb{E}\left[\left(\frac{1}{\rho} + |h_m|^2\right)^{\beta_m}\right]\right)\right) \\
&= \lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln\left(\rho^{\beta_m}\right) \\
&= \lim_{\rho\to\infty} \frac{1}{2\ln 2} \ln\rho,
\end{aligned} \tag{I.2}$$

which approaches infinity. From (I.1) and (I.2), we get $\lim_{\rho\to\infty}\left(E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m}\right) \to -\infty$.

As for the n^{th} user, $\lim_{\rho\to\infty} E_{\mathrm{c}}^{n} \to \infty$, and $\lim_{\rho\to\infty} \bar{E}_{\mathrm{c}}^{n} \to \infty$ can be easily proved, which are omitted here. To analyze the EC difference of the NOMA and OMA scheme for

the n^{th} user when $\rho \to \infty$, we have that

$$\lim_{\rho\to\infty} \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right) \tag{I.3a}$$

$$= \lim_{\rho\to\infty} -\frac{1}{\theta_n T_{\text{f}} B} \ln\left(\frac{\mathbb{E}\left[(1+\rho\alpha_n|h_n|^2)^{2\beta_n}\right]}{\mathbb{E}\left[(1+\rho|h_n|^2)^{\beta_n}\right]}\right) \tag{I.3b}$$

$$= \lim_{\rho\to\infty} -\frac{1}{\theta_n T_{\text{f}} B} \ln\left(\frac{\rho^{\beta_n}\mathbb{E}\left[(\alpha_n|h_n|^2)^{2\beta_n}\right]}{\mathbb{E}\left[(|h_n|^2)^{\beta_n}\right]}\right), \tag{I.3c}$$

which approaches infinity. This completes the proof that $\lim_{\rho\to\infty} \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right) \to \infty$. □

Appendix J

Proof of Lemma 9

Proof. To analyze the trends of E_{c}^m and $\bar{E}_{\mathrm{c}}^m$ with respect to ρ, we have

$$\frac{\partial E_{\mathrm{c}}^m}{\partial \rho} = -\frac{1}{\theta_m T_{\mathrm{f}} B} \frac{\left(\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]\right)'}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]} \tag{J.1a}$$

$$= \frac{1-\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m-1} \frac{|h_m|^2}{\left(\rho\alpha_n|h_m|^2+1\right)^2}\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]}, \tag{J.1b}$$

where $()'$ is the first derivative with respect to ρ. Apparently, (J.1b) is non-negative. Similarly, for the EC in OMA for the m^{th} user, we get

$$\frac{\partial \bar{E}_{\mathrm{c}}^m}{\partial \rho} = \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m-1}|h_m|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]}, \tag{J.2}$$

which is non-negative too.

We then start to analyze the trend of $E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m$ with respect to ρ, as follows.

$$\frac{\partial\left(E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m\right)}{\partial \rho} = \frac{\partial E_{\mathrm{c}}^m}{\partial \rho} - \frac{\partial \bar{E}_{\mathrm{c}}^m}{\partial \rho} \tag{J.3a}$$

$$= \frac{1-\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m-1} \frac{|h_m|^2}{\left(\rho\alpha_n|h_m|^2+1\right)^2}\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]} - \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m-1}|h_m|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]}. \tag{J.3b}$$

When $\rho \to 0$, we prove that

$$\lim_{\rho \to 0} \frac{\partial \left(E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m \right)}{\partial \rho} = \left(\frac{1 - 2\alpha_n}{2 \ln 2} \right) \mathbb{E} \left[|h_m|^2 \right] \geq 0, \tag{J.4}$$

due to the reason that $\alpha_n \in \left(0, \frac{1}{2} \right]$ and $\mathbb{E} \left[|h_m|^2 \right] \geq 0$.

When ρ is very large, we can prove that

$$\frac{\partial \left(E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m \right)}{\partial \rho} \tag{J.5a}$$

$$= \frac{1 - \alpha_n}{\ln 2} \frac{\mathbb{E} \left[\left(\frac{\rho |h_m|^2}{\rho \alpha_n |h_m|^2} \right)^{2\beta_m - 1} \frac{|h_m|^2}{\left(\rho \alpha_n |h_m|^2 \right)^2} \right]}{\mathbb{E} \left[\left(\frac{\rho |h_m|^2}{\rho \alpha_n |h_m|^2} \right)^{2\beta_m} \right]} - \frac{1}{2 \ln 2} \frac{\mathbb{E} \left[\left(\rho |h_m|^2 \right)^{\beta_m - 1} |h_m|^2 \right]}{\mathbb{E} \left[\left(\rho |h_m|^2 \right)^{\beta_m} \right]} \tag{J.5b}$$

$$= \frac{\frac{1 - \alpha_n}{\alpha_n \ln 2} \mathbb{E} \left[\frac{1}{|h_m|^2} \right] - \frac{1}{2 \ln 2} \rho}{\rho^2}. \tag{J.5c}$$

Since $\mathbb{E} \left[\frac{1}{|h_m|^2} \right]$ is a finite value, unrelated to ρ, therefore when ρ is very large, (J.5c) can be approximated by $-\frac{1}{2\rho \ln 2}$, which is smaller than 0. It gradually approaches 0 when $\rho \to \infty$. □

Appendix K

Proof of Lemma 10

Proof. Here, we analyze the trends of E_{c}^n and $\bar{E}_{\mathrm{c}}^n$ versus ρ.

$$\frac{\partial E_{\mathrm{c}}^n}{\partial \rho} = -\frac{1}{\theta_n T_{\mathrm{f}} B} \frac{\left(\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]\right)'}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]} = \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]}, \tag{K.1}$$

which is non-negative. As for the EC in OMA, we can also prove that $\frac{\partial \bar{E}_{\mathrm{c}}^n}{\partial \rho} \geq 0$, which is omitted here due to the page limit. To analyze the trend of $E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n$ versus ρ, we have that

$$\frac{\partial\left(E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n\right)}{\partial \rho} = \frac{\partial E_{\mathrm{c}}^n}{\partial \rho} - \frac{\partial \bar{E}_{\mathrm{c}}^n}{\partial \rho} \tag{K.2a}$$

$$= \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]} - \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_n|^2\right)^{\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_n|^2\right)^{\beta_n}\right]}. \tag{K.2b}$$

When $\rho \to 0$, we prove that

$$\lim_{\rho\to 0} \frac{\partial\left(E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n\right)}{\partial \rho} = \left(\frac{\alpha_n - \frac{1}{2}}{\ln 2}\right) \mathbb{E}\left[|h_n|^2\right] \leq 0, \tag{K.3}$$

due to the fact that $\alpha_n \in \left(0, \frac{1}{2}\right]$, and $\mathbb{E}\left[|h_n|^2\right] \geq 0$.

When ρ is very large, we can prove that

$$\frac{\partial\left(E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n\right)}{\partial \rho} = \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]} - \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(\rho|h_n|^2\right)^{\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(\rho|h_n|^2\right)^{\beta_n}\right]} \tag{K.4a}$$

$$= \frac{1}{2\rho\ln 2}, \tag{K.4b}$$

which is non-negative, and approaches 0 when $\rho \to \infty$. □

Appendix L

Proof of Lemma 11

Proof. From Lemma 8, we note that when $\rho \to 0$, $\mathrm{T}_N = E_{\mathrm{c}}^m + E_{\mathrm{c}}^n \to 0$, and $\lim_{\rho\to\infty} \mathrm{T}_N \to \infty$. For the sum EC in OMA scheme, T_O, we can also get that $\mathrm{T}_O \to 0$ when $\rho \to 0$, and $\lim_{\rho\to\infty} \mathrm{T}_O \to \infty$. In addition, for the sum EC in NOMA scheme, T_N, we can prove that

$$\frac{\partial \mathrm{T}_N}{\partial \rho} = \frac{\partial \left(E_{\mathrm{c}}^m + E_{\mathrm{c}}^n\right)}{\partial \rho} \tag{L.1a}$$

$$= \frac{1-\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m-1} \frac{|h_m|^2}{\left(\rho\alpha_n|h_m|^2+1\right)^2}\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]} + \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]}, \tag{L.1b}$$

which is non-negative because $\frac{\partial E_{\mathrm{c}}^m}{\partial \rho} \geq 0$, and $\frac{\partial E_{\mathrm{c}}^n}{\partial \rho} \geq 0$. When $\rho \to 0$, we have that

$$\lim_{\rho\to 0} \frac{\partial \mathrm{T}_N}{\partial \rho} = \frac{1-\alpha_n}{\ln 2}\mathbb{E}\left[|h_m|^2\right] + \frac{\alpha_n}{\ln 2}\mathbb{E}\left[|h_n|^2\right]. \tag{L.2}$$

When $\rho \to \infty$, we can prove that

$$\lim_{\rho\to\infty} \frac{\partial \mathrm{T}_N}{\partial \rho} = \frac{1-\alpha_n}{\alpha_n \ln 2 \rho^2}\mathbb{E}\left[\frac{1}{|h_m|^2}\right] + \frac{1}{\rho \ln 2}, \tag{L.3}$$

which equals to 0.

By following similar steps, we can also prove that $\frac{\partial \mathrm{T}_O}{\partial \rho} \geq 0$, $\lim_{\rho\to 0} \frac{\partial \mathrm{T}_O}{\partial \rho} = \frac{1}{2\ln 2}\mathbb{E}[|h_m|^2] + \frac{1}{2\ln 2}\mathbb{E}\left[|h_n|^2\right]$, and $\lim_{\rho\to\infty} \frac{\partial \mathrm{T}_O}{\partial \rho} = 0$. Hence, we complete the proof for Lemma 11. □

Appendix M

Proof of Lemma 12

Proof. When $\rho \to 0$, one can easily get that $\mathrm{T}_N - \mathrm{T}_O \to 0$. When $\rho \to \infty$, we get

$$\lim_{\rho\to\infty} \left(\mathrm{T}_N - \mathrm{T}_O\right) = \lim_{\rho\to\infty} \left(E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m + E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n\right) \tag{M.1a}$$

$$= \lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\mathbb{E}\left[\left(\dfrac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]}{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]} \right) - \frac{1}{\theta_n T_{\mathrm{f}} B} \ln \left(\frac{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]}{\mathbb{E}\left[\left(1+\rho|h_n|^2\right)^{\beta_n}\right]} \right) \tag{M.1b}$$

$$= \lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\alpha_n^{-2\beta_m}}{\mathbb{E}\left[\left(|h_m|^2\right)^{\beta_m}\right]} \rho^{-\beta_m} \right) - \frac{1}{\theta_n T_{\mathrm{f}} B} \ln \left(\frac{\alpha_n^{2\beta_n}\mathbb{E}\left[\left(|h_n|^2\right)^{2\beta_n}\right]}{\mathbb{E}\left[\left(|h_n|^2\right)^{\beta_n}\right]} \rho^{\beta_n} \right) \tag{M.1c}$$

$$= -\frac{1}{\theta_m T_{\mathrm{f}} B} \ln \left(\frac{\alpha_n^{-2\beta_m}}{\mathbb{E}\left[\left(|h_m|^2\right)^{\beta_m}\right]} \right) - \frac{1}{\theta_n T_{\mathrm{f}} B} \ln \left(\frac{{\alpha_n}^{2\beta_n}\mathbb{E}\left[\left(|h_n|^2\right)^{2\beta_n}\right]}{\mathbb{E}\left[\left(|h_n|^2\right)^{\beta_n}\right]} \right), \tag{M.1d}$$

which is a constant with respect to ρ.

From Lemma 11, we note that

$$\lim_{\rho\to 0} \frac{\partial\left(\mathrm{T}_N - \mathrm{T}_O\right)}{\partial\rho} = \lim_{\rho\to 0} \left(\frac{\partial \mathrm{T}_N}{\partial\rho} - \frac{\partial \mathrm{T}_O}{\partial\rho} \right) = \frac{\frac{1}{2}-\alpha_n}{\ln 2}\mathbb{E}\left[|h_m|^2\right] + \frac{\alpha_n - \frac{1}{2}}{\ln 2}\mathbb{E}\left[|h_n|^2\right], \tag{M.2}$$

which is non-positive because $\alpha_n - \dfrac{1}{2} \leq 0$, and $\mathbb{E}\left[|h_n|^2\right] \geq \mathbb{E}\left[|h_m|^2\right]$, since the instantaneous channel power gains $|h_n|^2$ is always larger than $|h_m|^2$.

When $\rho \to \infty$, one can easily prove that $\lim\limits_{\rho\to\infty} \dfrac{\partial\left(\mathrm{T}_N - \mathrm{T}_O\right)}{\partial\rho} = 0$, which is omitted here. □

Appendix N

Proof of Lemma 13

Proof. Recall that the EC expression in NOMA scheme, for the m^{th} user, is given by

$$E_{\text{c}}^{m} = -\frac{1}{\theta_m T_{\text{f}} B} \ln\left(\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]\right), \tag{N.1}$$

which gives an indeterminate form $\frac{0}{0}$, when $\theta_m \to 0$.

By applying L'Hopital's rule, $\lim\limits_{\theta_m\to 0} E_{\text{c}}^{m}$ becomes

$$\begin{aligned}
&\lim_{\theta_m\to 0} -\frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m} \ln\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)\left(-\frac{1}{\ln 2}\right)\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]} \\
=&\mathbb{E}\left[\log_2\left(1+\frac{\alpha_m|h_m|^2}{\alpha_n|h_m|^2+\frac{1}{\rho}}\right)\right],
\end{aligned} \tag{N.2}$$

which equals to $\mathbb{E}\left[R_m\right]$. In other words, when $\theta_m \to 0$, which refers to a user with no delay constraint, the EC in NOMA is equivalent to the ergodic capacity. Similarly, by using L'Hopital's rule, we can also conclude that $\lim\limits_{\theta_m\to 0} \bar{E}_{\text{c}}^{m} = \frac{1}{2}\mathbb{E}\left[\log_2\left(1+\rho|h_m|^2\right)\right]$, which equals to $\mathbb{E}\left[\bar{R}_m\right]$. Hence, when $\theta_m \to 0$, $E_{\text{c}}^{m} - \bar{E}_{\text{c}}^{m} = \mathbb{E}\left[R_m\right] - \mathbb{E}\left[\bar{R}_m\right]$. By following similar steps, we can get the same conclusion for the n^{th} user, i.e., $\lim\limits_{\theta_n\to 0} E_{\text{c}}^{n} = \mathbb{E}\left[R_n\right]$, $\lim\limits_{\theta_n\to 0} \bar{E}_{\text{c}}^{n} = \mathbb{E}\left[\bar{R}_n\right]$, and $\lim\limits_{\theta_n\to 0}\left(E_{\text{c}}^{n} - \bar{E}_{\text{c}}^{n}\right) = \mathbb{E}\left[R_n\right] - \mathbb{E}\left[\bar{R}_n\right]$.

Consider the m^{th} user with no delay constraint, i.e., $\theta_m \to 0$. By inserting $\rho \to \infty$ to (N.2), we can prove that $\lim\limits_{\substack{\theta_m\to 0\\ \rho\to\infty}} E_{\text{c}}^{m} = \mathbb{E}\left[\log_2\left(\frac{1}{\alpha_n}\right)\right]$. As for the EC

in OMA for the m^{th} user, we can get that $\lim\limits_{\substack{\theta_m \to 0 \\ \rho \to \infty}} \bar{E}_{\text{c}}^m \to \infty$, by inserting $\rho \to \infty$ into $\frac{1}{2}\mathbb{E}\left[\log_2\left(1+\rho|h_m|^2\right)\right]$. Henceforth, we can prove that $\lim\limits_{\substack{\theta_m \to 0 \\ \rho \to \infty}} \left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right) \to -\infty$.

Similarly, for the n^{th} user with $\theta_n \to 0$, when the transmit SNR ρ is very large, we can prove that $\lim\limits_{\substack{\theta_n \to 0 \\ \rho \to \infty}} E_{\text{c}}^n \to \infty$, and $\lim\limits_{\substack{\theta_n \to 0 \\ \rho \to \infty}} \bar{E}_{\text{c}}^n \to \infty$. As for $\lim\limits_{\substack{\theta_n \to 0 \\ \rho \to \infty}} \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right)$, we have that

$$\lim_{\substack{\theta_n \to 0 \\ \rho \to \infty}} \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right) \tag{N.3a}$$

$$= \lim_{\rho \to \infty} \mathbb{E}\left[\log_2\left(1+\rho\alpha_n|h_n|^2\right)\right] - \frac{1}{2}\mathbb{E}\left[\log_2\left(1+\rho|h_n|^2\right)\right] \tag{N.3b}$$

$$= \lim_{\rho \to \infty} \mathbb{E}\left[\log_2\left(\frac{1+\rho\alpha_n|h_n|^2}{\sqrt{1+\rho|h_n|^2}}\right)\right] \tag{N.3c}$$

$$= \lim_{\rho \to \infty} \mathbb{E}\left[\log_2\left(\frac{\frac{1}{\sqrt{\rho}}+\sqrt{\rho}\alpha_n|h_n|^2}{\sqrt{\frac{1}{\rho}+|h_n|^2}}\right)\right] \tag{N.3d}$$

$$= \lim_{\rho \to \infty} \mathbb{E}\left[\log_2\left(\sqrt{\rho}\alpha_n\sqrt{|h_n|^2}\right)\right], \tag{N.3e}$$

which approaches infinity. Hence, we complete the proof for Lemma 13. □

Bibliography

[1] GSMA Intelligence. (2014, Dec.) Understanding 5G: Perspectives on future technological advancements in mobile. [Online]. Available: https://gsmaintelligence.com/research/2014/12/understanding-5g/451/

[2] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5g mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[3] A. Osseiran, V. Braun, T. Hidekazu, P. Marsch, H. Schotten, H. Tullberg, M. A. Uusitalo, and M. Schellman, "The foundation of the mobile and wireless communications system for 2020 and beyond: Challenges, enablers and technology solutions," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Dresden, Germany, Jun. 2013.

[4] S. Zhang, Q. Wu, S. Xu, and G. Y. Li, "Fundamental green tradeoffs: Progresses, challenges, and impacts on 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 33–56, First Quarter 2017.

[5] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.

[6] C. Campolo, A. Molinaro, A. O. Berthet, and A. Vinel, "Full-duplex radios for vehicular communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 182–189, Jun. 2017.

[7] W. Viriyasitavat, M. Boban, H. Tsai, and A. Vasilakos, "Vehicular communications: Survey and challenges of channel and propagation models," *IEEE Veh. Technol. Mag.*, vol. 10, no. 2, pp. 55–66, Jun. 2015.

[8] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[9] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing the tactile internet: Haptic communications over next generation 5G cellular networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 82–89, Apr. 2017.

[10] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality-of-service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[11] J. Tang, "Qos-driven adaptive resource allocation for mobile wireless communications and networks," in *Ph.D. Dissertation, Texas A&M University*, Texas, USA, Dec. 2006.

[12] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058 – 3068, Aug. 2007.

[13] C. S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[14] ——, "Stability, queue length and delay, part i: Deterministic queueing networks," in *Proc. IEEE Conf. Decision Contr.*, vol. 1, Tucson, Arizona, USA, Dec. 1992, pp. 999–1004.

[15] ——, "Stability, queue length and delay, part ii: Stochastic queueing networks," in *Proc. IEEE Conf. Decision Contr.*, vol. 1, Tucson, Arizona, USA, Dec. 1992, pp. 1005–1010.

[16] C. S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 1995, pp. 1001–1009.

[17] C. S. Chang and J. A. Thomas, "Effective bandwidth in high speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[18] G. Kesidis, J. Walrand, and C. S. Chang, "Effective bandwidths for multiclass markov fluids and other ATM sources," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[19] A. Helmy, L. Musavian, and T. Le-Ngoc, "Energy-efficient power adaptation over a frequency-selective fading channel with delay and power constraints," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4529–4541, Sep. 2013.

[20] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 1536–1276, Dec. 2007.

[21] M. Gursoy, D. Qiao, and S. Velipasalar, "Analysis of energy efficiency in fading channels under QoS constraints," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4252–4263, Aug. 2009.

[22] L. Musavian and Q. Ni, "Delay-QoS-driven spectrum and energy efficiency tradeoff," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, Australia, Jun. 2014, pp. 4981 – 4986.

[23] ——, "Effective capacity maximization with statistical delay and effective energy efficiency requirements," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3824–3835, Jul. 2015.

[24] W. Cheng, X. Zhang, and H. Zhang, "Joint spectrum and power efficiencies optimization for statistical QoS provisionings over SISO/MIMO wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 903 – 915, Apr. 2013.

[25] X. Chen *et al.*, "Tradeoff between energy efficiency and spectral efficiency in a delay constrained wireless system," *Wirel. Commun. Mob. Comput.*, vol. 15, pp. 1945–1956, Mar. 2014.

[26] C. Xiong, G. Y. Li, Y. Liu, Y. Chen, and S. Xu, "Energy-efficient design for downlink OFDMA with delay-sensitive traffic," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 3085 – 3094, Jun. 2013.

[27] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[28] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[29] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.

[30] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy efficiency of resource scheduling for non-orthogonal multiple access (NOMA) wireless network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.

[31] Y. Zhang, H. Wang, T. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[32] X. Chen, Z. Zhang, C. Zhong, and D. W. K. Ng, "Exploiting multiple-antenna techniques for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2207–2220, Oct. 2017.

[33] Y. Zhang, H. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[34] Z. Qin, Y. Liu, Z. Ding, Y. Gao, and M. Elkashlan, "Physical layer security for 5G non-orthogonal multiple access in large-scale networks," in *IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May. 2016.

[35] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.

[36] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[37] D. Gamarnik. (2013, Fall) 15.070J advanced stochastic processes. [Online]. Available: https://ocw.mit.edu

[38] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communication and Computing.* Chapman and Hall/CRC, 1995.

[39] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications.* Springer-Verlag New York Inc., 2nd Edition, 1998.

[40] D. Gamarnik and J. Tsitsiklis. (2008, Fall) 6.436J fundamentals of probability. [Online]. Available: https://ocw.mit.edu

[41] O. Knill, *Probability Theory and Stochastic Processes with Applications.* Overseas Press, 2009.

[42] G. Grimmett and D. Welsh, *Probability: An Introduction.* Oxford University Press, 2014.

[43] M. Zukerman. (2016, Jan.) Introduction to queueing theory and stochastic teletraffic models. [Online]. Available: https://arxiv.org/pdf/1307.2968.pdf

[44] A. Weiss, "An introduction to large deviations for communication networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 938–952, Aug. 1995.

[45] J. K. Hunter, *An Introduction to Real Analysis.* Department of Mathematics, University of California at Davis, 2014.

[46] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective.* Cambridge University Press, 2013.

[47] J. T. Lewis and R. Russell, *An Introduction to Large Deviations for Teletraffic Engineers*, 1997.

[48] D. D. Kouvatsos, *Network Performance Engineering.* Springer-Verlag, 2011.

[49] M. J. Neely. (2008, Fall) Some background math notes on limsups, sets, and convexity. [Online]. Available: http://www-bcf.usc.edu/ mjneely/ee599/liminf-notes.pdf

[50] R. Ellis, "Large deviations for a general class of random vectors," *Ann. Probab.*, vol. 12, pp. 1–12, 1984.

[51] J. Gärtner, "On large deviations from invariant measure," *Theory Probab. Appl.*, vol. 22, pp. 24–39, 1977.

[52] J. A. Bucklew, *Introduction to Rare Event Simulation.* Springer-Verlag New York Inc., 2004.

[53] C. S. Chang, *Performance Guarantees in Communication Networks.* Springer-Verlag London, 2000.

[54] S. Mao and S. S. Panwar, "A survey of envelope processes and their applications in quality of service provisioning," *IEEE Commun. Surveys Tuts*, vol. 8, no. 3, pp. 2–20, Third Quarter 2006.

[55] R. L. Cruz, "A calculus for network delay, part i: Network elements in isolation," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 114–131, Jan. 1991.

[56] Y. Jiang and Y. Liu, *Stochastic Network Calculus.* Springer-Verlag London, 2008.

[57] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts*, vol. 17, no. 1, pp. 92–105, First Quarter 2015.

[58] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals. Math. Statist.*, vol. 23, pp. 493–507, 1952.

[59] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge Univ. Press, 2004.

[60] J. Nocedal and S. J. Wright, *Numerical Optimization.* Springer-Verlag New York, 2006.

[61] S. Schaible, "Minimization of ratios: Techinical notes," *J. Optim. Theory Appl.*, vol. 19, no. 2, pp. 347–352, Jun. 1976.

[62] ——, "Parameter-free convex equivalent and dual programs of fractional programming problems," *Zeitschrift fur Oper. Res.*, vol. 18, no. 5, pp. 187–196, Oct. 1974.

[63] ——, "Fractional programming," *Zeitschrift fur Oper. Res.*, vol. 27, no. 1, pp. 39–54, Dec. 1983.

[64] ITU Statistics. (2014, May) The world in 2014: ICT facts and figures. [Online]. Available: http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

[65] C. Gunaratne *et al.*, "Reducing the energy consumption of ethernet with adaptive link rate (ALR)," *IEEE Trans. Comput.*, vol. 57, no. 4, pp. 448–461, Apr. 2008.

[66] Smart2020, "Enabling the low-carbon economy in the information age," The Climate Group, London, U.K., Tech. Rep., 2008.

[67] W. D. Nordhaus, "To slow or not to slow: The economics of the greenhouse effect," *The Econ. J.*, vol. 101, no. 407, pp. 920–937, Jul. 1991.

[68] A. P. Bianzino *et al.*, "A survey of green networking reasearch," *IEEE Commun. Surveys Tuts*, vol. 14, no. 1, pp. 3–20, Feb. 2012.

[69] C. W. Tan, D. P. Palomar, and M. Chiang, "Energy-robustness tradeoff in cellular network power control," *IEEE/ACM Trans. Netw.*, vol. 17, no. 3, pp. 912–925, Jun. 2009.

[70] L. Zhang and C. W. Tan, "Cognitive radio network duality and algorithms for utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 3, pp. 500 – 513, Mar. 2013.

[71] Y. Li *et al.*, "Energy-efficient subcarrier assignment and power allocation in ofdma systems with max-min fairness guarantees," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3183 – 3195, Sep. 2015.

[72] X. Zhai, L. Zheng, and C. W. Tan, "Energy-infeasibility tradeoff in cognitive radio networks: Price-driven spectrum access algorithms," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 3, pp. 528–538, Mar. 2014.

[73] M. R. Mili, L. Musavian, K. A. Hamdi, and F. Marvasti, "How to increase energy efficiency in cognitive radio networks," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1829 – 1843, May. 2016.

[74] Q. Wu, W. Chen, M. Tao, J. Li, H. Tang, and J. Wu, "Resource allocation for joint transmitter and receiver energy efficiency maximization in downlink OFDMA systems," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 416 – 430, Feb. 2015.

[75] C. Bae and W. E. Stark, "End-to-end energy-bandwidth tradeoff in multihop wireless networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4051–4066, Sept. 2009.

[76] C. Xiong, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy- and spectral-efficiency tradeoff in downlink OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3874–3886, Nov. 2011.

[77] Y. Li *et al.*, "Energy efficiency and spectral efficiency tradeoff in interference-limited wireless networks," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1924–1927, Oct. 2013.

[78] X. Chen and S. Ouyang, "Energy- and spectral-efficiency trade-off in OFDMA-based cooperative cognitive radio networks," *Int. J. Distrib. Sens. Netw.*, vol. 2014, Feb. 2014.

[79] C. He *et al.*, "Energy- and spectral-efficiency tradeoff for distributed antenna systems with proportional fairness," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 894–902, May 2013.

[80] O. Amin, E. Bedeer, M. H. Ahmed, and O. A. Dobre, "Energy efficiency and spectral efficiency trade-off for OFDM systems with imperfect channel estimation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, Australia, Jun. 2014, pp. 3553–3558.

[81] F. Fu and M. van der Schaar, "Decomposition principles and online learning in cross-layer optimization for delay-sensitive applications," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1401 – 1415, Mar. 2010.

[82] W. Chen, M. J. Neely, and U. Mitra, "Energy-efficient transmissions with individual packet delay constraints," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2090 – 2109, May 2008.

[83] X. Zhang and J. Tang, "Power-delay tradeoff over wireless networks," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3673 – 3684, Sep. 2013.

[84] M. Ozmen and M. C. Gursoy, "Wireless throughput and energy efficiency with random arrivals and statistical queuing constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1375 – 1395, Mar. 2016.

[85] M. Sinaie, A. Zappone, E. A. Jorswieck, and P. Azmi, "A novel power consumption model for effective energy efficiency in wireless networks," *IEEE Wireless Commun. Lett.*, vol. 5, no. 2, pp. 2162 – 2337, Apr. 2016.

[86] W. Yu, L. Musavian, and Q. Ni, "Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3339–3353, Jan. 2016.

[87] T. Abrao, L. D. H. Sampaio, S. Yang, K. T. K. Cheung, P. J. E. Jeszensky, and L. Hanzo, "Energy efficient OFDMA networks maintaining statistical QoS guarantees for delay-sensitive traffic," *IEEE Access*, vol. 4, pp. 774 – 791, Feb. 2016.

[88] S. V. Hanly and D. N. C. Tse, "Multiaccess fading channels - part ii: delay-limited capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816 – 2831, Nov. 1998.

[89] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1468 – 1489, Jul. 1999.

[90] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[91] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Annu. Symp. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, London, UK, Sep. 2013, pp. 611–615.

[92] Z. Ma, Z. Zhang, Z. Ding, P. Fan, and H. Li, "Key techniques for 5G wireless communications: Network architecture, physical layer, and MAC layer perspectives," *Science China Information Sciences*, vol. 58, no. 4, pp. 1–20, Feb. 2015.

[93] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, no. 99, Oct. 2016.

[94] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[95] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 152–10 157, Dec. 2016.

[96] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[97] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1849–1858, Jan. 2017.

[98] L. Musavian and T. Le-Ngoc, "Energy-efficient power allocation over nakagmim fading channels under delay-outage constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4081–4091, Aug. 2014.

[99] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*. John Wiley and Sons Inc., 2000.

[100] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions*. New York: Dover, 1965.

[101] A. J. Goldsmith, *Wireless Communications*. Cambridge: Cambridge Univ. Press, 2005.

[102] C. Xiong, L. Lu, and G. Y. Li, "Energy efficiency tradeoff in downlink and uplink TDD OFDMA with simultaneous wireless information and power transfer," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Aug. 2014.

[103] S. Ruzika and M. M. Wiecek, "Survey paper: Approximation methods in multiobjective programming," *J. Optim. Theory Appl.*, vol. 126, no. 3, pp. 473–501, Sep. 2005.

[104] J. S. A. R. T. Marler, "Survey of multi-objective optimization methods for engineering," *Struct. Multidiscipl. Optim.*, vol. 26, no. 6, pp. 369–395, Apr. 2004.

[105] L. Zhang and C. W. Tan, "Maximizing sum rates in cognitive radio networks: Convex relaxation and global optimization algorithms," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 3, pp. 667 – 680, Mar. 2014.

[106] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[107] A. Helmy and T. Le-Ngoc, "Low-complexity QoS-aware frequency provisioning in downlink multi-user multicarrier systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Istanbul, Turkey, Apr. 2014, pp. 1785 – 1790.

[108] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171 – 178, Feb. 2003.

[109] 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.

[110] W. Yu, L. Musavian, and Q. Ni, "Statistical delay qos driven energy efficiency and effective capacity tradeoff for uplink multi-user multi-carrier systems," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3494–3508, Aug. 2017.

[111] 3GPP TD RP-150496, "Study on downlink multiuser superposition transmission," Tech. Rep.

[112] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley, New York, 3rd ed., 2003.

[113] M. Ehrgott, *Multicriteria Optimization*. Springer, 2005.

[114] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic Press, 6th ed., 2000.